Statistical Arbitrage and Algorithmic Trading : Overview and Applications

Miquel Noguer Alonso - Licenciado y Master en Administración y Dirección de Empresas – Universitat Ramon Llull - ESADE

Facultad de Ciencias Económicas y Empresariales

Departamento de Economía Aplicada Cuantitativa II

UNED

2010

Facultad de Ciencias Económicas y Empresariales

Departamento de Economía Aplicada Cuantitativa II

Statistical Arbitrage and Algorithmic Trading : Overview and Applications

Miquel Noguer Alonso - Licenciado y Master en Administración y Dirección de Empresas – Universitat Ramon Llull – ESADE - UBS AG

Director : Andreu Pacheco Pages IFAE/CERN

Codirector : Manuel Jose Sanchez Sanchez UNED

This PhD Thesis is a tribute to my family.

To my mother and brother, whose confidence, support in me has been a constant in my life, to my wife Mima, whose love, patience, and understanding have been the foundations of this work.

To the memory of my father, wherever he is, I love him so much.

To my son Jordi who came early this year to inspire my work, bringing to our family such happiness that cannot be imagined.

Acknowledgements

This PhD thesis is the result of more than a decade of work with my talented colleagues in UBS, Andbanc as many other people in the financial industry.

Thanks to my Thesis Director, Andreu Pacheco Pages, for his knowledge, patience, help with ideas and devoting his precious time to my work!

Thanks to Alberto Alvarez and Jose Manuel Sanchez for their guidance, dedication and support through all my PhD time at their university.

Thanks to Christian Mazza, Jean Pierre Gabriel and Ales Janka for their collaboration in the research in Fribourg University. I am greatly indebted to Yi-Chen-Zhang and Damien Challet for bringing me there.

Thanks to Lorenzo Moneta from CERN for his collaboration in the machine learning chapter of this PhD thesis.

Thanks to Jose Miguel Dominguez for his collaboration in the factor models section of this work and the interesting discussions we have on quantitative investing.

Thanks to Stephen Wolfram and Jason Cawley from Wolfram Research for their support and the research we did together in the NKS summer school back in 2008.

My discussions and research with Martin Schaden have extremely useful to explore new ideas like quantum finance.

My work wouldn't have been the same without my conversations with Jean-Phillipe Bouchaud, Conrad Perez Vicente and other econophysicists. Whose contribution to economics science is changing the foundations of Finance. I am highly indebted to the Fisica i Finances research group from the Physics department at Universitat de Barcelona for their inspiring papers and books.

Table of Figures and Tables

1. Introduction

1.1. Scope

In this PhD thesis I present the most successful approaches in the exciting world of quantitative investing or algorithmic trading, introducing new concepts and applications to achieve superior risk adjusted returns. Two important aspects of quantitative trading will be covered: statistical arbitrage and algorithmic trading. In general both disciplines try to find exploitable regularities, trends, anomalies in the financial data (alone or using factors).

The thesis starts discussing in chapter 1 the theory of forecasting financial markets and our ultimate goal: achieving high sharpe ratios

In chapter 2 we introduce the Algorithmic Trading Strategies: track record, categories and disciplines.

The historical performance and risk metrics of the systematic traders index is shown since 1987 to april 2010, the track record of quantitative strategies. In this chapter the categories of quantitative strategies are described. The topics covered are mean reversion, momentum, regime switching and factor models, I present in this chapter the analytical proof of mean reversion given some conditions on the time series are met, this theorem can be extended to some of the well known stochastic process driving the markets. Unfortunately as we see in the tests I performed, it is not easy to find the right set of conditions in the real life markets to make this simple beautiful strategy work.

In Chapters 3 to 4 the main processes used to model financial variables are reviewed. The parallel between discrete-time processes, mainly used by econometricians for risk- and portfolio-management, and their continuous-time counterparts, mainly used by mathematicians to price derivatives is defined. The thesis highlights the relationship of such processes with the building blocks of stochastic dynamics and statistical inference, namely the invariants.

In chapter 5 the thesis introduces the multivariate Ornstein-Uhlenbeck process and discusses how it generalizes a vast class of continuous-time and discrete-time multivariate processes. Relying on the simple geometrical interpretation of the dynamics of the Ornstein-Uhlenbeck process I introduce cointegration and its relationship to statistical arbitrage. I show an application to swap contract strategies.

Statistical model estimation main concepts and techniques are described in chapter 6.

In Chapter 7 we analyse nonlinear dynamical systems, including apparently simple cases that can exhibit chaotic behaviour. I only outline the cases that may be relevant to our subject of interest.

Technical analysis is reviewed in chapter 8, sometimes misunderstood, always controversial, we see here the concepts that in my view can add value in the trading arena, I skip the chart concepts because in my view cannot be analysed properly in a rigourous backtesting. In the

applications part of this thesis I show that some indicators once optimized show impressive results.

In these first 8 chapters in introduce the most relevant theoretical concepts in the following chapters I concentrate in empirical applications of these concepts.

Finance is an empirical science, all concepts that can be useful to our ultimate goal should be backtested and analysed in depth.

In chapter 9 I analyse the historical performance (1928 - 6/2009) of 2 different price momentum and 2 value strategies ranking stocks both across and within industries in the US. The gross Sharpe ratios of the various strategies show interesting results. The market exposure of value and momentum strategies has changed significantly over time but a combined signal appears to be uncorrelated with market returns and indeed market neutral in a broader sense. This work extends and expands the results of some of the most well known Fama-French portfolios/factors.

In this chapter I show a momentum strategy of 60 highly liquid futures and currency forwards during the period from January 1985 to December 2009 applied in a long-short portfolio context, the historical sharpe ratio of 1.42 shows that momentum defined in that way add significant value to invest in many asset classes.

A short machine learning review is provided in chapter 10.

In chapter 11, the performance of ordinal GA-based trading strategies is evaluated under six classes of time series model, namely, the linear ARMA model, the bilinear model, the ARCH model, the GARCH model, the threshold model and the chaotic model. The performance criteria employed are the winning probability, accumulated returns, Sharpe ratio and luck coefficient. Asymptotic test statistics for these criteria are derived. The hypothesis as to the superiority of GA over a benchmark, say, buy-and-hold, can then be tested using Monte Carlo simulation. From this rigorously established evaluation process, we find that simple genetic algorithms can work very well in linear stochastic environments, and that they also work very well in nonlinear deterministic (chaotic) environments. However, they may perform much worse in pure nonlinear stochastic cases. These results shed light on the superior performance of GA when it is applied to the two tick-by-tick time series of foreign exchange rates: EUR/USD and USD/JPY.

In chapter 12 I show a new application in which we create a "machine" that learns how to trade different regimes in the SP500 and Oil using Artificial Neural Networks to detect the trends and Genetic Algorithms to classify the regimes. The results show that these non-linear engines maybe able to give some interesting Sharpe ratios. I use the idea of finding trends in different regimes. The caveat here is obvious, when the regimes are massively different from the past, this machine learning technique isn't adding much value.

In Section 12.6 I present the results of using Support Vector Machines and macroeconomic variables to forecast the direction of the NIKKEI 225 Index.

This chapter opens up the door to show some other interesting non-linear techniques.

In chapter 13 other non-linear applications are shown: takes theorem – delay embedding theorem, the application of VARMA / reduced form VAR to forecast GDP and a markov switching application to SP500 and VIX.
In chapter 14 we show what I call Asymmetric Algorithmic trading strategies that exploit asymmetric statistical distributions. I show an application of Asymmetric Algorithmic Trading Strategies that extracts value from mean-reverting or trend following type of assets that show asymmetric return distribution and high volatility, this strategy uses Exponentially Weighted Moving Average (EWMA) indicators to generate buying signals and selling signals. Sharpe ratio is around 1.43 from 1997 to may 2010.
In chapter 15, the Integrated Algorithmic Trading Strategy Theory is introduced: a systematic methodology/framework to invest in a liquid asset class performing a set of tests and then using the most appropriate techniques and a combination of these in a portfolio of systematic strategies. 2 applications of the theory are shown a GOLD (commodity) IATST and a Multi Strategy IATST.
The fact that markets are so difficult to model makes all the work necessary.
I use in this PHD thesis algorithmic trading to define strategies based in algorithms other than the ones based in statistical concepts included in statistical arbitrage. In other contexts algorithmic trading is sometimes used to define the discipline of using computer algorithms to execute orders, to mitigate market impact. Here we define it as the use of computer algorithms to make trading decisions.

1.2. Definitions

1.2.1 Forecasting

To predict (or forecast) is to form an expectation of what will happen in the future. The actual notion of forecastability hinges on how we can forecast the future given what we know today. Forecasting is the relationship between present information and future events. If the future returns of a financial asset do not depend on any variable known today, then returns are unpredictable.
The difference between financial predictions and others, say, weather predictions is that they influence the markets themselves; weather forecasts do not influence the weather itself.
Stock return forecasts are not certain; uncertain predictions are embodied in probability distributions (Bayesian statistics).
The idea of predicting the future has always fascinated people and has been the subject, successively, of magic, philosophical enquiry, and scientific debate. Already in ancient times, it was clear that the notion of predicting the future is subject to potential inconsistencies. If, for example, we received a “credible” prediction that tomorrow we will have a car accident on our way to work, we might either decide to leave the car at home or be extra careful with our driving. In either case, our behavior will have been influenced by the prediction, thus potentially invalidating

the prediction. It is because of inconsistencies of this type that Samuelson (1965)[1] and Fama (1965) arrived at the apparently paradoxical conclusion that “properly anticipated prices fluctuate randomly.”

It is widely acknowledged that financial time series modelling and forecasting is an arduous task. These time series behave very much like a random walk process and several studies have concluded that their serial correlation is economically and statistically insignificant. The same studies seem to confirm the efficient market hypothesis (EMH)[2] developed by Professor Eugene Fama at the University of Chicago Booth School of Business as an academic concept of study through his published Ph.D. thesis in the early 1960s at the same school, which maintains that the current market price of a stock fully reflects—at any time—the available information assimilated by traders.

As new information enters the system, the imbalance is immediately detected and promptly redressed by a counteracting change in market price. Depending on the type of information examined, three forms of EMH exist: weak, semi-strong, and strong. We are particularly concerned with the weak EMH, which only takes into account past stock price data. In this case, the underlying assumption is that no predictions can be made based on stock price data alone, as they follow a random walk in which successive changes have zero correlation.

The Adaptive Market Hypothesis, as proposed by Andrew Lo (2004,2005)[3], is an attempt to reconcile theories that imply that the markets are efficient with behavioral alternatives, by applying the principles of evolution - competition, adaptation, and natural selection - to financial interactions. Under this approach the traditional models of modern financial economics can coexist alongside behavioral models. He argues that much of what behavioralists cite as counterexamples to economic rationality - loss aversion, overconfidence, overreaction, and other behavioral biases - are, in fact, consistent with an evolutionary model of individuals adapting to a changing environment using simple heuristics.

This hypothesis implies that future changes in stock market prices cannot be predicted from information about past prices. Notwithstanding these difficulties, most stock market investors seem convinced that they can statistically predict price trends and make a profit. This is done by exploiting technical or fundamental analysis rules; mean reversion, as well as ‘‘momentum strategies’’ (i.e., buying when the market is bullish and selling when it is bearish). For these reasons, many attempts have been made to model and forecast financial markets, using all the computational tools available for studying time series and complex systems: linear auto-regressive models, principal component analysis, artificial neural networks (ANNs), genetic algorithms (GAs), and others.

The forecastability of stock returns continues to be at the center of a heated debate. It is believed that (1) predictable processes allow investors (or asset managers on behalf of their clients) to earn excess returns whereas (2) unpredictable processes do not allow one to earn excess returns. Neither is necessarily true. Understanding why will shed some light on the crucial issues in modelling. In a nutshell: (1) predictable expectations do not necessarily mean profit if they are

associated with unfavourable risk and (2) unpredictable expectations can be profitable if their expected value is favourable (positive alpha).

Most of our knowledge is uncertain; our forecasts are also uncertain. The development of probability theory gave us the conceptual tools to represent and measure the level of uncertainty. Probability theory assigns a number—the probability—to every possible event. This number, the probability, might be interpreted in one of two ways:

• The probability of an event is a quantitative measure of the strength of our beliefs that a particular event will happen, where 1 represents certainty;

• Probability is the percentage of times (i.e., frequency) that we observe a particular event in a large number of observations.

The second interpretation is the one normally used in econometrics and in science at large. When we make a probabilistic forecast of an event, we assess the percentage of times that we expect to observe that event.

1.2.2 The ultimate goal: achieving high Sharpe ratios

The Sharpe ratio or Sharpe index or Sharpe measure or reward-to-variability ratio is a measure of the excess return (or Risk Premium) per unit of risk in an investment asset or a trading strategy, named after William Forsyth Sharpe. Since its revision by the original author in 1994[4], it is defined as:

$$SR = \frac{R - Rf}{\sigma} = \frac{E\left(R - R_f\right)}{\sqrt{VAR\left(R - R_f\right)}} \quad (1)$$

The Sharpe ratio is used to characterize how well the return of an asset compensates the investor for the risk taken, the higher the Sharpe ratio number the better. When comparing two assets each with the expected return E[R] against the same benchmark with return Rf, the asset with the higher Sharpe ratio gives more return for the same risk.

As we know our forecasts is uncertain the quality of our forecasts will be ultimately measured and assessed according its sharpe ratio, so adjusting the returns obtained by risk, risk defined as standard deviation. In doing so what we are doing is forecasting not only returns but also risk.

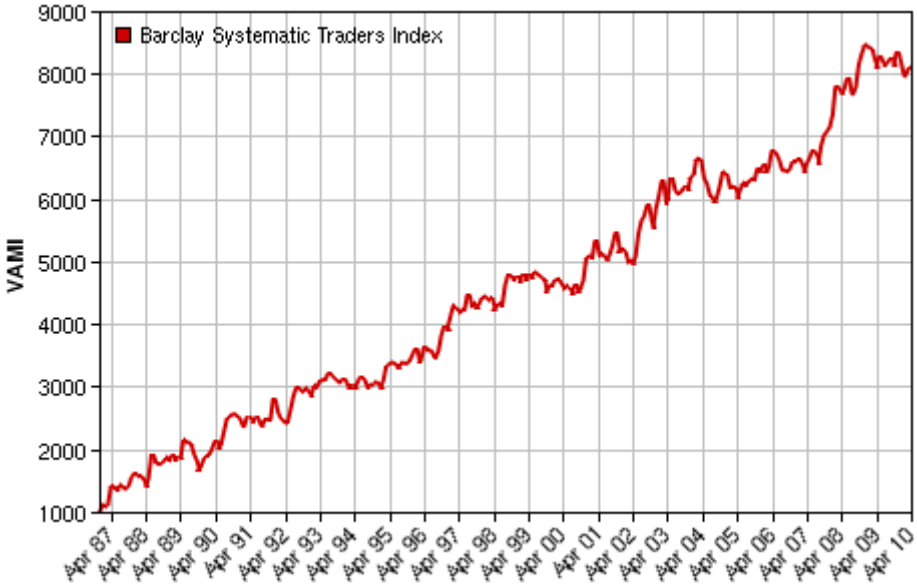
2. Algorithmic Trading strategies: track record, categories and disciplines

2.1. The industry of systematic traders: Barclay Systematic Traders Index

To start our analysis we provide some historical returns of the systematic traders index, the universe of systematic traders, managers that use systematic approaches to invest, systematic and algorithmic are in what follows the same thing.

