Statistical Arbitrage

Algorithmic Trading Insights and Techniques

ANDREW POLE

John Wiley & Sons, Inc.

Statistical Arbitrage

Founded in 1807, John Wiley & Sons is the oldest independent publishing company in the United States. With offices in North America, Europe, Australia, and Asia. Wiley is globally committed to developing and marketing print and electronic products and services for our customers' professional and personal knowledge and understanding.

The Wiley Finance series contains books written specifically for finance and investment professionals as well as sophisticated individual investors and their financial advisors. Book topics range from portfolio management to e-commerce, risk management, financial engineering, valuation, and financial instrument analysis, as well as much more.

For a list of available titles, visit our Web site at www.WileyFinance.com.

Statistical Arbitrage

Algorithmic Trading Insights and Techniques

ANDREW POLE

John Wiley & Sons, Inc.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Wiley Bicentennial logo: Richard J. Pacifico.

Library of Congress Cataloging-in-Publication Data:

Pole, Andrew, 1961–
Statistical arbitrage : algorithmic trading insights and techniques / Andrew Pole.
p. cm. — (Wiley finance series)
Includes bibliographical references and index.
ISBN 978-0-470-13844-1 (cloth)
1. Pairs trading. 2. Arbitrage---Mathematical models. 3. Speculation--Mathematical models. I. Title.
HG4661.P65 2007
332.64'5 — dc22

2007026257

ISBN 978-0-470-13844-1

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Eliza and Marina

Contents

Preface xiii

Foreword xix

Acknowledgments xxiii

CHAPTER 1
Monte Carlo or Bust 1

Beginning 1
Whither? And Allusions 4

CHAPTER 2
Statistical Arbitrage 9

Introduction 9
Noise Models 10
Reverse Bets 11
Multiple Bets 11
Rule Calibration 12
Spread Margins for Trade Rules 16
Popcorn Process 18
Identifying Pairs 20
Refining Pair Selection 21
Event Analysis 22
Correlation Search in the Twenty-First Century 26
Portfolio Configuration and Risk Control 26
Exposure to Market Factors 29
Market Impact 30
Risk Control Using Event Correlations 31
Dynamics and Calibration 32
Evolutionary Operation: Single Parameter Illustration 34

CHAPTER 3
Structural Models **37**

Introduction 37
Formal Forecast Functions 39
Exponentially Weighted Moving Average 40
Classical Time Series Models 47
Autoregression and Cointegration 47
Dynamic Linear Model 49
Volatility Modeling 50
Pattern Finding Techniques 51
Fractal Analysis 52
Which Return? 52
A Factor Model 53
Factor Analysis 54
Defactored Returns 55
Prediction Model 57
Stochastic Resonance 58
Practical Matters 59
Doubling: A Deeper Perspective 61
Factor Analysis Primer 63
Prediction Model for Defactored Returns 65

CHAPTER 4
Law of Reversion **67**

Introduction 67
Model and Result 68
The 75 percent Rule 68
Proof of the 75 percent Rule 69
Analytic Proof of the 75 percent Rule 71
Discrete Counter 73
Generalizations 73
Inhomogeneous Variances 74
Volatility Bursts 75
Numerical Illustration 76
First-Order Serial Correlation 77
Analytic Proof 79
Examples 82
Nonconstant Distributions 82
Applicability of the Result 84
Application to U.S. Bond Futures 85

Summary 87
Appendix 4.1: Looking Several Days Ahead 87

CHAPTER 5
Gauss Is Not the God of Reversion 91

Introduction 91
Camels and Dromedaries 92
Dry River Flow 95
Some Bells Clang 98

CHAPTER 6
Interstock Volatility 99

Introduction 99
Theoretical Explanation 103
Theory versus Practice 105
Finish the Theory 105
Finish the Examples 106
Primer on Measuring Spread Volatility 108

CHAPTER 7
Quantifying Reversion Opportunities 113

Introduction 113
Reversion in a Stationary Random Process 114
Frequency of Reversionary Moves 117
Amount of Reversion 118
Movements from Quantiles Other Than the Median 135
Nonstationary Processes: Inhomogeneous Variance 136
Sequentially Structured Variances 136
Sequentially Unstructured Variances 137
Serial Correlation 138
Appendix 7.1: Details of the Lognormal Case in Example 6 139