An equal weighted composite of managed programs whose approach is at least 95% systematic. In 2010 there are 417 systematic programs included in the index.

Figure 1: : Barclay Systematic Traders Index 05/1987- 4/2010



An equal weighted composite of managed programs whose approach is at least 95% systematic. In 2010 there are 417 systematic programs included in the index.

Table 1: Annual returns - Barclay Systematic Traders Index

1980	-	**1990**	34.58%	**2000**	9.89%
1981	-	**1991**	13.37%	**2001**	2.99%
1982	-	**1992**	3.25%	**2002**	12.09%
1983	-	**1993**	8.19%	**2003**	8.71%
1984	-	**1994**	-3.18%	**2004**	0.54%
1985	-	**1995**	15.27%	**2005**	0.95%
1986	-	**1996**	11.58%	**2006**	2.10%
1987	63.01%	**1997**	12.76%	**2007**	8.72%
1988	12.22%	**1998**	8.12%	**2008**	18.16%
1989	1.18%	**1999**	-3.71%	**2009**	-3.38%

Table 2: Perfomance metrics 1987-04/2010

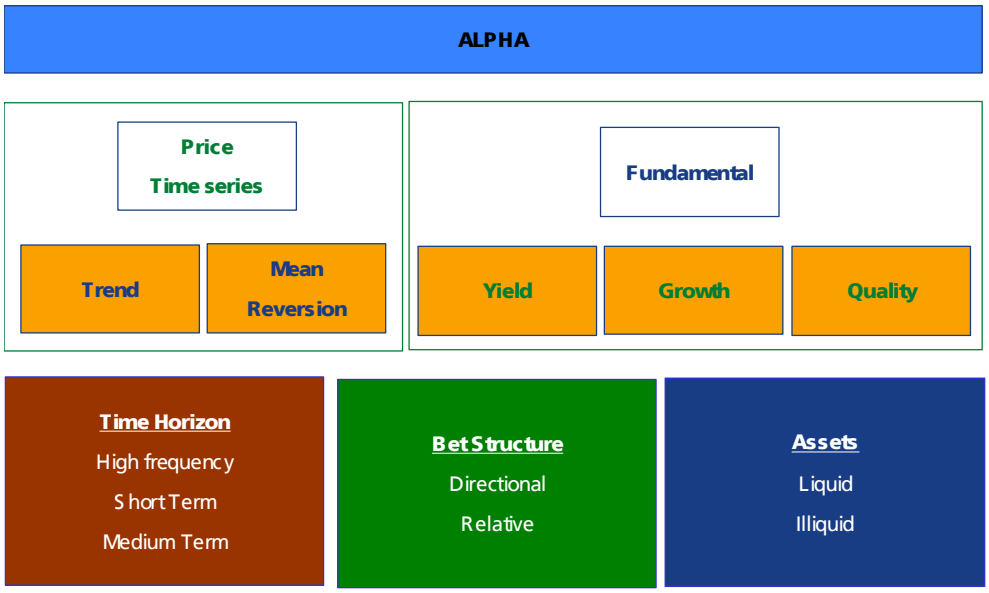
Compound Annual Return	9.39%
Sharpe Ratio	0.38
Worst Drawdown	22.07%
Correlation vs S&P 500	-0.04
Correlation vs US Bonds	0.07
Correlation vs World Bonds	-0.05

Here we have in these two tables the performance metrics of the systematic traders index. These metrics tell us two stories, reasonable but not outstanding sharpe ratios and good feature an almost 0 correlation to other assets classes, so that these strategies combine well with the traditional long only bonds and stocks.

2.2. Trading strategies categories : Mean reversion, momentum / Regime switching / Factor models

We will describe the two basic categories of trading strategies: mean-reverting versus momentum strategies. Periods of mean reverting and trending behaviors are examples of what some traders call regimes, and the switch between different regimes is a topic of discussion here. Mean-reverting strategies derive their mathematical justification from the concepts of stationarity and cointegration of time series, which we will cover next. Then we will describe a the theory of factor models. Other categories of strategies that traders frequently discuss are seasonal trading and high-frequency strategies. All trading strategies require a way to exit their positions; I will describe the different logical ways to do this. Finally, we ponder the question of how to best enhance the returns of a strategy: through higher leverage or trading higher-beta stocks?

Figure 2: Taxonomy of Systematic strategies



2.2.1 Mean-reverting versus momentum strategies

Trading strategies can be profitable only if securities prices are either mean-reverting or trending. Otherwise, they are randomwalking, and trading will be futile. If you believe that prices are mean reverting and that they are currently low relative to some reference price, you should buy now and plan to sell higher later. However, if you believe the prices are trending and that they are currently low, you should (short) sell now and plan to buy at an even lower price later. The opposite is true if you believe prices are high.

Academic research has indicated that stock prices are on average very close to random walking. However, this does not mean that under certain special conditions, they cannot exhibit some

degree of mean reversion or trending behavior. Furthermore, at any given time, stock prices can be both mean reverting and trending depending on the time horizon you are interested in. Constructing a trading strategy is essentially a matter of determining if the prices under certain conditions and for a certain time horizon will be mean reverting or trending, and what the initial reference price should be at any given time. (When the prices are trending, they are also said to have "momentum," and thus the corresponding trading strategy is often called a momentum strategy.)

Some people like to describe the phenomenon that prices can be both mean reverting and trending at the same time as the "fractal" nature of stock prices. Technical analysts or chartists like to use the so-called Elliott wave theory to analyze such phenomena.

Still others like to use the discipline of machine learning or artificial intelligence (in particular, techniques such as hidden Markov models, Kalman filter, neural networks, etc.) to discover whether the prices are in a mean-reverting or trending "regime."

In fact, financial researchers (Khandani and Lo, 2007)[5] have constructed a very simple short-term mean reversal model that is profitable (before transaction costs) over many years. Of course, whether the mean reversion is strong enough and consistent enough such that we can trade profitably after factoring in transaction costs is another matter, and it is up to you, the trader, to find those special circumstances when it is strong and consistent.

Though mean reversion is quite prevalent, backtesting a profitable mean-reverting strategy can be quite perilous.

Many historical financial databases contain errors in price quotes. Any such error tends to artificially inflate the performance of mean-reverting strategies. It is easy to see why: a mean-reverting strategy will buy on a fictitious quote that is much lower than some moving average and sell on the next correct quote that is in line with the moving average and thus makes a profit. One must make sure the data is thoroughly cleansed of such fictitious quotes before one can completely trust your backtesting performance on a mean-reverting strategy.

Survivorship bias also affects the backtesting of mean-reverting strategies disproportionately, Stocks that went through extreme price actions are likely to be either acquired (the prices went very high) or went bankrupt (the prices went to zeros). A mean-reverting strategy will short the former and buy the latter, losing money in both cases. However, these stocks may not appear at all in your historical database if it has survivorship bias, thus artificially inflating your backtest performance.

Momentum can be generated by the slow diffusion of information—as more people become aware of certain news, more people decide to buy or sell a stock, thereby driving the price in the same direction. I suggested earlier that stock prices may exhibit momentum when the expected earnings have changed. This can happen when a company announces its quarterly earnings, and investors either gradually become aware of this announcement or they react to this change by incrementally executing a large order (so as to minimize market impact). And indeed, this leads to a momentum strategy called post earnings announcement drift, or PEAD.

Essentially, this strategy recommends that you buy a stock when its earnings exceed expectations, and short a stock when it falls short. More generally, many news announcements have the potential of altering expectations of a stock's future earnings, and therefore have the potential to trigger a trending period. As to what kind of news will trigger this, and how long the trending period will last, it is again up to you to find out.

Besides the slow diffusion of information, momentum can be caused by the incremental execution of a large order due to the liquidity needs or private investment decisions of a large investor.

This cause probably accounts for more instances of short-term momentum than any other causes. With the advent of increasingly sophisticated execution algorithms adopted by the large brokerages, it is, however, increasingly difficult to ascertain whether a large order is behind the observed momentum. Momentum can also be generated by the herdlike behavior of investors: investors interpret the (possibly random and meaningless) buying or selling decisions of others as the sole justifications of their own trading decisions. As Yale economist Robert Schiller said in the New York Times (Schiller, 2008)[6], nobody has all the information they need in order to make a fully informed financial decision. One has to rely on the judgment of others. There is, however, no sure way to discern the quality of the judgment of others. More problematically, people make their financial decisions at different times, not meeting at a town hall and reaching a consensus once and for all. The first person who paid a high price for a house is "informing" the others that houses are good investments, which leads another person to make the same decision, and so on. Thus, a possibly erroneous decision by the first buyer is propagated as "information" to a herd of others.

Unfortunately, momentum regimes generated by these two causes (private liquidity needs and herdlike behavior) have highly unpredictable time horizons. How could you know how big an order an institution needs to execute incrementally? How do you predict when the "herd" is large enough to form a stampede? Where is the infamous tipping point? If we do not have a reliable way to estimate these time horizons, we cannot execute a momentum trade profitably based on these phenomena. In a later section on regime switch, I will examine some attempts to predict these tipping or "turning" points.

There is one last contrast between mean-reverting and momentum strategies that is worth pondering. What are the effects of increasing competition from traders with the same strategies?

For mean-reverting strategies, the effect typically is the gradual elimination of any arbitrage opportunity, and thus gradually diminishing returns down to zero. When the number of arbitrage opportunities has been reduced to almost zero, the mean-reverting strategy is subject to the risk that an increasing percentage of trading signals are actually due to fundamental changes in stocks' valuation and thus is not going to mean revert. For momentum strategies, the effect of competition is often the diminishing of the time horizon over which the trend will continue. As news disseminates at a faster rate and as more traders take advantage of this trend earlier on, the equilibrium price will be reached sooner. Any trade entered after this equilibrium price is reached will be unprofitable.

The academic rationale underlying momentum effects: the Lifecycle of a trend

Start of the Trend: Under-reaction to information

Research has linked this under-reaction to a number of behavioral tendencies and market frictions that lead to actions that slow down the process of price discovery:

1) Anchor-and-insufficient-adjustment

Edwards (1968)[7], Tversky and Kahneman (1974)[8] find that people anchor their views to historical data and adjust their views insufficiently to new information. This behavior can cause prices to under-react to news.

2) The disposition effect Shefrin and Statman (1985)[9], Frazzini (2006)[10] observe that people tend to sell winners too early and ride losers too long. They sell winners too early because they like to realize their gains. This selling creates downward price pressure, which slows down the upward price adjustment to the new fundamental level. On the other hand, people hang on to losers for too long since realizing losses is painful. Instead, they try to "make back" what has been lost. In this case, the absence of willing sellers keeps prices from adjusting downward as fast as they should.

3) Non-profit-seeking market participants who fight

Trends Silber (1994)[11] argues that central banks operate in the currency and fixed-income markets to reduce exchange-rate volatility and manage inflation expectations, thus potentially slowing down the price-adjustment to news. As another example, hedging activity in commodity markets can also slow down price discovery. These effects make the price initially move too little in response to news, which creates a continued price drift as the market realizes the full importance of the news over time. A managed futures strategy will tend to position itself in relation to the initial news and therefore profit if the trend continues.

Trend Continuation: Over-reaction

Once a trend has started, a number of other phenomena exist which have the potential to extend the trend:

1) Herding and feedback trading

De Long et al. (1990)[12] and Bikhchandani et al. (1992)[13] argue that when prices have moved up or down for a while, some traders may jump on the bandwagon, and this herding effect can feed on itself. Herding has been documented among equity analysts in their recommendations and earnings forecasts, in institutional investors' investment decisions, and in mutual fund investors who tend to move from funds with recent poor performance and herd into funds that have recently done well. 2) Confirmation bias and representativeness Wason (1960)[14] and Tversky and Kahneman (1974) show that people tend to look for information that confirms what they already believe and look at recent price moves as representative of the future. This can lead investors to move capital into investments that have recently made money, and conversely out of investments that have declined, causing trends to continue.

3) Risk management

Garleanu and Pedersen (2007)[15] argue that some risk management schemes imply selling in down markets and buying in up markets, in line with the trend. For instance, stop-losses get

triggered causing buying/selling in the same direction of the movement. Another example is that a drop in price is often associated with higher volatility (or Value at Risk), leading traders to reduce positions.

End of the Trend

Obviously, trends cannot go on forever. At some point, prices extend beyond underlying fundamental value. As people start to realize that prices have gone too far, they revert towards fundamental value and the trend dies out. The market may become range bound until new events cause price moves and set off new trends. One of the main challenges for managed futures strategies is to minimize losses associated with the ending of trends and to preserve capital in range bound markets that do not exhibit trends.

The law of mean reversion

The first result, presented in this section, is a simple probability theorem that evidences a basic law guaranteeing the presence of reversion in prices in an efficient market. We present a model for forecasting prices of financial instruments that guarantees 75 percent forecasting accuracy. The chosen setting is prediction about the daily spread range of a pair but a little reflection will reveal a much wider applicability. Specifically, we focus on predicting whether the spread tomorrow will be greater or smaller than the spread today.

The model is quite simple. If the spread today is greater than the expected average spread, then predict that the spread tomorrow will be smaller than the spread today. On the other hand, if the spread today was less than the expected average spread, then predict that the spread tomorrow will be greater than the spread today.

The model just described is formalized as a probability model as follows. Define a sequence of identically distributed, independent continuous random variables {Pt, t = 1, 2, . . .} with support on the nonnegative real line and median m. Then:

$$\Pr\left[\left(\boldsymbol{P}_t > \boldsymbol{P}_{t-1} \cap \boldsymbol{P}_{t-1} < \boldsymbol{m}\right) \cup \left(\boldsymbol{P}_t < \boldsymbol{P}_{t-1} \cap \boldsymbol{P}_{t-1} > \boldsymbol{m}\right)\right] = 0.75 \quad (2)$$

In the language of the motivating spread problem, the random quantity Pt is the spread on day t (a nonnegative value), and days are considered to be independent. The two compound events comprising the probability statement are straightforwardly identified with the actions specified in the informal prediction model above. But a word is in order regarding the details of each event. It is crucial to note that each event is a conjunction, and, and not a conditional, given that, as might initially be considered appropriate to represent the if –then nature of the informal model. The informal model is a prescription of the action that will be taken; the probability in which we are interested is the probability of how often those actions (predictions) will be correct. Thus, looking to expected performance, we want to know how often the spread on a given day will exceed the spread on the previous day when at the same time the spread on that previous day does not exceed the median value. Similarly, we want to know how often the spread on a given day will not exceed the spread on the previous day when at the same time the spread on that previous day

exceeds the median. Those distinctions may seem arcane but proper understanding is critical to the correct evaluation of expected result of a strategy.

Suppose that on eight days out of ten the spread is precisely equal to the median. Then the scheme makes a prediction only for 20 percent of the time. That understanding flows directly from the conjunction/disjunction distinction. With the wrong understanding a five-to-one ratio of expected return to actual return of a scheme would ensue.

Operationally, one may bet on the outcome of the spread tomorrow once today's spread is confirmed (close of trading). On those days for which the spread is observed to be greater than the median spread, the bet for tomorrow is that the exhibited spread tomorrow will be less than the spread seen today. The proportion of winning bets in such a scheme is the conditional given that probability:

$$\Pr\left[\boldsymbol{P}_{t+1} < \boldsymbol{P}_t \middle| \boldsymbol{P}_t > \boldsymbol{m}\right] = \frac{3}{4} \quad (3)$$

See on the appendix the analytic proof.

Similarly, bets in the other direction will be winners three quarters of the time. Does this mean that we win ‘‘1.5 of the time?’’ Now that really would be a statistical arbitrage! The missing consideration is the relative frequency with which the conditioning event occurs. Now, Pt < m occurs half of the time by definition of the median. Therefore, half of the time we will bet on the spread decreasing relative to today and of those bets, three quarters will be winners. The other half of the time we will bet on the spread increasing relative to today and of those bets, three quarters will also be winners. Thus, over all bets, three quarters will be winners. (In the previous illustration, the conditioning events occur only 20 percent of the time and the result would be 3/4 × 1/5 or just 3/20) .Note that the assumption of continuity is crucial (hence the emphasis in the model statement). It is trivial to show that the result is not true for discrete variables

Empirical tests

Unfortunately this rather easy to prove and to trade model impeccable mathematically does not always hold true in reality, the identically distributed, independent continuous random variable condition to be met does not hold for many markets in reality. Obviously one of the biggest difficulties is finding the median/average to which the serie is reverting to.

SP500 - EUR/USD

At least the concept is not tradable using local median and averages for 20 and 10 days for the SP500 (daily data : 1/1928 / 4/2010) – EUR / USD (daily data : 1/1980 / 4/2010) in where the law in holding well below 50% instead of 75% telling us a complete different story so suggesting that interesting autocorrelations maybe in place.

2.2.2 Regime switching

The concept of regimes is most basic to financial markets. What else are “bull” and “bear” markets if not regimes? The desire to predict regime switches, which are also commonly known

as turning points, is also as old as financial markets themselves. If our attempts to predict the switching from a bull to a bear market were even slightly successful, we could focus our discussion to this one type of switching and call it a day. If only it were that easy. The difficulty with predicting this type of switching encourages researchers to look more broadly at other types of regime switching in the financial markets, hoping to find some that may be more amenable to existing statistical tools.

We have already described two regime switches (or "shifts," for these two examples did not switch back to their former regimes) that are due to changes in market and regulatory structures: decimalization of stock prices in 2003 and the elimination of the short sale plus-tick rule in 2007. These regime shifts are preannounced by the government, so no predictions of the shifts are necessary, though few people can predict the exact consequences of the regulatory changes.

Some of the other most common financial or economic regimes studied are inflationary vs. recessionary regimes, high- vs. low volatility regimes, and mean-reverting vs. trending regimes. Among these, volatility regime switching seems to be most amenable to classical econometric tools such as the generalized autoregressive conditional heteroskedasticity (GARCH) model (See Klaassen, 2002)[16]. That is not surprising, as there is a long history of success among financial economists in modelling volatilities as opposed to the underlying stock prices themselves. While such predictions of volatility regime switches can be of great value to options traders, they are unfortunately of no help to stock traders.

Academic attempts to model regime switches in stock prices generally proceed along these lines:

1. Propose that the two (or more) regimes are characterized by different probability distributions of the prices. In the simplest cases, the log of the prices of both regimes may be represented by normal distributions, except that they have different means and/or standard deviations.
2. Assume that there is some kind of transition probability among the regimes.
3. Determine the exact parameters that specify the regime probability distributions and the transition probabilities by fitting the model to past prices, using standard statistical methods such as maximum likelihood estimation.
4. Based on the fitted model above, find out the expected regime of the next time step and, more importantly, the expected stock price.

This type of approach is usually called Markov regime switching or hidden Markov models, and it is generally based on a Bayesian probabilistic framework. Readers who are interested in reading more about some of these approaches may peruse Nielsen and Olesen (2000)[17], van Norden and Schaller (1993)[18], or Kaufmann and Scheicher (1996)[19].

Despite the elegant theoretical framework, such Markov regime switching models are generally useless for actual trading purposes. The reason for this weakness is that they assume constant transition probabilities among regimes at all times. In practice, this means that at any time (as illustrated by the Nielsen and Olesen paper), there is always a very small probability for the stock to transition from a normal, quiescent regime to a volatile regime. But this is useless to traders who want to know when—and under what precise conditions—the transition probability will

suddenly peak. This question is tackled by the turning points models. Turning points models take a data mining approach (Chai, 2007)[20]: Enter all possible variables that might predict a turning point or regime switch. Variables such as current volatility; last-period return; or changes in macroeconomic numbers such as consumer confidence, oil price changes, bond price changes, and so on can all be part of this input. In fact, in a very topical article about turning points in the real estate market by economist Robert Schiller (2007)[21], it was suggested that the crescendo of media chatter about impending boom or bust may actually be a good predictor of a coming turning point.

2.2.3 Stationarity and cointegration

A time series is "stationary" if it never drifts farther and farther away from its initial value. In technical terms, stationary time series are "integrated of order zero," or I(0). (See Alexander, 2001.)[22] It is obvious that if the price series of a security is stationary, it would be a great candidate for a mean-reversion strategy. Unfortunately, most stock price series are not stationary—they exhibit a geometric random walk that gets them farther and farther away from their starting (i.e., initial public offering) values. However, you can often find a pair of stocks such that if you long one and short the other, the market value of the pair is stationary. If this is the case, then the two individual time series are said to be cointegrated. They are so described because a linear combination of them is integrated of order zero. Typically, two stocks that form a cointegrating pair are from the same industry group. Traders have long been familiar with this so-called pair-trading strategy. They buy the pair portfolio when the spread of the stock prices formed by these pairs is low, and sell/short the pair when the spread is high—in other words, a classic mean-reverting strategy. If a price series (of a stock, a pair of stocks, or, in general, a portfolio of stocks) is stationary, then a mean-reverting strategy is guaranteed to be profitable, as long as the stationarity persists into the future (which is by no means guaranteed). However, the converse is not true. You don't necessarily need a stationary price series in order to have a successful mean-reverting strategy. Even a nonstationary price series can have many short-term reversal opportunities that one can exploit, as many traders have discovered. Many pair traders are unfamiliar with the concepts of stationarity and cointegration. But most of them are familiar with correlation, which superficially seems to mean the same thing as cointegration. Actually, they are quite different. Correlation between two price series actually refers to the correlations of their returns over some time horizon (for concreteness, let's say a day). If two stocks are positively correlated, there is a good chance that their prices will move in the same direction most days. However, having a positive correlation does not say anything about the long-term behaviour of the two stocks. In particular, it doesn't guarantee that the stock prices will not grow farther and farther apart in the long run even if they do move in the same direction most days. However, if two stocks were cointegrated and remain so in the future, their prices (weighted appropriately) will

be unlikely to diverge. Yet their daily (or weekly, or any other time horizon) returns may be quite uncorrelated.

Stationarity is not limited to the spread between stocks: it can also be found in certain currency rates. For example, the Canadian dollar/Australian dollar (CAD/AUD) cross-currency rate is quite stationary, both being commodities currencies. Numerous pairs of futures as well as well as fixed-income instruments can be found to be cointegrating as well. (The simplest examples of cointegrating futures pairs are calendar spreads: long and short futures contracts of the same underlying commodity but different expiration months. Similarly for fixed-income instruments, one can long and short bonds by the same issuer but of different maturities.)

2.2.4 Factor models

A factor is a common characteristic among a group of assets. Factors should be founded on sound economic intuition, market insight or an anomaly.

Factors fall into 3 categories – macroeconomic, cross-sectional and statistical factors.

How do we quantify these and other common drivers of returns? There is a well-known framework in quantitative finance called factor models (also known as arbitrage pricing theory [APT]) that attempts to capture the different drivers of returns such as earnings growth rates, interest rate, or the market capitalization of a company. These drivers are called factors. Mathematically, we can write the excess returns (returns minus risk-free rate) R of N stocks as

$$\boldsymbol{R} = \boldsymbol{Xb} + \boldsymbol{u} \quad (4)$$

where X is an N × N matrix of factor exposures (also known as factor loadings), b is an N vector of factor returns, and u an N vector of specific returns. (Every one of these quantities is time dependent, but I suppress this explicit dependence for simplicity.)

The terms factor exposure, factor return, and specific return are commonly used in quantitative finance, and it is well worth our effort to understand their meanings. Factor returns are the common drivers of stock returns, and are therefore independent of a particular stock. Factor exposures are the sensitivities to each of these common drivers. Any part of a stock's return that cannot be explained by these common factor returns is deemed a specific return (i.e., specific to a stock and essentially regarded as just random noise within the APT framework). Each stock's specific return is assumed to be uncorrelated to another stock's.

Let's illustrate these using a simple factor model called the Fama-French Three-Factor model (Fama and French, 1992)[23]. This model postulates that the excess return of a stock depends linearly on only three factor exposures: its beta (i.e., its sensitivity to the market index), its market capitalization, and its book-to-price ratio. These factor exposures are obviously different for each stock and for each time period. (Factor exposures are often normalized such that the average of the factor exposures within a universe of stocks is zero, and the standard deviation is 1.) Now that we know how to calculate the factor exposures, what about the factor returns and specific returns? We cannot directly compute the factor returns and specific returns—we have to infer

their values by running a multivariate linear regression of the excess returns of stocks against the factor exposures. Note that each stock represents one data point in this linear regression, and we have to either run a separate linear regression for each time period or, if we want an average value over many time periods, aggregate the values from all these time periods into one training set and run one regression against them all.

If you perform this linear regression fit over many time periods for the Fama-French Three-Factor model, you will find that the market capitalization factor return is usually negative (meaning that small-cap stocks usually outperform large-cap stocks), and the book-to-price ratio factor return is usually positive (meaning value stocks usually outperform growth stocks). And since most stocks are positively correlated with the market index, the beta factor return is positive as well.

The Fama-French model has no monopoly on the choice of factors. In fact, you can construct as many factors as creativity and rationality allow. For example, you can choose return on equity as a factor exposure, or the correlation of the stock return with the prime rate as another. You can choose any number of other economic, fundamental, or technical factors. Whether the factor exposures you have chosen are sensible or not will determine whether the factor model explains the excess returns of the stocks adequately. If the factor exposures (and consequently the model as a whole) are poorly chosen, the linear regression fit will produce specific returns of significant sizes, and the R2 statistic of the fit will be small. According to experts (Grinold and Kahn, 1999)[24], the R2 statistic of a good factor model with monthly returns of 1,000 stocks and 50 factors is typically about 30 percent to 40 percent. It may appear that these factor models are only explanatory in retrospect—that is, given historical returns and factor exposures, we can compute the factor returns of those historical periods. But what good are those historical factor returns for our trading? It turns out that often factor returns are more stable than individual stock returns. In other words, they have momentum. You can therefore assume that their values remain unchanged from the current period (known from the regression fit) to the next time period. If this is the case, then, of course, you can also predict the excess returns, as long as the factor exposures are well chosen and therefore the time-varying specific returns are not significant. Let me clarify one point of potential confusion. Even though I stated that factor models can be useful as a predictive model (and therefore for trading) only if we assume the factor returns have momentum, it does not mean that factor models cannot capture mean reversion of stock returns. You can, in fact, construct a factor exposure that captures mean reversion, such as the negative of the previous period return. If stock returns are indeed mean reverting, then the corresponding factor return will be positive. If you are interested in building a trading model based on fundamental factors, there are a number of vendors from whom you can obtain historical factor data.

Until now we have seen the static factor models. Dynamic factor models are models that allow the asset manager to specify dynamics for factors themselves.

Sargent[25] and Geweke[26] proposed a dynamic factor model of the type:

$$\boldsymbol{r}_t = \sum_{i=0}^{\infty} \beta_i \boldsymbol{f}_{t-i} + \varepsilon_t \quad (5)$$

where returns are an N x 1 vector, the β, are the N x Q matrices, $\boldsymbol{f}_t$ is a K x 1 vector for each t, and ε_t is a N x 1 vector. It is assumed that N is finite, K<<N and T tends to inifinity. It is assumed that factors and residuals are uncorrelated and that residuals are mutually uncorrelated though possibly autocorrelated. This model is the dynamic equivalent of the strict factor model. Estimation is performed with maximum likelihood in the frequency domain. The number of factors is determined with a likelihood ratio test.

How good are the performances of factor models in real trading?

Naturally, it mostly depends on which factor model we are looking at. But one can make a general observation that factor models that are dominated by fundamental and macroeconomic factors have one major drawback—they depend on the fact that investors persist in using the same metric to value companies. This is just another way of saying that the factor returns must have momentum for factor models to work. For example, even though the value (book-to-price ratio) factor returns are usually positive, there are periods of time when investors prefer growth stocks such as during the Internet bubble in the late 1990s, and in August and December of 2007. As The Economist noted, one reason growth stocks were back in favor in 2007 is the simple fact that their price premium over value stocks has narrowed significantly (Economist, 2007b). Another reason is that as the U.S. economy slowed, investors increasingly opted for companies that still managed to generate increasing earnings instead of those that were hurt by the recessionary economy.

Therefore, it is not uncommon for factor models to experience steep drawdown during the times when investors' valuation method shifts, even if only for a short duration. But then, this problem is common to practically any trading model that holds stocks overnight.

2.2.5 High-frequency trading strategies

In general, if a high Sharpe ratio is the goal of your trading strategy then you should be trading at high frequencies, rather than holding stocks overnight.

What are high-frequency trading strategies, and why do they have superior Sharpe ratios? Many experts in high-frequency trading would not regard any strategy that holds positions for more than a few seconds as high frequency, but here I would take a more pedestrian approach and include any strategy that does not hold a position overnight. Many of the early high-frequency strategies were applied to the foreign exchange market, and then later on to the futures market, because of their abundance of liquidity. In the last six or seven years, however, with the increasing liquidity in the equity market, the availability of historical tick database for stocks, and mushrooming computing power, this type of strategies has become widespread for stock trading as well.