CHAPTER 8
Nobel Difficulties 141

Introduction 141
Event Risk 142
Will Narrowing Spreads Guarantee Profits? 144
Rise of a New Risk Factor 145

Redemption Tension 148
Supercharged Destruction 150
The Story of Regulation Fair Disclosure (FD) 150
Correlation During Loss Episodes 151

CHAPTER 9
Trinity Troubles 155

Introduction 155
Decimalization 156
European Experience 157
Advocating the Devil 158
Stat. Arb. Arbed Away 159
Competition 160
Institutional Investors 163
Volatility Is the Key 163
Interest Rates and Volatility 165
Temporal Considerations 166
Truth in Fiction 174
A Litany of Bad Behavior 174
A Perspective on 2003 178
Realities of Structural Change 179
Recap 180

CHAPTER 10
Arise Black Boxes 183

Introduction 183
Modeling Expected Transaction Volume and Market Impact 185
Dynamic Updating 188
More Black Boxes 189
Market Deflation 189

CHAPTER 11
Statistical Arbitrage Rising 191

Catastrophe Process 194
Catastrophic Forecasts 198
Trend Change Identification 200
Using the Cuscore to Identify a Catastrophe 202
Is It Over? 204
Catastrophe Theoretic Interpretation 205
Implications for Risk Management 209

Sign Off 211
Appendix 11.1: Understanding the Cuscore 211

Bibliography 223

Index 225

Preface

These pages tell the story of statistical arbitrage. It is both a history, describing the first days of the strategy's genesis at Morgan Stanley in the 1980s through the performance challenging years of the early twenty-first century, and an exegesis of how and why it works. The presentation is from first principles and largely remains at the level of a basic analytical framework. Nearly all temptation to compose a technical treatise has been resisted with the goal of contributing a work that will be readily accessible to the larger portion of interested readership. I say "nearly all": Chapter 7 and the appendix to Chapter 11 probably belong to the category of "temptation not resisted." Much of what is done by more sophisticated practitioners is discussed in conceptual terms, with demonstrations restricted to models that will be familiar to most readers. The notion of a pair trade—the progenitor of statistical arbitrage—is employed to this didactic end rather more broadly than actual trading utility admits. In adopting this approach, one runs the risk of the work being dismissed as a pairs trading manual; one's experience, intent, and aspirations for the text are more extensive, but the inevitability of the former is anticipated. In practical trading terms, the simple, unelaborated pair scheme is no longer very profitable, nonetheless it remains a valuable tool for explication, retaining the capacity to demonstrate insight, modeling, and analysis while not clouding matters through complexity. After a quarter century in the marketplace, for profitable schemes beyond paper understanding and illustration, one needs to add some structural complexity and analytical subtlety.

One elaboration alluded to in the text is the assembling of a set of similar pairs (without getting into much detail on what metrics are used to gauge the degree of similarity), often designated as a group. Modeling such groups can be done in several ways, with some practitioners preferring to anchor a group on a notional archetype, structuring forecasts in terms of deviation of tradable pairs from the archetype; others create a formal implementation of the cohort as

a gestalt or a synthetic instrument. Both of those approaches, and others, can be formally analyzed as a hierarchical model, greatly in vogue (and greatly productive of insight and application) in mainstream statistical thinking for two decades; add to the standard static structure the dynamic element in a time series setting and one is very quickly building an analytical structure of greater sophistication than routinely used as the didactic tool in this book. Nonetheless, all such modeling developments rely on the insight and techniques detailed herein.

Those readers with deeper knowledge of mathematical and statistical science will, hopefully, quickly see where the presentation can be taken.