The reason why these strategies have Sharpe ratio is simple: Based on the “law of large numbers,” the more bets you can place, the smaller the percent deviation from the mean return you will experience. With high-frequency trading, one can potentially place hundreds if not thousands of bets all in one day. Therefore, provided the strategy is sound and generates positive mean return, you can expect the day-to-day deviation from this return to be minimal.

With this high Sharpe ratio, one can increase the leverage to a much higher level than longer term strategies can, and this high leverage in turn boosts the return-on-equity of the strategy to often stratospheric levels.

Of course, the law of large numbers does not explain why a particular high-frequency strategy has positive mean return in the first place. In fact, it is impossible to explain in general why high frequency strategies are often profitable, as there are as many such strategies as there are fund managers. Some of them are mean reverting, while others are trend following. Some are market-neutral pair traders, while others are long-only directional traders. In general, though, these strategies aim to exploit tiny inefficiencies in the market or to provide temporary liquidity needs for a small fee. Unlike betting on macroeconomic trends or company fundamentals where the market environment can experience upheavals during the lifetime of a trade, such inefficiencies and need for liquidity persists day to day, allowing consistent daily profits to be made. Furthermore, high-frequency strategies typically trade securities in modest sizes. Without large positions to unwind, risk management for high-frequency portfolios is fairly easy: “Deleveraging” can be done very quickly in the face of losses, and certainly one can stop trading and be completely in cash when the going gets truly rough. The worst that can happen as these strategies become more popular is a slow death as a result of gradually diminishing returns. Sudden drastic losses are not likely, nor are contagious losses across multiple accounts. Though successful high-frequency strategies have such numerous merits, it is not easy to backtest such strategies when the average holding period decreases to minutes or even seconds.

Transaction costs are of paramount importance in testing such strategies. Without incorporating transactions, the simplest strategies may seem to work at high frequencies. As a consequence, just having high-frequency data with last prices is not sufficient—data with bid, ask, and last quotes is needed to find out the profitability of executing on the bid versus the ask. Sometimes, we may even need historical order book information for backtesting. Quite often, the only true test for such strategies is to run it in real-time unless one has an extremely sophisticated simulator. Backtesting is only a small part of the game in high-frequency trading. High-speed execution may account for a large part of the actual profits or losses. Professional high-frequency trading firms have been writing their strategies in C instead of other, more user-friendly languages, and locating their servers next to the exchange or a major Internet backbone to reduce the microsecond delays. So even though the Sharpe ratio is appealing and the returns astronomical, truly high-frequency trading is not by any means easy for an independent trader to achieve in the beginning. But there is no reason not to work toward this goal gradually as expertise and resources accrue.

2.2.6 What is your exit strategy?

While entry signals are very specific to each trading strategy, there isn't usually much variety in the way exit signals are generated. They are based on one of these:

- A fixed holding period
- A target price or profit cap
- The latest entry signals
- A stop price

A fixed holding period is the default exit strategy for any trading strategy, whether it is a momentum model, a reversal model, or some kind of seasonal trading strategy, which can be either momentum or reversal based. (More on this later.) I said before that one of the ways momentum is generated is the slow diffusion of information. In this case, the process has a finite lifetime. The average value of this finite lifetime determines the optimal holding period, which can usually be discovered in a backtest. One word of caution on determining the optimal holding period of a momentum model: As I said before, this optimal period typically decreases due to the increasing speed of the diffusion of information and the increasing number of traders who catch on to this trading opportunity. Hence, a momentum model that has worked well with a holding period equal to a week in the backtest period may work only with a one-day holding period now. Worse, the whole strategy may become unprofitable a year into the future. Also, using a backtest of the trading strategy to determine holding period can be fraught with data-snooping bias, since the number of historical trades may be limited. Unfortunately, for a momentum strategy where the trades are triggered by news or events, there are no other alternatives. For a mean-reverting strategy, however, there is a more statistically robust way to determine the optimal holding period that does not depend on the limited number of actual trades. The mean reversion of a time series can be modelled by an equation called the Ornstein-Uhlenbeck formula (Unlenbeck, 1930)[27].

Let's say we denote the mean-reverting spread (long market value minus short market value) of a pair of stocks as z(t). Then we can write :

$$dz(t) = -\phi(z(t) - \mu)dt + dW \quad (6)$$

where μ is the mean value of the prices over time, and dW is simply some random Gaussian noise. Given a time series of the daily spread values, we can easily find θ (and μ) by performing a linear regression fit of the daily change in the spread dz against the spread itself. Mathematicians tell us that the average value of z(t) follows an exponential decay to its mean μ, and the half-life of this exponential decay is equal to ln(2)/θ, which is the expected time it takes for the spread to revert to half its initial deviation from the mean. This half-life can be used to determine the optimal holding period for a mean-reverting position. Since we can make use of the entire time series to find the best estimate of θ, and not just on the days where a trade was triggered, the estimate for the half-life is much more robust than can be obtained directly from a trading model.

If you believe that your security is mean reverting, then you also have a ready-made target price—the mean value of the historical prices of the security, or μ in the Ornstein-Uhlenbeck formula. This target price can be used together with the half-life as exit signals (exit when either criterion is met).

Target prices can also be used in the case of momentum models if you have a fundamental valuation model of a company. But as fundamental valuation is at best an inexact science, target prices are not as easily justified in momentum models as in mean-reverting models. If it were that easy to profit using target prices based on fundamental valuation, all investors have to do is to check out stock analysts' reports every day to make their investment decisions.

Suppose you are running a trading model, and you entered into a position based on its signal. Some time later, you run this model again. If you find that the sign of this latest signal is opposite to your original position (e.g., the latest signal is "buy" when you have an existing short position), then you have two choices. Either you simply use the latest signal to exit the existing position and become flat or you can exit the existing position and then enter into an opposite position. Either way, you have used a new, more recent entry signal as an exit signal for your existing position. This is a common way to generate exit signals when a trading model can be run in shorter intervals than the optimal holding period.

Notice that this strategy of exiting a position based on running an entry model also tells us whether a stop-loss strategy is recommended. In a momentum model, when a more recent entry signal is opposite to an existing position, it means that the direction of momentum has changed, and thus a loss (or more precisely, a drawdown) in your position has been incurred. Exiting this position now is almost akin to a stop loss. However, rather than imposing an arbitrary stop-loss price and thus introducing an extra adjustable parameter, which invites data-snooping bias, exiting based on the most recent entry signal is clearly justified based on the rationale for the momentum model. Consider a parallel situation when we are running a reversal model. If an existing position has incurred a loss, running the reversal model again will simply generate a new signal with the same sign. Thus, a reversal model for entry signals will never recommend a stop loss. (On the contrary, it can recommend a target price or profit cap when the reversal has gone so far as to hit the opposite entry threshold.) And, indeed, it is much more reasonable to exit a position recommended by a mean-reversal model based on holding period or profit cap than stop loss, as a stop loss in this case often means you are exiting at the worst possible time. (The only exception is when you believe that you have suddenly entered into a momentum regime because of recent news.)

2.2.7 Event trading

Event trading refers to strategies that place trades on the markets' reaction to events, the events maybe economic, industry-specific and security-specific occurrences that consistently affect the securities of interest on a recurring manner. For example, the Fed Funds rates consistently raise

the value of the US dollar, simultaneously raising the rate for USD/CAD. The Fed announcements of the Fed funds decisions are events potentially suitable for profitable strategies.
Most event strategies follow a three stage development process:

1. For each event type, identify dates and times of past events in historical data.
2. Compute historical price changes and desired frequencies pertaining to securities of interest and surrounding the events identified in first step.
3. Estimate expected price responses base on historical price behaviour surrounding past events.

2.2.8 Volatility arbitrage

Last subject we want to discuss regarding quantitative trading strategies categories is volatility arbitrage.
Supposing that the volatility pricing model we are dealing with is correct, and if the options are mistaken in evaluating the stock distribution during their lifetime, there should be an arbitrage opportunity to take advantage of. The ninth chapter of the Härdle et al. book has a description of these strategies. Note that both these strategies are European and cannot be changed until maturity. At this point we should reiterate that the profit and loss of this trade could be used as an empirical and model-free measure of how consistent or inconsistent the information embedded in the options is with the one in the underlying stocks.

Skewness Trades

To capture an undervalued third moment, we can buy OTM calls and sell OTM puts. Note that Aït-Sahalia[28] says that the options are overly skewed, which means that the options skew is larger in absolute value. However, given the negative sign of the skew, the cross-sectional skew is actually lower than the one implied by the time series, hence the described strategy.
Note that in order to be immune to parallel shifts of the volatility curve, we should make the trade as vega-neutral as possible. The correspondence between the call and the put is usually not one-to-one. Therefore, calling V the vega, Δ's the hedge ratios for C the call and P the put option, then the hedged portfolio π will be:

$$\prod = \boldsymbol{C}(\boldsymbol{S}_t, \boldsymbol{K}_C) - \frac{\boldsymbol{V}_C}{\boldsymbol{V}_P}\boldsymbol{P}(\boldsymbol{S}_t, \boldsymbol{K}_P) - \left(\Delta_C - \frac{\boldsymbol{V}_C}{\boldsymbol{V}_P}\Delta_P\right)\boldsymbol{S}_t \quad (7)$$

and the positions in the options should be dynamically adjusted in theory. However, that would cause too much transaction cost and exposure to the bid-ask spread. As we shall see in the paragraph on "exact replication," more-elaborate strategies are available to better exploit the third-moment differences.
Kurtosis Trades

To capture an overvalued fourth moment, we need to use the “fat tails” of the distribution. For this we can, for instance, sell ATM and far OTM options, and buy close OTM options.

Directional Risks

Despite the delta-hedging, the skewness trade applied to an undervalued third moment has an exposure to the direction of the markets. A bullish market is favorable to it, and a bearish one unfavorable. The kurtosis trade applied to an overvalued fourth moment generates a profit if the market stays at one point and moves sideways but loses money if there are large movements.
The skewness and kurtosis trading strategies above are profitable given the options’ implied moments, unless the options were actually right in factoring in a large and sudden downward movement. This also makes sense because the way the options were priced changed only after the crash of 1987.Prior to that, the volatility negative skew was practically absent altogether.
Note that as the skew formula in bibliography shows, the volatility-of-volatility ξ affects the skew as much as the correlation ρ does. This explains why sudden upward movements can hurt us as well. If the overall correlation is negative but there are large movements in both directions, we will have large third (in absolute value) and fourth moments, which would make the options expectations correct. In fact, as we will see in the following example, a large upward movement can make us lose on our hedge account. The trade will generate moderate and consistent profits if no crash happens. But if the crash does happen we could suffer a large loss.

Skewness vs. Kurtosis

The skewness trade seems to be a simpler one and has a better chance to be realized. Indeed, in order to have a large negative skew, we need a large volatility-of-volatility ξ (as we do for the kurtosis trade) and a large negative correlation ρ. In other words, if for a given stock time series we have a large volatility-of-volatility but a weak correlation, we will not have a kurtosis trade opportunity but we will have a skewness trade opportunity. The historic skew will be small and the historic kurtosis high. Graphically, we could have the following interpretation. For these assets, the historic distribution does have fat tails, but remains symmetric, whereas the implied distribution has a fatter left tail. This is why we have a skewness trade opportunity, even if we do not have a kurtosis trade opportunity. Finally, as we previously mentioned, the estimation of the skewness from a time series is more reliable because it depends only on the product of the volatility of-volatility and the correlation.

An Exact Replication

These trading strategies can be refined using a Carr-Madan [29]replication.

$$E[f(S_T)] = f(F) + e^{rT}\int_0^F f''(K)P(S_0, K, t=0, T)dK + e^{rT}\int_0^F f''(K)C(S_0, K, t=0, T)dK \quad (8)$$

with $F = S_0 e^{rT}$ (9) the forward price.

In order to get the Das skew and kurtosis calculations, we need to take for the nth moment

$f(S_T) = (Z_T - E(Z_T))^n$ (10)

With

$$Z_T = \ln\left(\frac{S_T}{S_0}\right) \quad (11)$$

The Mirror Trades

Should we see the opposite conditions in the market, that is, having the skew (in absolute value) or kurtosis undervalued by the options given a historic path, we could obviously put on the mirror trades. The inverse of the peso theory would be as follows. The stock in question has already had a crash and the options are supposing there probably will not be another one in the near future. Setting up the overvalued kurtosis trade in the previous paragraph, we picked up a premium and made an immediate profit and hoped that there would not be a sudden movement. Here we start by losing money and hope a crash will happen within the life of the option so that we can generate a profit. Because jumps and crashes are rare by nature, this trade does not seem very attractive. Moreover, if there was a recent crash, the possibility of another one is indeed reduced and we should believe the options prediction. However, these mirror trades could be considered as buying insurance and therefore as a protection against a possible crash.

Other algorithmic trading strategies

A small number of execution strategies have become de facto standards and are offered by most technology providers, banks, and institutional broker/ dealers. However, even among these standards, the large number of input parameters makes it diffi cult to compare execution strategies directly. Typically, a strategy is motivated by a theme, or style of trading. The objective is to minimize either absolute or risk-adjusted costs relative to a benchmark. For strategies with mathematically defined objectives, an optimization is performed to determine how to best use the strategy to maximize a trader's or portfolio manager's utility. A trade schedule—or trajectory— is planned for strategies with a target quantity of shares to execute. The order placement engine—sometimes called the microtrader—translates from a strategy's broad objectives to individual orders. User defined input parameters control the trade schedule and order placement strategy.
In this section we review some of the most common algorithmic trading strategies.

Volume-Weighted Average Price

Six or seven years ago, the Volume Weighted Average Price (VWAP) execution strategy represented the bulk of algorithmic trading activity. Currently, it is second in popularity only to arrival price. The appeal of benchmarking to VWAP is that the benchmark is easy to compute and intuitively accessible. The typical parameters of a VWAP execution are the start time, the end time, and the number of shares to execute. Additionally, optimized forms of this strategy require a choice of risk aversion.

The most basic form of VWAP trading uses a model of the fractional daily volume pattern over the execution period. A trade schedule is calculated to match this volume pattern. For example, if the execution period is one day, and 20% of a day's volume is expected to be transacted in the first hour, a trader using this basic strategy would trade 20% of his target accumulation or liquidation in the first hour of the day. Since the daily volume pattern has a U shape—with more trading in the morning and afternoon and less in the middle of the day—the volume distribution of shares executed in a VWAP pattern would also have this U shape. VWAP is an ideal strategy for a trader who meets all of the following criteria: His trading has little or no alpha during the execution period. He is benchmarked against the volume weighted average price. He believes that market impact is minimized when his own rate of trading represents the smallest possible fraction of all trading activity. He has a set number of shares to buy or sell. Deviation from these criteria may make VWAP strategies less attractive. For example, market participants who trade over the course of a day and have strong positive alpha may prefer a front-weighted trajectory, such as those that are produced by an arrival price strategy. The period of a VWAP execution is most typically a day or a large fraction of a day. Basic VWAP models predict the daily volume pattern using a simple historical average of fractional volume. Several weeks to several months of data are commonly used. However, this forecast is noisy. On any given day, the actual volume pattern deviates substantially from its historical average, complicating the strategy's objective of minimizing its risk adjusted cost relative to the VWAP benchmark. Some models of fractional volume attempt to increase the accuracy of volume pattern prediction by making dynamic adjustments to the prediction based on observed trading results throughout the day. Several variations of the basic VWAP strategy are common. The ideal VWAP user (as defined previously) can lower his expected costs by increasing his exposure to risk relative to the VWAP benchmark. For example, assuming an alpha of zero, placing limit orders throughout the execution period and catching up to a target quantity with a market order at the end of the execution period will lower expected cost while increasing risk. This is the highest risk strategy. Continuously placing small market orders in the fractional volume pattern is the lowest risk strategy, but has a higher expected cost. For a particular choice of risk aversion, somewhere between the highest and lowest risk strategies, is a compromise optimal strategy that perfectly balances risk and costs.

For example, a risk-averse VWAP strategy might place one market order of 100 shares every 20 seconds while a less risk-averse strategy might place a limit order of 200 shares, and, 40 seconds later, place a market order for the difference between the desired fill of 200 and the actual fi ll (which may have been smaller). The choice of the average time between market orders in a VWAP execution implies a particular risk aversion. For market participants with a positive alpha, a frequently used rule-of thumb optimization is compressing trading into a shorter execution period. For example, a market participant may try to capture more profits by doing all of his VWAP trading in the first half of the day instead of taking the entire day to execute. In another variant of VWAP—guaranteed VWAP—a broker commits capital to guarantee his client the VWAP price in return for a predetermined fee. The broker takes on a risk that the difference between his execution and VWAP will be greater than the fee he collects. If institutional trading volume and individual stock returns were uncorrelated, the risk of guaranteed VWAP trading could be diversified away across many clients and many stocks. In practice, managing a guaranteed VWAP book requires some complex risk calculations that include modelling the correlations of institutional trading volume.

Time-Weighted Average Price

The Time-Weighted Average Price execution strategy (TWAP) attempts to minimize market impact costs by maintaining an approximately constant rate of trading over the execution period. With only a few parameters—start time, end time, and target quantity—TWAP has the advantage of being the simplest execution strategy to implement. As with VWAP, optimized forms of TWAP may require a choice of risk aversion. Typically, the VWAP or arrival price benchmarks are used to gauge the quality of a TWAP execution. TWAP is hardly ever used as its own benchmark. The most basic form of TWAP breaks a parent order into small child orders and executes these child orders at a constant rate. For example, a parent order of 300 shares with an execution period of 10 minutes could be divided into three child orders of 100 shares each. The child orders would An ideal TWAP user has almost the same characteristics as an ideal VWAP user, except that he believes that the lowest trading rate—not the lowest participation rate—incurs the lowest market impact costs. TWAP users can benefit from the same type of optimization as VWAP users by placing market orders less frequently, and using resting limit orders to attempt to improve execution quality.

Participation

The participation strategy attempts to maintain a constant fractional trading rate. That is, its own trading rate as a fraction of the market's total trading rate should be constant throughout the execution period. If the fractional trading rate is maintained exactly, participation strategies cannot guarantee a target fill quantity. The parameters of a participation strategy are the start time, end

time, fraction of market volume the strategy should represent, and max number of shares to execute. If the max number of shares is specified, the strategy may complete execution before the end time. Along with VWAP and TWAP, participation is a popular form of non optimized strategies, though some improvements are possible with optimization.

VWAP and arrival price benchmarks are often used to gauge the quality of a participation strategy execution. The VWAP benchmark is particularly appropriate because the volume pattern of a perfectly executed participation strategy is the market's volume pattern during the period of execution. An ideal user of participation strategies has all of the same characteristics as an ideal user of VWAP strategies, except that he is willing to forego certain execution to maintain the lowest possible fractional participation rate. Participation strategies do not use a trade schedule. The strategy's objective is to participate in volume as it arises. Without a trade schedule, a participation strategy can't guarantee a target fill quantity. The most basic form of participation strategies waits for trading volume to show up on the tape, and follows this volume with market orders. For example, if the target fractional participation rate is 10%, and an execution of 10,000 shares is shown to have been transacted by other market participants, a participation strategy would execute 1,000 shares in response. Unlike a VWAP trading strategy, which for a given execution may experience large deviations from an execution period's actual volume pattern, participation strategies can closely track the actual—as opposed to the predicted— volume pattern. However, close tracking has a price. In the preceding example, placing a market order of 1,000 shares has a larger expected market impact than slowly following the market's trading volume with smaller orders. An optimized form of the participation strategy amortizes be executed at the 3:20, 6:40, and 10:00 minute marks. Between market the trading shortfall over some period of time. Specifically, if an execution of 10,000 shares is shown to have been transacted by other market participants, instead of placing 1,000 shares all at once, a 10% participation strategy might place 100 share orders over some period of time to amortize the shortfall of 1,000 shares. The result is a lower expected shortfall, but a higher dispersion of shortfalls.

Market-on-Close

The market-on-close strategy is popular with market participants who either want to minimize risk-adjusted costs relative to the closing price of the day, or want to manipulate—game—the close to create the perception of a good execution. The ideal market-on-close user is benchmarked to the close of the day and has low or negative alpha. The parameters of a market-on-close execution are the start time, the end time, and the number of shares to execute.

Optimized forms of this strategy require a risk-aversion parameter. When market-on-close is used as an optimized strategy, it is similar in its formulation to an arrival price strategy. However, with market-on-close, a back-weighted trade schedule incurs less risk than a front-weighted one. With arrival price, an infinitely risk averse trader would execute everything in the opening seconds of the execution period. With market-on-close, an infinitely risk averse trader would execute

everything at the closing seconds of the day. For typical levels of risk aversion, some trading would take place throughout the execution period. As with arrival price optimization, positive alpha increases urgency to trade and negative alpha encourages delayed execution. In the past, market-on-close strategies were used to manipulate—or game—the close, but this has become less popular as the use of VWAP and arrival price benchmarks have increased. Gaming the close is achieved by executing rapidly near the close of the day. The trade print becomes the closing price or very close to it, and hence shows little or no shortfall from the closing price benchmark. The true cost of the execution is hidden until the next day when temporary impact dissipates and prices return to a new equilibrium.

Arrival Price

The arrival price strategy—also called the implementation shortfall strategy— attempts to minimize risk-adjusted costs using the arrival price benchmark. Arrival price optimization is the most sophisticated and popular of the commonly used algorithmic trading strategies.
The ideal user of arrival price strategies has the following characteristics.
Orders, the strategy may place limit orders in an attempt to improve execution

- He is benchmarked to the arrival price.
- He is risk averse and knows his risk-aversion parameter.
- He has high positive or high negative alpha.
- He believes that market impact is minimized by maintaining a constant rate of trading over the maximum execution period while keeping trade size small.

Most implementations are based on some form of the risk-adjusted cost minimization introduced by Almgren and Chriss[30] that we discussed earlier. In the most general terms, an arrival price strategy evaluates a series of trade schedules to determine which one minimizes risk-adjusted costs relative to the arrival price benchmark. As discussed in the section on optimal execution, under certain assumptions, this problem has a closed form solution. The parameters in an arrival price optimization are alpha, number of shares to execute, start time, end time, and a risk aversion parameter. For buyers (sellers) positive (negative) alpha encourages faster trading. For both buyers and sellers, risk encourages faster trading, while market impact costs encourage slower trading. For traders with positive alpha, the feasible region of trade schedules lies between the immediate execution of total target quantity and a constant rate of trading throughout the execution period. A more general form of arrival price optimization allows for both buyers and sellers to have either positive or negative alpha. For example, under the assumption of negative alpha, shares held long and scheduled for liquidation are—without considering one's own trading—expected to go up in price over the execution period. This would encourage a trader to delay execution or stretch out trading. Hence, the feasible region of solutions that account for both positive and negative alpha includes back-weighted as well as front-weighted trade schedules.

Other factors that necessitate back-weighted trade schedules in an arrival price optimization are expected changes in liquidity and expected crossing opportunities. For example, an expectation of a cross later in the execution period may provide enough cost savings to warrant taking on some price risk and the possibility of a compressed execution if the cross fails to materialize. Similarly, if market impact costs are expected to be lower later in the execution period, a rational trader may take on some risk to obtain this cost savings.

A variant of the basic arrival price strategy is adaptive arrival price. A favorable execution may result in a windfall in which an accumulation of a large number of shares takes place at a price significantly below the arrival quality. price. This can happen by random chance alone. Almgren and Lorenz[31] demonstrated that a risk-averse trader should use some of this windfall to reduce the risk of the remaining shares. He does this by trading faster and thus incurring a higher market impact. Hence, the strategy is adaptive in that it changes its behavior based on how well it is performing.

Crossing

Though crossing networks have been around for some time, their use in algorithmic trading strategies is a relatively recent development. The idea behind crossing networks is that large limit orders—the kind of orders that may be placed by large institutional traders—are not adequately protected in a public exchange. Simply displaying large limit orders in the open book of an electronic exchange may leak too much information about institutional traders' intentions. This information is used by prospective counter-parties to trade more passively in the expectation that time constraints will force traders to replace some or all of large limit orders with market orders. In other words, information leakage encourages gaming of large limit orders. Crossing networks are designed to limit information leakage by making their limit books opaque to both their clients and the general public.

A popular form of cross is the mid-quote cross, in which two counterparties obtain a mid-quote fill price. The mid-quote is obtained from a reference exchange, such as the NYSE or other public exchange. Regulations require that the trade is then printed to a public exchange to alert other market participants that it has taken place. The cross has no market impact but both counterparties pay a fee to the crossing network. These fees are typically higher than the fees for other types of algorithmic trading because the market impact savings are significant while the fee is contingent on a successful cross. More recently, crossing networks have offered their clients the ability to place limit orders in the crossing networks' dark books. Placing a limit order in a crossing network allows a cross to occur only at a certain price. This makes crossing networks much more like traditional exchanges, with the important difference that their books are opaque to market participants.

To protect their clients from price manipulation, crossing networks implement antigaming logic. As previously explained, opaqueness is itself a form of antigaming, but there are other strategies.

For example, some crossing networks require orders above a minimum size, or orders that will remain in the network longer than a minimum time commitment. Other networks will cross only orders of similar size. This prevents traders from pinging—sending small orders to the network to determine which side of the network's book has an order imbalance. Another approach to antigaming prevents crosses from taking place during periods of unusual market activity. The assumption is that some of this unusual activity is caused by traders trying to manipulate the spread in the open markets to get a better fill in a crossing network. Some networks also attempt to limit participation by active traders, monitoring their clients' activities to see if their behavior is more consistent with normal trading than with gaming. There are several different kinds of crossing networks. A continuous crossing network constantly sweeps through its book in an attempt to match buy orders with sell orders. A discrete crossing network specifies points in time when a cross will take place, say every half hour. This allows market participants to queue up in the crossing network just prior to a cross instead of committing resting orders to the network for extended periods of time. Some crossing networks allow scraping—a one-time sweep to see if a single order can find a counterparty in the crossing network's book—while others allow only resting orders. In automated crossing networks, resting orders are matched according to a set of rules, without direct interaction between the counterparties. In negotiated crossing networks, the counterparties first exchange indications of interest, then negotiate price and size via tools provided by the system.

Some traditional exchanges now allow the use of invisible orders, resting orders that sit in their order books but are not visible to market participants. These orders are also referred to as dark liquidity. The difference between these orders and those placed in a crossing network is that traditional exchanges offer no special antigaming protection.

Private dark pools are collections of orders that are not directly available to the public. For example, a bank or pension manager might have enough order flow to maintain an internal order book that, under special circumstances is exposed to external scraping by a crossing network or crossing aggregator.

A crossing aggregator charges a fee for managing a single large order across multiple crossing networks. Order placement and antigaming rules differ across networks, making this task fairly complex. A crossing aggregator may also use information about historical and real-time fi lls to direct orders. For example, failure to fill a small resting buy order in a crossing network may betray information of a much larger imbalance in the network's book. This makes the network a more attractive destination for future sell orders. In general, the management of information across crossing networks should give crossing aggregators higher fill rates than exposure to any individual network.

Crossing lends itself to several optimization strategies. Longer exposure to a crossing network increases the chances of an impact-free fill, but also increases the risk of a large and compressed execution if an order fails to obtain a fill. Finding an optimal exposure time is one type of crossing optimization. A more sophisticated version of this approach is solving for a trade-out, a schedule

for trading shares out of the crossing network into the open markets. As time passes and a cross is not obtained, the strategy mitigates the risk of a large, compressed execution by slowly trading parts of the order into the open markets.

Other Algorithms

Two other algorithms are typically included in the mix of standard algorithmic trading offerings. The first is liquidity seeking where the objective is to soak up available liquidity. As the order book is depleted, trading slows. As the order book is replenished, trading speeds up.
The second algorithm is financed trading. The idea behind this strategy is to use a sale to finance the purchase of a buy with the objective of obtaining some form of hedge. This problem has all of the components of a full optimization. For example, if, after a sell, a buy is executed too quickly, it will obtain a less favourable fill price. On the other hand, executing a buy leg too slowly increases the tracking error between the two components of the hedge and increases the dispersion of costs required to complete the hedge.

2.3. Statistics and Finance: Econophysics and behavioral finance

2.3.1 Statistics, econophysics and behavioural finance

Statistical finance, sometimes called econophysics, is an empirical attempt to shift finance from its normative roots to a positivist framework using exemplars from statistical physics with an emphasis on emergent or collective properties of financial markets. The starting point for this approach to understanding financial markets are the empirically observed stylized facts (Bouchaud and Potters 2003)[32] .

Stylized facts

Macrostructure
1. Stock markets are characterised by bursts of price volatility.
2. Price changes are less volatile in bull markets and more volatile in bear markets.
3. Price change correlations are stronger with higher volatility, and their auto-correlations die out quickly.
4. Almost all real data have more extreme events than suspected.
5. Volatility correlations decay slowly.
6. Trading volumes have memory the same way that volatilities do.
7. Past price changes are negatively correlated with future volatilities.
Stock Individual and collective stock dynamics: intra-day seasonalities

Stylized facts concerning the intra-day seasonalities of stock dynamics.