Maintaining focus on the structurally simple pair scheme invites readers to treat this book as an explicit "how to" manual. From this perspective, one may learn a reasonable history of the *what* and the *how* and a decent knowledge of *why it is possible*. Contemporary successful execution will require from the reader some additional thought and directed exploration as foregoing remarks have indicated. For that task, the book serves as a map showing major features and indicating where the reader must get out a compass and notebook. The old cartographers' device "Here be dragons" might be usefully remembered when you venture thus.

The text has, unashamedly, a statistician's viewpoint: Models can be useful. Maintaining a model's utility is one theme of the book. The statistician's preoccupation with understanding variation—the appreciation of the knowledge that one's models are wrong, though useful, and that the nature of the wrongness is illuminated by the structure of "errors" (discrepancies between observations and what a model predicts) is another theme of the book. Or, rather, not a distinct theme, but an overriding, guiding context for the material.

The notion of a pair trade is introduced in Chapter 1 and elaborated upon in Chapter 2. Following explication and exemplification, two simple theoretical models for the underlying phenomenon exploited by pairs, reversion, are proposed. These models are used throughout the text to study what is possible, illuminate how the possibilities might be exploited, consider what kinds of change would have negative impact on exploitation, and characterize the nature of the impact. Approaches for selecting a universe of instruments for modeling and trading are described. Consideration of change is

introduced from this first toe dipping into analysis, because temporal dynamics underpin the entirety of the project. Without the dynamic there is no arbitrage.

In Chapter 3 we increase the depth and breadth of the analysis, expanding the modeling scope from simple observational rules[1] for pairs to formal statistical models for more general portfolios. Several popular models for time series are described but detailed focus is on weighted moving averages at one extreme of complexity and factor analysis at another, these extremes serving to carry the message as clearly as we can make it. Pair spreads are referred to throughout the text serving, as already noted, as the simplest practical illustrator of the notions discussed. Where necessary to make our urgencies sensible, direct mention is made of other aspects of the arbitrageur's concern, including portfolio optimization and factor exposures. For the most part though, incursions into multivariate territory are avoided. Volatility modeling (and the fascinating idea of stochastic resonance) are treated separately here and in Chapter 6; elsewhere discussion is subsumed in that of the mean forecast process.

Chapter 4 presents a probability theorem that illuminates the prevalence of price moves amenable to exploitation by the simple rules first applied in the late 1980s. The insight of this result guides evaluation of exploitation strategies. Are results borne of brilliance on the part of a modeler or would a high school graduate perform similarly because the result is driven by structural dynamics, long in the public domain, revealed by careful observation alone? Many a claim of a "high" proportion of winning bets by a statistical arbitrageur has more to do with the latter than any sophistication of basic spread modeling or (importantly) risk management. When markets are disrupted and the conditions underlying the theoretical result are grossly violated, comparative practitioner performance reveals much about basic understanding of the nature of the process

[1]There is no pejorative intent in the use of the term: The rules were effective. Statistical content was limited to measurement of range of variation; no distributional study, model formulation, estimation, error analysis, or forecasting was undertaken prior to milking the observational insight. Those activities came soon enough—after the profits were piling up. With the expanded statistical study, adding trading experience to historical data, came insight into subtleties of the stock price motions exploited and the market forces driving repetitious occurrence of opportunities.

being exploited. Knowledge of the theoretical results often reveals itself more when assumptions are violated than when things are hunky dory and managers with solid understanding and those operating intellectually blind generate positive returns in equal measure. (Tony O'Hagan suggested that the basic probability result is long known, but I have been unable to trace it. Perhaps the result is too trivial to be a named result and exists as a simple consequence, a textbook exercise, of basic distribution theory. No matter, the implication remains profoundly significant to the statistical arbitrage story.)

Chapter 5 critiques a published article (whose authors remain anonymous here to protect their embarrassment) to clarify the broad conditions under which the phenomenon of reversion occurs. A central role for the normal distribution is dismissed. The twin erroneous claims that (a) a price series must exhibit a normal marginal distribution for reversion to occur, and (b) a series exhibiting a normal marginal distribution necessarily exhibits reversion are unceremoniously dispelled. There is reversion anywhere and everywhere, as Chapter 4 demonstrates.

Chapter 6 answers the question, important for quantifying the magnitude of exploitable opportunities in reversion gambits, "How much volatility is there in a spread?"