1. U-shaped pattern of the volatility
2. Average correlation between stocks increases throughout the day, leading to a smaller relative dispersion between stocks. Somewhat paradoxically, the kurtosis (a measure of volatility surprises) reaches a minimum at the open of the market, when the volatility is at its peak.
3. Dispersion kurtosis is a markedly decreasing function of the index return. This means that during large market swings, the idiosyncratic component of the stock dynamics becomes sub-dominant.
4. In a nutshell, early hours of trading are dominated by idiosyncratic or sector specific effects with little surprises, whereas the influence of the market factor increases throughout the day, and surprises become more frequent.

Research objectives

Statistical finance is focused on three areas:

1. Empirical studies focused on the discovery of interesting statistical features of financial time-series data aimed at extending and consolidating the known stylized facts.

2. The use of these discoveries to build and implement models that better price derivatives and anticipate stock price movement with an emphasis on non-Gaussian methods and models.

3. The study of collective and emergent behaviour in simulated and real markets to uncover the mechanisms responsible for the observed stylized facts with an emphasis on agent based models (see agent-based model).

Behavioral finance and statistical finance

Behavioural finance attempts to explain price anomalies in terms of the biased behaviour of individuals, mostly concerned with the agents themselves and to a lesser degree aggregation of agent behaviour. Statistical finance is concerned with emergent properties arising from systems with many interacting agents and as such attempts to explain price anomalies in terms of the collective behaviour. Emergent properties are largely independent of the uniqueness of individual agents because they are dependent on the nature of the interactions of the agents rather that the agents themselves. This approach has draw strongly on ideas arising from complex systems, phase transitions, criticality, self-organized criticality, non-extensivity (see Tsallis entropy), q-Gaussian models, and agents based models (see agent based model); as these are known to be able to recover some of phenomenology of financial market data, the stylized facts, in particular the long-range memory and scaling due to long-range interactions.

Criticism

Within the subject the description of financial markets blindly in terms of models of statistical physics has been argued as flawed because it has transpired these do not fully correspond to what we now know about real finance markets. First, traders create largely noise, not long range correlations among themselves. A market is not at an equilibrium critical point, the resulting non-

equilibrium market must reflect details of traders' interactions (universality applies only to a limited very class of bifurcations, and the market does not sit at a bifurcation). Even if the notion of a thermodynamics equilibrium is considered not at the level of the agents but in terms of collections of instruments stable configurations are not observed. The market does not 'self-organize' into a stable statistical equilibrium, rather, markets are unstable. Although markets could be 'self-organizing' in the sense used by finite-time singularity models these models are difficult to falsify. Although Complex systems have never been defined in a broad sense; financial markets do satisfy reasonable criterion of being considered complex adaptive systems The Tallis doctrine has been put into question as it is apparently a special case of markov dynamics so questioning the very notion of a "non-linear Fokker-Plank equation". In addition, the standard 'stylized facts' of financial markets, fat tails, scaling, and universality are not observed in real FX markets even if they are observed in equity markets.

From outside the subject the approach has been considered by many as a dangerous view of finance which has drawn criticism from some economists because of:

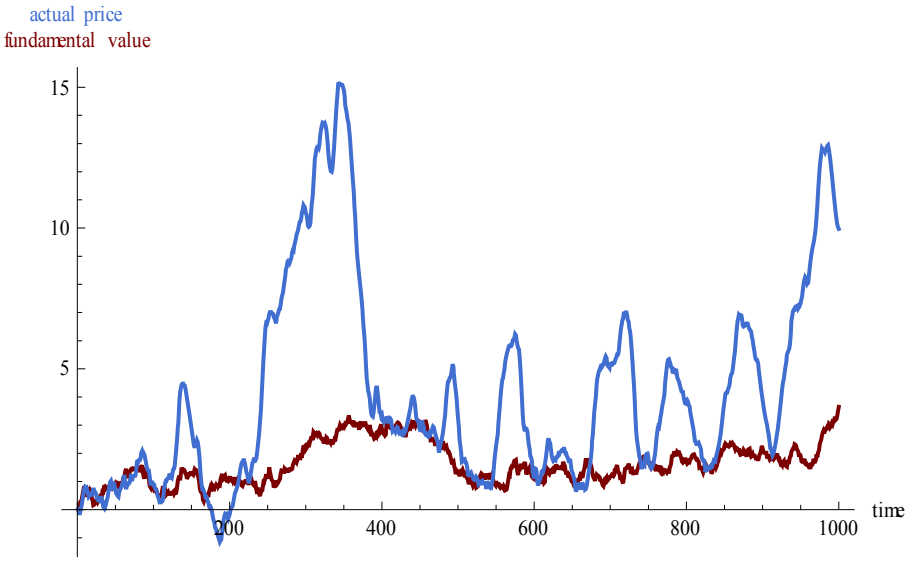
1. "A lack of awareness of work which has been done within economics itself."
2. "Resistance to more rigorous and robust statistical methodology."
3. "The belief that universal empirical regularities can be found in many areas of economic activity."
4. "The theoretical models which are being used to explain empirical phenomena."

In response to this criticism there are claims of a general maturing of these non-traditional empirical approaches to Finance. This defense of the subject does not flatter the use of physics metaphors but does defend the alternative empirical approach of "econophysics" itself.

Some of the key data claims have been questioned in terms of methods of data analysis. Some of the ideas arising from nonlinear sciences and statistical physics have been helpful in shifting our understanding financial markets, and may yet be found useful, but the particular requirements of stochastic analysis to the specific models useful in finance is apparently unique to finance as a subject. There is much lacking in this approach to finance yet it would appear that the canonical approach to finance based optimization of individual behaviour given information and preferences with assumptions to allow aggregation in equilibrium are even more problematic.

It has been suggested that what is required is a change in mindset within finance and economics that moves the field towards methods of natural science. Perhaps finance needs to be thought of more as an observational science where markets are observed in the same way as the observable universe in cosmology, or the observable ecosystems in the environmental sciences. Here local principles can be uncovered by local experiments but meaningful global experiments are difficult to envision as feasible without reproducing the system being observed. The required science becomes that based largely on pluralism (see scientific pluralism), as in most sciences that deal with complexity, rather than a singled unified mathematical framework that is to be verified by experiment.

Figure 3: Bubbles in a Simple Behavioral Finance Model



Suppose participants in a financial market adopt either a "fundamentalist" or "chartist" trading strategy based on the historical, risk-adjusted profitability of each strategy. Fundamentalists bet that price will adjust (at a speed determined by the "speed-of-adjustment parameter") towards a value justified by economic fundamentals (the "fundamental value", assumed to follow a random walk). Chartists bet that price will follow its historical trend, such that a percentage (determined by the "extrapolation parameter") of the previous period's price movement will occur again. This Demonstration shows how deviations of price from fundamental value ("bubbles") are affected by the parameters of the trading strategy and the stochastic nature of the model.

2.3.2 Agent-Based Modelling of Financial Markets

Agent-based modelling has become a popular methodology in social sciences, particularly in economics. Here we focus on the agent-based modelling of financial markets. The very idea of describing markets with models of interacting agents (traders, investors) does not fit well with the classical financial theory that is based on the notions of efficient markets and rational investors. However, it has become obvious that investors are neither perfectly rational nor have homogeneous expectations of the market trends. Agent-based modelling proves to be a flexible framework for a realistic description of the investor adaptation and decision-making process.

The paradigm of agent-based modelling applied to financial markets implies that trader actions determine price. This concept is similar to that of statistical physics within which the thermodynamic (macroscopic) properties of the medium are described via molecular interactions. A noted expansion of the microscopic modelling methodology into social systems is the minority game[33]. Its development was inspired by the famous El Farol's bar problem[34]. The minority game is a simple binary choice problem in which players have to choose between two sides, and those on the minority side win. Similarly to the El Farol's bar problem, in the minority game there is no communication among players and only a given set of forecasting strategies defines player

decisions. The minority game is an interesting stylized model that may have some financial implications.

But we shall focus further on the models derived specifically for describing financial markets. In the known literature, early work on the agent-based modelling of financial markets can be traced back to 1980. In this paper, Beja and Goldman[35] considered two major trading strategies, value investing and trend following. In particular, they showed that system equilibrium may become unstable when the number of trend followers grows. Since then, many agent-based models of financial markets have been developed. We divide these models into two major groups. In the first group, agents make decisions based on their own predictions of future prices and adapt their beliefs using different predictor functions of past returns. The principal feature of this group is that price is derived from the supply-demand equilibrium. Therefore, we call this group the adaptive equilibrium models. In the other group, the assumption of the equilibrium price is not employed. Instead, price is assumed to be a dynamic variable determined via its empirical relation to the excess demand. We call this group the nonequilibrium price models.

The author of this PhD thesis researched with Jason Cawley in a model called New Kind of Science Artificial Market Model[36]. The model can reproduce stylized facts of real markets that are absent from lognormal random walks. The internal mechanisms creating those stylized facts are at least plausible. These may be taken as evidence that non - uniform trading strategies and their different feedback effects on prices may be operative and important in real markets.

2.3.3 Game theory

Game theory is a branch of applied mathematics that is used in the social sciences, most notably in economics, as well as in biology (most notably evolutionary biology and ecology), engineering, political science, international relations, computer science, and philosophy. Game theory attempts to mathematically capture behavior in strategic situations, in which an individual's success in making choices depends on the choices of others. While initially developed to analyze competitions in which one individual does better at another's expense (zero sum games), it has been expanded to treat a wide class of interactions, which are classified according to several criteria. Today, "game theory is a sort of umbrella or 'unified field' theory for the rational side of social science, where 'social' is interpreted broadly, to include human as well as non-human players (computers, animals, plants)" (Aumann 1987)[37].

Traditional applications of game theory attempt to find equilibrium in these games. In an equilibrium, each player of the game has adopted a strategy that they are unlikely to change. Many equilibrium concepts have been developed (most famously the Nash equilibrium) in an attempt to capture this idea. These equilibrium concepts are motivated differently depending on the field of application, although they often overlap or coincide.

This methodology and the appropriateness of particular equilibrium concepts remain controversial as some other research shows, so further proof is needed to accept it as general.

Although some advances were made before it, the field of game theory came into being with the 1944 book Theory of Games and Economic Behavior by John von Neumann and Oskar Morgenstern[38]. This theory was developed extensively in the 1950s by many scholars. Game theory was later explicitly applied to biology in the 1970s, although similar developments go back at least as far as the 1930s. Game theory has been widely recognized as an important tool in many fields. Eight game theorists have won the Nobel Memorial Prize in Economic Sciences, and John Maynard Smith was awarded the Crafoord Prize for his application of game theory to biology.

2.3.4 Microstructure – Are the dynamics of financial markets endogenous or exogenous ?

There is important evidence that the erratic dynamics of markets is to a large extent of endogenous origin, i.e. determined by the trading activity itself and not due to the rational processing of exogenous news. In order to understand why and how prices move, the joint fluctuations of order flow and liquidity – and the way these impact prices – become the key ingredients.

Impact is necessary for private information to be reflected in prices, but by the same token, random fluctuations in order flow necessarily contribute to the volatility of markets. Our thesis is that the latter contribution is in fact dominant, resulting in a decoupling between prices and fundamental values, at least on short to medium time scales. We argue that markets operate in a regime of vanishing revealed liquidity, but large latent liquidity, which would explain their hyper-sensitivity to fluctuations.

More precisely, we can identify a dangerous feedback loop between bid-ask spread and volatility that may lead to micro-liquidity crises and price jumps. It exist several other unstable feedback loops that should be relevant to account for market crises: imitation, unwarranted quantitative models, pro-cyclical regulation, etc.

News seem to play a minor role in market volatility; most jumps appear to be unrelated to news, but seem to appear spontaneously as a result of the market activity itself; the stylised facts of price statistics (fat-tails in the distribution of returns, long-memory of the volatility) are to a large extent universal, independent of the particular nature of the traded asset, and very reminiscent of endogenous noise in other complex systems (turbulence, Barkhausen noise, earthquakes, fractures, etc.). In all these examples, the intermittent, avalanche nature of the dynamics is an emergent property, unrelated to the exogenous drive which is slow and regular.

In search of a purely endogenous interpretation of these effects, it is natural to investigate to high-frequency, microstructure ingredients that generate price changes. We have discussed the remarkable long-range correlations in order flow that has far-reaching consequences and forces us to revise many preconceived ideas about equilibrium.

First of all, these correlations reflect the fact that even “liquid” markets are in fact very illiquid, in the sense that the total volume in the order book available for an immediate transaction is extremely small (10^{-5} of the market capitalisation for stocks). The immediate consequence is

that the trades of medium to large institutions can only be executed incrementally, explaining the observed correlation in the order flow. By the same token, the information motivating these trades (if any) cannot be instantaneously reflected by prices. Prices cannot be in equilibrium, but randomly evolve as the icebergs of latent supply and demand progressively reveal themselves (and possibly evolve with time).

This feature is an unavoidable consequence of the fact that sellers and buyers must hide their intentions, while liquidity providers only post small volumes in fear of adverse selection.

The observation that markets operate in a regime of vanishing revealed liquidity, but large latent liquidity is crucial to understand their hyper-sensitivity to fluctuations, potentially leading to instabilities. Liquidity is necessarily a dynamical phenomenon that reacts to order flow such as to dampen the trending effects and keep price returns unpredictable, through the subtle 'tug-of-war' equilibrium mentioned above. Such a dynamical equilibrium can however easily break down. For example, an upward fluctuation in buy order flow might trigger a momentary panic, with the opposing side failing to respond immediately. Similarly, the strong structural link between spread and volatility can ignite a positive feedback loop whereby increased spreads generate increased volatility, which itself causes liquidity providers to cancel their orders and widen the spread. Natural fluctuations in the order flow therefore lead, in some cases, to a momentary lapse of liquidity, explaining the frequent occurrence of price jumps without news.

An extreme incarnation of this feedback loop probably took place during the "flash crash" of May 6th, 2010. We believe that the formal limit of zero liquidity is a critical point [39], which would naturally explain the analogy between the dynamics of markets and that of other complex systems, in particular the universal tails and the intermittent bursts of activity. We are however lacking a precise model that would allow one to formalise these ideas (see [40], [41] for work in that direction).

In summary, the picture of markets we advocate is such that the lion's share of high frequency dynamics is due to fluctuations in order flow. News and information about fundamental values only play the role of "stirring" the system, i.e. slowly changing the large latent supply and demand, except on relatively rare occasions where these events do indeed lead to violent shocks. Most of the market activity comes from the slow execution of these large latent orders, that cascades into high frequency fluctuations under the action of the use of liquidity providers and liquidity takers, who compete to exploit all statistical regularities.

The end product of this activity is a white noise signal. Prices are, in a first approximation, statistically efficient in the sense that there is little predictability left in the time series. But this does not necessarily mean that these prices reflect in any way some true underlying information about assets. We believe, as Keynes and Black did, that the uncertainty in fundamental values is so large that there is no force to anchor the price against random high frequency perturbations. It is quite remarkable indeed that the high frequency value of the volatility approximately coincides with the volatility on the scale of weeks, showing that there is very little mean-reverting effects to rein the high frequency tremor of markets. Only when prices reach values that are – say – a

factor 2 away from their "fundamental value" will mean-reverting effects progressively come into play. In the context of stocks, this only happens on the scale of months to years, see [42] and the discussion in [43]. From this point of view, as emphasised by Lyons [44], "micro-structure implications may be long-lived" and "are relevant to macroeconomics". For a review read also Hasbrouck[45].

2.3.5 Endogenous-Exogenous Market model

Using the Bouchaud [46]framework we introduce a price at time t that can be decomposed on the exogenous and endogenous past impacts:

$$p_t = p_{-\infty} + \sum_{t'=-\infty}^{t-1} G(t-t')\varepsilon_{t'} S_{t'} V_{t'}^{\psi} + \sum_{t'=-\infty}^{t-1} Q(t-t') EA_{t'}^{\eta} + \sum_{t'=-\infty}^{t-1} C(t-t') ON_{t'}^{\alpha} \quad (12)$$

where

- $G(l)$ is trading impact propagation function
- $\varepsilon_{t'}$ sign on the trade
- $S_{t'}$ spread at time t
- $V_{t'}^{\psi}$ volume of the trade
- $Q(l)$ is earnings announcement propagation function
- $EA_{t'}^{\eta}$ Earnings announcements
- $C(l)$ Other news propagation function
- $ON_{t'}^{\alpha}$ Other news

2.3.6 Statistics and Finance

Since the financial markets evolve apparently in a random fashion, the natural object to model its evolution are stochastic processes. Stochastic processes are thus essential in risk management, portfolio management and trade optimization to model the evolution of the risk factors that determine the price of a set of securities at a given point in the future. We denote by $\boldsymbol{X}_t$ the value of one such factor $\boldsymbol{X}$ at a generic time t, i.e. the stochastic process for $\boldsymbol{X}$. This stochastic process is fully described by a multivariate function, the process cumulative distribution function, i.e. the joint probability of its realizations at any set of times

$$\boldsymbol{F}_{t_1,t_2,\ldots}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots) \equiv \mathrm{P}\left\{\boldsymbol{X}_{t_1} \le \boldsymbol{x}_1, \boldsymbol{X}_{t_2} \le \boldsymbol{x}_2, \ldots\right\} \quad (13)$$

For the above mentioned purposes of risk management and portfolio management, the monitoring times t1, t2, . . . are typically chosen on a discrete, equally spaced, grid t, t+Δ, t+2Δ, . ..: for instance Δ can be one day, or one week, or one month, etc.

The probability measure "P" is the "real" probability that governs the evolution of the process. In reality, such probability is not known and needs to be estimated: indeed, estimation is the main concern of the real-measure P-world.

P : estimate the future

There exists a parallel, much developed, area of application of stochastic processes in finance: derivatives pricing. According to the fundamental theorem of pricing, the normalized price $\widetilde{P}_t$ of a security is arbitrage-free only if there exists a fictitious stochastic process describing its future evolution such that its expected value is always constant and equals the current price:

$$\widetilde{P}_t = \mathrm{E}\left\{\widetilde{P}_{t+\tau}\right\},\ \tau \geq 0 \quad (14)$$

A process satisfying (14) a martingale: since a martingale does not reward risk, the fictitious probability of the pricing process is called risk-neutral and is typically denoted by "Q". Since some prices are liquid and readily observable,(14) becomes a powerful tool to assign a fair price to a non-liquid security,once a fictitious martingale has been calibrated to other traded securities. In other words, derivatives pricing is a very high-tech interpolation exercise which lives in the risk neutral Q-world

Q : interpolate the present

Notice that martingales represent a very restrictive class of processes. A tractable and much more general class of processes are the so-called semimartingales, on which a theory of integration can be defined that makes sense for financial modelling. However, for risk and portfolio management, we do not need to restrict our attention to semimartingales: any process that suitably models the behaviour of financial time series is a-priori admissible.

To summarize, risk management, portfolio management and trade optimization rely on discrete-time processes that live in the real-world probability measure P. Derivatives pricing relies on continuous-time processes that live in a fictitious risk-neutral probability measure Q. Interactions between the P-world and the Q-world occur frequently. For instance, the sensitivities of a security to a set of risk factors, the so-called Greeks, are computed using Q models, but then they are applied in the P-world for hedging purposes. Similarly, the evolution through time of the Q parameters of a given pricing model can be used to predict the P-world distribution of the same prices in the future and generate such statistics as the value at risk. Here we present an overview of the main processes in finance, highlighting the relationships between the P-world, discrete-time models and their Q-world, continuous-time counterparts. We start in Chapter 3 with the P-world discrete-time processes. In Section 3.1 we introduce the random walk, which is the cumulative sum of invariants. This process represents the benchmark for buy-side modelling: often

considered too simplistic by Q quants and P econometricians not involved in risk and portfolio management, the random walk proves to be very hard to outperform when estimation error is accounted for. In Section 3.2 we relax the hypothesis that the increments of the process be invariant: we introduce autocorrelation, thereby obtaining ARMA processes. We will account in Section 3.3 also for the empirical observations that at times autocorrelation, when present, decays very slowly with the number of lags: this behavior is suitably modeled by long memory-processes. In Section 3.4 we discuss volatility clustering: the scatter of the process, rather than the process itself, displays autocorrelation: GARCH and generalizations thereof capture this feature.

In Chapter 4 we discuss the Q-world continuous-time counterparts of the above P-world models. In Section 4.1 we present Levy processes, the continuous-time version of the random walk. In Section 4.2 we model autocorrelation with the Ornstein-Uhlenbeck and related processes. In Section 5.3 we model long memory by means of the fractional Brownian motion. In Section 4.4 we tackle volatility clustering in two flavours: stochastic volatility and subordination. We summarize in the table below the processes covered and their key characteristics

Table 3 Finance Processes

	Discrete time (P)	**Continuous time (Q)**
Base case	Random walk	Levy processes
Autocorrelation	ARMA	Ornstein-Uhlenbeck
Long memory	Fractional integration	Fractional Brownian motion
Volatility Clustering	GARCH	Stochastic Volatility Subordination

We analyze also markov switching models 4.5 and fractals and multifractals 4.6 as interesting models that have received a lot of interest lately by econometrics and econophysics.

An important concept is that of a stationary time series. A series is stationary in the “strict sense” if all finite dimensional distributions are invariant with respect to a time shift. A series is stationary in a “weaker sense” if only the moments up to a given order are invariant with respect to a time shift. In this chapter, time series will be considered (weakly) stationary if the first two moments are time-independent. Note that a stationary series cannot have a starting point but must extend over the entire infinite time axis. Note also that a series can be strictly stationary (that is, have all distributions time-independent, but the moments might not exist). Thus a strictly stationary series is not necessarily weakly stationary.

3. P: discrete-time processes

We assume that we monitor a financial process at discrete, equally spaced time intervals. Without loss of generality, we can assume this interval to have unit length: therefore, denoting time by t, we assume that $t \in \mathrm{Z}$.

3.1. Random walk

The random walk with drift is a process that cumulates invariants

$$\mathrm{X}_{\mathrm{t+1}} = \mathrm{X}_{\mathrm{t}} + \varepsilon_{\mathrm{t+1}} \quad (16)$$

The invariants ε_t, are independent and identically distributed (i.i.d.) shocks with any distribution. Since the distribution of the invariants is fully general, the random walk has a much richer structure than one would expect. Two very simple tests to spot invariance in a realized time series are discussed in the references.

First, due to the Glivenko-Cantelli theorem, if ε_t is an invariant the histogram of its time series represents its probability density function (pdf). Therefore, also any portion of its time series represents the pdf. As a consequence, the histogram of the first half of the time series of ε_t and that of the second half should look alike, see the top portion of Figure 1 and the discussion in the examples below.

Second, if ε_t is an invariant, then ε_t and ε_{t+1} are independent and identically distributed. Therefore the mean-covariance ellipsoid of ε_t and ε_{t+1} must be a circle. A scatter plot of ε_t versus ε_{t+1} with the plot of the mean-covariance ellipsoid should convey this impression, see the bottom portion of Figure 4 and the discussion in the examples below.

Consider a security that trades at time t at the price $\boldsymbol{P}_t$ and consider the compounded return

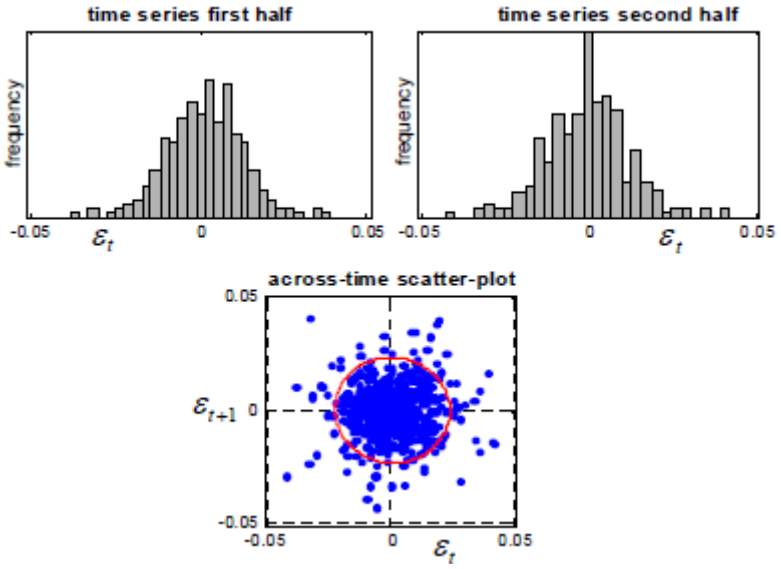
$$C_t \equiv \ln\left(\frac{P_t}{P_{t-1}}\right) \quad (17)$$

In the equity world the benchmark assumption is that the compounded returns are invariants, i.e. $\boldsymbol{C}_t \equiv \varepsilon_t$. This amounts to saying that $\boldsymbol{X}_t \equiv \ln \boldsymbol{P}_t$ evolves according to a random walk:

$$\ln\left(P_{t-1}\right) \equiv \ln\left(P_t\right) + \varepsilon_{t+1} \quad (18)$$

The plots in Figure 4 refer to this case.

Figure 4: Invariance check



As another standard example, denote by $\boldsymbol{Z}_t^{(E)}$ the price at time t of a zero coupon bond that matures at time E. As discussed in Meucci (2005)[47], the raw bond price cannot give rise to an invariant, because the approaching maturity dates breaks the time translation invariance of the analysis. Therefore, one introduces the yield to maturity, which is the annualized compounded return over the life of the bond

$$Y_t^{(v)} \equiv -\frac{1}{v}\ln\left(Z_t^{(t+v)}\right) \quad (19)$$

The time series of the yield to maturity, as opposed to the yield of maturity, is invariant under time translations and therefore we can hope to use it to generate invariants. Notice however that this time series does not correspond to the price of one single bond throughout, but rather to the price of a different bond at each time. In the fixed-income world the benchmark assumption is that the changes in yield to maturity are invariants

$$Y_{t+1}^{(v)} = Y_t^{(v)} + \varepsilon_{t+1}^{(v)} \quad (20)$$

which is clearly in the format (16). One can check this assumption with the simple test described in Figure 5.

3.1.1 Continuous invariants

As a base case, the invariants can have a normal distribution

$$\varepsilon_t \sim N\left(\mu, \sigma^2\right) \quad (21)$$

where μ is the expectation and σ^2 is the variance. Independent normal distributions are closed under the sum. With slight abuse of notation we write:

$$\left(\mu_1, {\sigma_1}^2\right) + \left(\mu_2, {\sigma_2}^2\right) \overset{d}{=} \left(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2\right) \quad (22)$$

Since the sum of independent normal variables is normal, from normal invariants follows a normally distributed random walk, i.e. a Brownian motion, see Section 5.1.

Different continuous distributions for the invariants give rise to different distributions for the random walk (16). For instance, one can model the invariants using stable distributions, Student t distributions, skewed distributions, etc.

3.1.2 Discrete invariants

The invariants can also take on a set $\boldsymbol{a} \equiv \{\boldsymbol{a}_0, ..., \boldsymbol{a}_K\}$ of discrete values with respective probabilities $\boldsymbol{p} \equiv \{\boldsymbol{p}_0, ..., \boldsymbol{p}_K\}$ that sum to one. This is the generalized Bernoulli distribution:

$$\varepsilon_t \sim Be(\mathbf{p}, \mathbf{a}) \quad (23)$$

If the invariant has a generalized Bernoulli distribution, the ensuing random walk also has a generalized Bernoulli distribution with different parameters at each time step.

When the number of values in a is large, a more parsimonious parameterization becomes necessary. An important case is the Poisson distribution:

$$\varepsilon_t \sim Po(\lambda, \Delta) \quad (24)$$

This represents a special case of (23) where the variable takes values on an infinite, equally spaced grid $\boldsymbol{a}_K \equiv \kappa\Delta$, also denoted as $\Delta\mathrm{N}$, with probabilities

$$p_k \equiv P\{\varepsilon_t = k\Delta\} \equiv \frac{\lambda^k e^{-\lambda}}{k!} \quad (25)$$

It follows that λ represents both the expectation and the variance of the invariant (24), as one can prove using the interchangeability of summation and derivation in the power series expansion of the exponential function.

Independent Poisson distributions are closed under the sum. With slight abuse of notation we write:

$$Po(\lambda_1; \Delta) + Po(\lambda_2; \Delta) \overset{d}{=} Po(\lambda_1 + \lambda_2; \Delta) \quad (26)$$

Therefore the random walk (16) ensuing from a Poisson invariant is Poisson distributed on the same grid.

3.1.3 Generalized representations

Since any function of invariants is an invariant, we can create invariants by aggregation.

Stochastic volatility

A flexible formulation in this direction is the stochastic volatility formulation:

$$\varepsilon_t \overset{d}{=} \mu_t + \sigma_t Z_t \quad (27)$$

In this expression μ_t is a fully general invariant, σ_t is a positive invariant and $\boldsymbol{Z}_t$ is a general invariant with unit scatter and null location: these invariants represent respectively the location, the scatter and the shape of ε_t specification. For instance, for specific choices of the building blocks we recover the Student t distribution:

$$\left.\begin{array}{l} \mu_t \equiv \mu \\ v/\sigma_t^2 \sim \chi_v^2 \\ Z_t \sim N(0,1) \end{array}\right\} \Rightarrow \varepsilon_t \sim S_t\left(v, \mu, \sigma^2\right) \quad (28)$$

We stress that the term stochastic volatility is used mostly when σ_t in (27) is not an invariant, see Section 3.4.