Chapter 7 is for the enthusiast not easily dissuaded by the presence of the many hieroglyphs of the probability calculus. Anyone with a good first course in probability theory can follow the arguments, and most can manage the detailed derivations, too. The mechanics are not enormously complicated. Some of the conceptual distinctions may be challenging at first—read it twice! The effort will be repaid as there is significant practical insight in the examples considered at length. Knowledge of how close theoretical abstractions come to reflecting measurable features of actual price series is invaluable in assessing modeling possibilities and simulation or trading results. Notwithstanding that remark, it is true that the remainder of the book does not rely on familiarity with the material in Chapter 7. While you may miss some of the subtlety in the subsequent discussions, you will not lack understanding for omitting attention to this chapter.

Chapters 8 through 10 might have been labeled "The Fall," as they characterize the problems that beset statistical arbitrage beginning in 2000 and directly caused the catastrophic drop in return during 2002–2004. An important lesson from this history is that there was not a single condition or set of conditions that abruptly

changed in 2000 and thereby eliminated forecast performance of statistical arbitrage models. What a story that would be! Far more dramatic than the prosaic reality, which is a complex mix of multiple causes and timings. All the familiar one liners, including decimalization, competition, and low volatility, had (and have) their moment, but none individually, nor the combination, can have delivered a blow to financial markets. Fundamentally altering the price dynamics of markets in ways that drastically diminish the economic potential in reversion schemes, mining value across the spectrum from the very high frequency hare of intra-day to the venerable tortoise of a month or more, requires a more profound explanation.

Change upon change upon change cataloged in Chapter 9 is at the root of the dearth of return to statistical arbitrage in 2002–2004. (Performance deterioration in 2000–2002 was evident but limited to a subset of practitioners.) This unusual episode in recent U.S. macroeconomic history is over, but the effects linger in the financial markets reflecting emergent properties of the collective behavior of millions of investors; and surely those investors continue to embody, no matter how lingering, those changes and the causes thereof.

The shift of trading from the floor of the New York Stock Exchange to internal exchanges, in the guise of computer algorithms designed by large brokerage houses and investment banks, has cumulatively become a change with glacier-like implacability. Slow. Massive. Irresistible. Crushing. Reforming.[2] A frequently remarked facet of the evolving dynamics is the decline of market volatility. Where has market volatility gone? In large part the algorithms have eaten it. Reduce the voice of a single participant yelling in a crowd and the babel is unaffected. Quite a significant proportion of participants and the reduced babel is oddly deafening. Now that computer programs (Chapter 10) ''manage'' over 60 percent of U.S. equity trades among ''themselves'' the extraodinary result is akin to administering a dose of ritalin to the hyperactive market child. In the commentary on low volatility two themes stand out: one is a lament over the lack

[2]One major structural consequence, fed also by technical advance in the credit markets and the development of Exchange Traded Funds, is literally the forming anew of patterns of price behavior detemined by the interaction of computer algorithms as agents for share dealings. In addition to this re-forming, reform is simultaneously underway with changes to Securities Exchange Commission regulations and NYSE rules.

of Keynes' animal spirits, a concern that the entrepreneurial genius of America is subdued even as Asian giants are stirring; the other is a fear that investors have forgotten the risks inherent in investment decisions, that inadvisable decisions are therefore being made that will have negative consequences in the near future. The inconsistency in those two characterizations is stark, but it can be rationalized. Contrary to the first notion, the spirit is quite animated—with a billion and a half shares changing ownership daily on the NYSE mart alone, what other conclusion should one draw? There is plenty of spirit: simply its animus is satisfied with less overt fuss. Algorithms don't have emotions. So there is plenty of innovative risk taking, but low volatility by historical standards, induced by trading technologies, has not yet been properly internalized by many market participants. Viewing contemporary volatility levels in the manner to which historical experience has been accustomed ineluctably leads to excessive risk taking.

Chapter 10 is interesting in its own right, notwithstanding any relationship to the evolution of statistical arbitrage opportunities. Algorithms and computer driven trading are changing the financial world in many ways. Electronic exchanges have already been seen off most of the world's peopled trading places—and who among us believes that the floor of the NYSE will be more than a museum, parking lot, or memory in a year or two?

Chapter 11 describes the phoenix of statistical arbitrage, rising out of the ashes of the fire created and sustained by the technological developments in algorithmic trading. New, sustained patterns of stock price dynamics are emerging. The story of statistical arbitrage has returned to a new beginning. Will this fledgling fly?

The renaissance predicted in Chapter 11, drafted in 2005, is already coming to pass. Since at least early 2006 there has been a resurgence of performance from those practitioners who persisted through the extremely challenging dynamic changes of 2003–2005. Interestingly, while there are new systematic patterns in the movements of relative equity prices, some old patterns have also regained potency. Adoption of algorithmic trading is accelerating, with tools now offered by more than 20 vendors. In another technology driven development, beginning with Goldman Sachs in late 2006, at least two offerings of general hedge fund replication by algorithmic means have been brought to market. This is an exciting as well as exacting time for statistical arbitrageurs.

Foreword

Mean reversion in prices, as in much of human activity, is a powerful and fundamental force, driving systems and markets to homeostatic relationships. Starting in the early 1980s, statistical arbitrage was a formal and successful attempt to model this behavior in the pursuit of profit. Understanding the arithmetic of statistical arbitrage (sometimes abbreviated as stat. arb.) is a cornerstone to understanding the development of what has come to be known as complex financial engineering and risk modeling.

The trading strategy referred to as statistical arbitrage is generally regarded as an opaque investment discipline. The view is that it is being driven by two complementary forces, both deriving from the core nature of the discipline: the vagueness of practitioners and the lack of quantitative knowledge on the part of investors. Statistical arbitrage exploits mathematical models to generate returns from systematic movements in securities prices. Granted, no investment manager is inclined to divulge the intricate "how-tos" of his business. While stock pickers can tell a good story without revealing the heart of their decision making, that is not the case with model-based strategies developed by "quants." A description with any meaningful detail at all quickly points to a series of experiments from which an alert listener can try to reverse-engineer the strategy. That is why quant practitioners talk in generalities that are only understandable by the mathematically trained.

Opacity has also increased the need for mathematical maturity on the part of investors seeking to assess managers. To comprehend what a statistical arbitrageur is saying beyond a glib level, one needs to understand advanced mathematics beyond the college level. This, naturally, limits the audience. The limitation is perpetuated by the lack of reference material from which to learn. *Statistical Arbitrage* now fills that void.

Statistical arbitrage has been in existence for approximately 25 years. During that time, the general concepts have been widely

disseminated via the storytelling of early implementers to interested investment bank analysts and academics. Nevertheless, opacity remains because practitioners have steadily increased the sophistication of their modeling—and for good commercial reasons remained obscure about their innovations. In the wide dissemination of basic stat. arb. concepts, the term mean reversion as well as its variant, reversion to the mean, looms very large. Reversion to the mean is a simple concept to illustrate: Children of unusually tall parents are typically shorter than their parents; children of unusually short parents are typically taller than their parents. This is a concept that is easy for most people to grasp. Translating this idea to the motions of security prices means that securities prices return to an average value. So far, so good. But then we hit a problem. Height reversion is an intergenerational phenomenon, while price reversion is an entity dynamic.

Prices returning from where? And to what average value? The average height of adults is a familiar concept, even if the precise quantification requires a little work. Even children as young as grade-school age can give a reasonable estimate of the average height of the adults they know, and by extension, of the average height of local adult populations. There is no such common grounding of observation or experience to apply to securities prices. They are all over the map. Scaling is arbitrary. They can grow many fold. And they can collapse to zero. People do not grow to the sky and then revert back to some average, but security prices can.

Even if we suppose that the questions have been reasonably answered, other technicalities immediately pose themselves: How does one identify when a price is away from the mean and by how much? How long will the return to the mean take?

Here is where the opacity enters the discussion and makes its permanent home. The language of mathematical models compounds the unfamiliarity of the notions, generating a sense of disquiet, a fear of lack of understanding.

In *Statistical Arbitrage,* Pole has given his audience a didactic tour of the basic principles of statistical arbitrage, eliminating opacity at the Statistical Arbitrage 101 level. In the 1980s and early 1990s, Stat. Arb. 101 was, for the most part, all there was (exceptions such as D.E. Shaw and Renaissance aside). Today, more than a decade later, there is a much more extensive and complex world of statistical arbitrage.

This is not unlike the natural world, which is now populated by incredibly complex biological organisms after four billion years of evolution. Yet the simplest organisms thrive everywhere and still make up by far the largest part of the planet's biomass. So is it true in statistical arbitrage, where the basics underpin much of contemporary practice.

Statistical Arbitrage describes the phenomena, the driving forces generating those phenomena, the patterns of dynamic development of exploitable opportunities, and models for exploitation of the basic reversion to the mean in securities prices. It also offers a good deal more, from hints at more sophisticated models to valuable practical advice on model building and performance monitoring—advice applicable far beyond statistical arbitrage.

Chapters 1 and 2 speak to the genesis of statistical arbitrage, the venerable pairs trading schemes of the 1980s, with startling illustration of the enormous extent and productivity of the opportunities. This demonstration sets the scene for theoretical development, providing the first step to critical understanding of practical exploitation with rules for calibrating trade placement. More penetration of opacity follows in Chapter 5 where the relationship between (a) reversion in securities prices watched day-by-day and (b) statistical descriptions (distributions) of collections of such daily prices viewed as a glob devoid of the day-by-day context, is clearly spelled out.

Chapters 8 and 9 tell of the midlife crisis of statistical arbitrage. The roiling of United States financial markets for many months, beginning with the Enron debacle in 2000 and running through the terrorist attacks of 2001 and what Pole calls "an appalling litany" of corporate misconduct, is dissected for anticipated impact on statistical arbitrage performance. Adding to that mix have been technical changes in the markets, including decimalization and the decline of independent specialists on the floor of the NYSE. Pole draws a clear picture of why statistical arbitrage performance was disrupted. Very clearly the impression is made that the disruption was not terminal.

Chapters 10 and 11 speak to the arriving future of statistical arbitrage. Trading algorithms, at first destroyers of classical stat. arb. are now, Pole argues, progenitors of new, systematically exploitable opportunities. He labels one of the new motions the "catastrophe move"; a detailed exposition of modeling the dynamics follows a

catastrophe-theory explication of a possible rationale for the behavioral pattern. The unmistakable impression is that statistical arbitrage is rising once again.

The tone of *Statistical Arbitrage* is direct and thorough. Obfuscation is in short supply. Occasionally, the tone is relieved with a bit of lightheartedness—the tadpole-naming story in a note to Chapter 11 is a gem—and throughout, refreshing prose is to be found.

In describing mathematical models, authors readily produce unmemorable, formulaic wording offering nothing by way of interpretation or explanation beyond what is provided by the algebra itself. *Statistical Arbitrage* is an exception—a break in the cloud of opacity—a mean that Pole has avoided reverting to!

Gregory van Kipnis

April 23, 2007
New York City

Acknowledgments

I was introduced to statistical arbitrage by Gregg van Kipnis. In many ways, the contents of this volume are directly the result of our collaboration and it is a pleasure to acknowledge the intellectual debt. Our conversations often strayed far beyond statistical arbitrage to macroeconomics and general science and very often to politics, none of which is reproduced here in recognizable form. Those discussions were not always motivated by statistical arbitrage considerations, though occasionally we would hit on a useful metaphor from an unrelated topic that subsequently proved fruitful in thinking about statistical arbitrage. It is not in the nature of things that individual such recollections can now be pointed to with certainty to say whose idea ABC was. Credit is rightfully due to van Kipnis; the rendition in these pages is entirely my responsibility.

The editorial and production staff at Wiley, in particular Bill Falloon, Emilie Herman, Laura Walsh, and Stacey Fischkelta though we never physically met, have been helpful and courteous throughout the project.