Mixture models

Another flexible aggregate representation is the mixture model

$$\varepsilon_t \overset{d}{=} (1 - B_t) Y_t + B_t Z_t \quad (29)$$

where $\boldsymbol{Y}_t$ and $\boldsymbol{Z}_t$ are generic invariants and $\boldsymbol{B}_t$ are invariants with a standard Bernoulli distribution (23) with values in {0, 1}: the ensuing model for ε_t is a mixture of $\boldsymbol{Y}_t$ and $\boldsymbol{Z}_t$ respectively, with a given probability: it is useful, for instance, to model the regime switch between a regular market environment $\boldsymbol{Y}_t$ and a market crash model $\boldsymbol{Z}_t$. More general mixture models select among S states

$$\varepsilon_t \overset{d}{=} \sum_{s=1}^{S} B_t^{(s)} Y_t^{(s)} \quad (30)$$

where $\boldsymbol{B}_t$ is a multiple-selection process with values in the canonical basis of $\boldsymbol{R}^S$.

Stochastic volatility as mixture models

Stochastic volatility models can be seen as special cases of multi-mixture models with a continuum of components. Indeed Consider a generic stochastic volatility model (27), which we report here

$$\varepsilon_t \overset{d}{=} \mu_t + \sigma_t Z_t \quad (31)$$

Denote by $f\mu,\sigma$ the joint pdf of the location and scatter invariants μ_t and σ_t respectively. Define the following multi-mixture model

$$\varepsilon_t \overset{d}{=} \iint B_t^{(\mu,\sigma)} Y_t^{(\mu,\sigma)} d\mu d\sigma \quad (32)$$

where

$$Y_t^{(\mu,\sigma)} \overset{d}{=} \mu + \sigma Z_t \quad (33)$$

and $\boldsymbol{B}_t^{(\mu,\sigma)}$ is a continuous multiple-selection process with values in the indicators of infinitesimal intervals of $\boldsymbol{R}\times\boldsymbol{R}^+$ such that

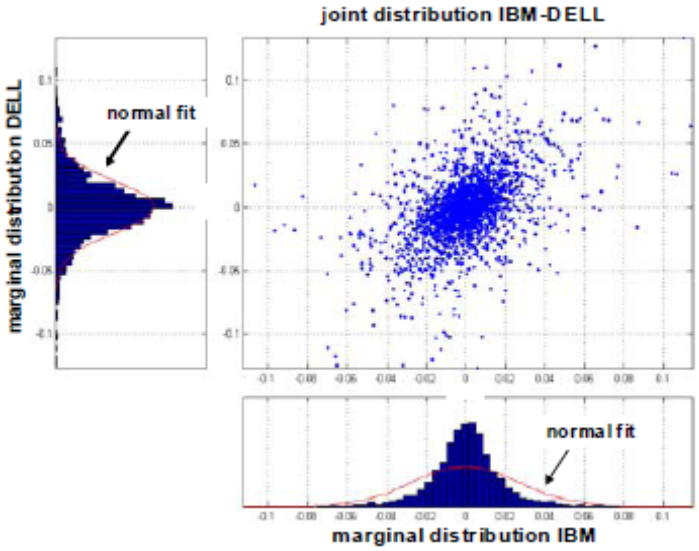
$$P\left\{B_t^{(\mu,\sigma)} = \mathrm{I}_{[d\mu\times d\sigma]}\right\} = f_{\mu,\sigma}(\mu,\sigma) d\mu d\sigma \quad (34)$$

Then the multi-mixture model (32) is distributed as the stochastic volatility model (31).
From (28) the Student t distribution is a special instance of the stochastic volatility representation (27) and thus it is also a multi-mixture distribution with a continuum of components: $\boldsymbol{Z}_t$ in (33) is standard normal, and $f\mu,\sigma \equiv f\mu\, f\sigma$, where $f\mu$ is a Dirac delta and $f\sigma$ is the density of the square root of the inverse chi-square, suitably rescaled.

3.1.4 Heavy tails

Heavy-tailed distributions are those whose pdf decreases away from the location of the distribution slower than exponentially. Suitable models for heavy tails can be found among the continuous distributions, the discrete distributions, or the generalized representations discussed above.

Figure 5 Heavy tails in empirical distribution of compounded stock returns



Stable distributions are continuous models popular for their tractability, see e.g. Rachev (2003)[48] and Meucci (2005): if the invariants ε_t are stable, the random walk X_t has the same distribution as ε_t, modulo rescaling.

Another popular parametric model for heavy tails is the Student t distribution (28), which for v > 2 has finite variance. For v ≡ 1 the Student t distribution is stable and it is known as Cauchy distribution.

In Figure 5 we display the empirical distribution of the daily compound returns (17) of two technology stocks. It is apparent that a normal fit is unsuitable to describe the heavy-body, heavy tail empirical behavior of these returns.

The heavy tails of a random walk tend to disappear with aggregation. More precisely, consider the aggregate invariant

$$\widetilde{\varepsilon}_{t,\tau} \equiv \varepsilon_t + \varepsilon_{t-1} + \ldots + \varepsilon_{t,\tau+1} \quad (35)$$

If the one-step dynamics is the random walk (16), the τ-step dynamics is also a random walk:

$$X_{t+\tau} = X_t + \widetilde{\varepsilon}_{t+,\tau,\tau} \quad (36)$$

However, the central limit theorem prescribes that, if the invariant ε_t has finite variance, the aggregate invariant $\widetilde{\varepsilon}_{t,\tau}$ becomes normal for large enough an aggregation size. In particular It follows that stable distributions must have infinite variance, or else they would violate the central limit theorem.

Therefore, random walks for large enough an aggregation size behave like discrete samples from the Brownian motion, see Section 5.1. From (35) it follows that the variance of the steps of a random walk is a linear function of the time interval of the step. This is the square root rule propagation of risk: risk, defined as the standard deviation of the step, increases as the square root of the time interval:

$$Sd\{\tilde{\varepsilon}_{t,\tau}\}\alpha\sqrt{\tau} \quad (37)$$

For stable distributions, where the variance is not defined, one can use other definitions of risk, such as the inter-quantile range. In particular, for the Cauchy distribution one can check that the propagation law of risk does not grow as the square root of the time interval, but instead it grows linearly with the time interval.

3.2. ARMA processes

The steps of a random walk display no dependence across time, because they are the invariants. This feature makes random walks not stationary: if $\boldsymbol{X}_t$ evolves as in (16), its distribution never stabilizes. To model stationarity we consider the one-lag autoregressive process:

$$\mathrm{X}_{\mathrm{t+1}} = \mathrm{aX}_{\mathrm{t}} + \varepsilon_{\mathrm{t+1}} \quad (38)$$

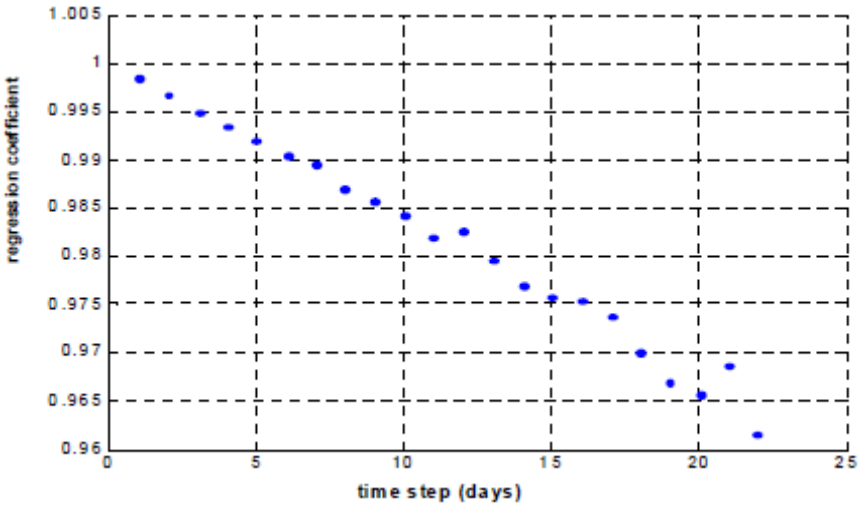
where ε_t are the invariants. The limit a ≡ 1 is the random walk (16). If |a| < 1 the distribution of $\boldsymbol{X}_t$ eventually stabilizes and thus the process becomes stationary.

The autocorrelation of this process reads:

$$Cor\{X_t, X_{t+\tau}\} = e^{(\ln a)\tau} \quad (39)$$

see Appendix 17.3.

Figure 6: AR(1) fit of five-year swap rate



For instance, interest rates cannot evolve as random walks at any scale. If (20) held true exactly then for any aggregation size τ the following would hold true

$$Y_{t+\tau}^{(v)} = Y_t^{(v)} + \tilde{\varepsilon}_{t+\tau,\tau}^{(v)} \quad (40)$$

where the aggregate invariant $\varepsilon_{t+\tau}^{(v)}$ is defined as in (35). However, rates cannot diffuse indefinitely. Therefore, for aggregation size τ of the order of a month or larger mean-reverting effects must become apparent. We see this in Figure 6 where we plot the fitted value of a as a function of the aggregation size for the time series of the five-year par swap rate.

We can generalize (38) by adding more lags of the process and of the invariant.

The so-called autoregressive moving average process, or ARMA (p, q) process is defined as

$$\prod_{j=1}^{p}(1-a_j L)X_t = D_t + \prod_{j=1}^{q}(1-b_j L)\varepsilon_t \quad (41)$$

where L denotes the lag operator $\boldsymbol{L}\boldsymbol{X}_t \equiv \boldsymbol{X}_{t+1}$, and Dt is a deterministic component that we have added in such a way that we can assume the location of ε_t to be zero. It can be proved that the process $\boldsymbol{X}_t$ stabilizes in the long run, and therefore it is stationary, only if all the a's in (41) lie inside the unit circle, see Hamilton (1994)[49]. The fast, exponential decay (39) of the autocorrelation is common to all ARMA(p, q) processes with finite p and q.

It is always possible to switch from an ARMA(p, q), to an ARMA(0, ∞) or an ARMA(∞, 0) representation. Indeed, the following identity holds

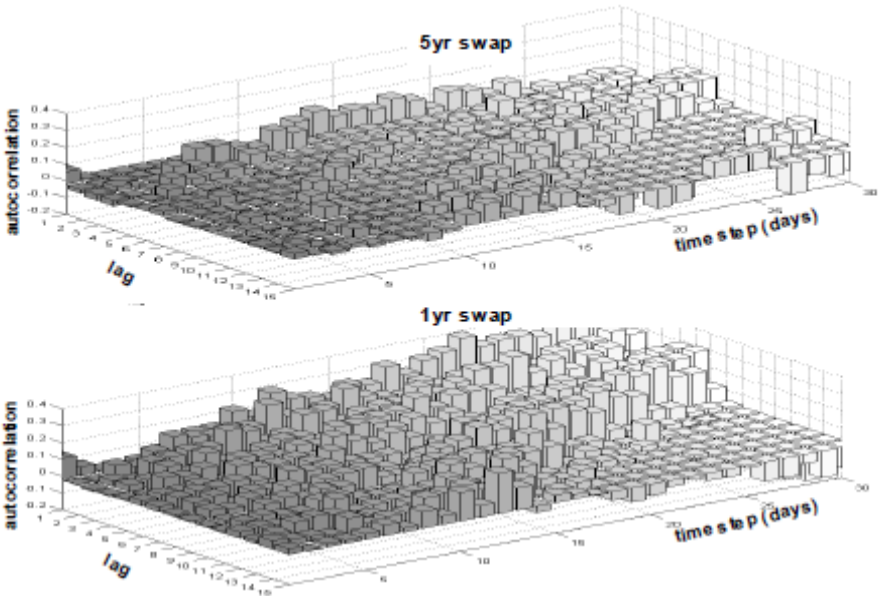
$$(1-\gamma L)^{-1} \equiv \sum_{k=0}^{\infty} (\gamma L)^{k} \quad (42)$$

which can be checked by applying the operator $(1-\gamma L)$ to the right hand side. Therefore, we can remove iteratively all the lags from the MA portion of an ARMA(p, q) process, making it ARMA(∞, 0), as long as q < ∞. Similarly, we can remove all the lags from the AR portion of an ARMA(p, q) process, making it ARMA(0, ∞), as long as p < ∞.

Although the converse is not true, the ARMA(0, ∞) is extremely important because of Wold's theorem, which states that any stationary process can be expressed as ARMA(0, ∞). The infinite parameters of such a process are impossible to estimate, but a parsimonious ARMA(p, q) should in most cases provide a good approximation.

3.3. Long memory

Figure 7: Autocorrelation decay properties of swap rate changes



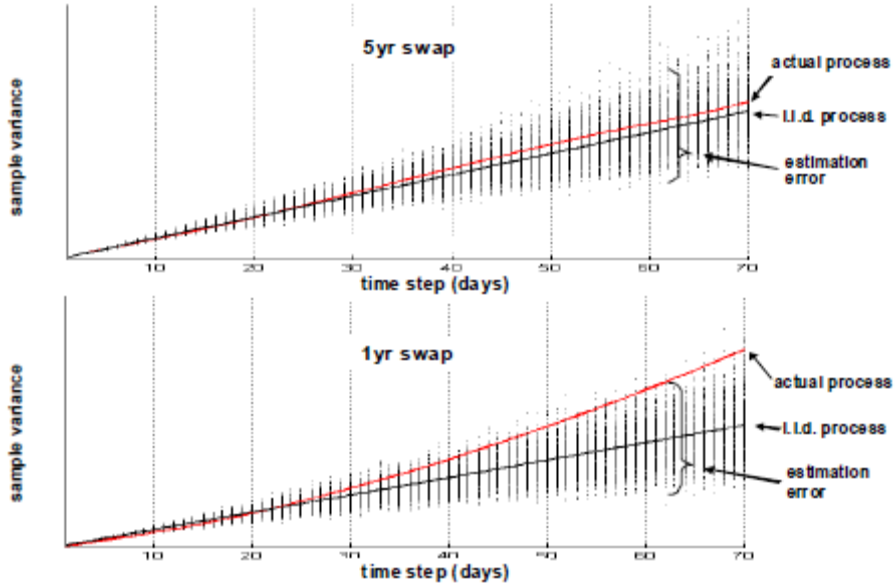
Long memory is a phenomenon whereby the autocorrelation decays slower than exponentially, see e.g. Beran (1994)[50].

For instance, in Figure 7 we display the autocorrelation decay of the fiveyear and the one year swap rates changes. The five year autocorrelation quickly decays to zero. This is consistent with (43) and comments thereafter. On the other hand, the one-year autocorrelation consistently displays a small, yet non-zero pattern. This pattern is persistent, or else the non-linear aggregation properties of the sample variance displayed in Figure 7 would not be justified.

Wold's theorem states that any stationary process can be represented as an ARMA process. However, a parsimonious, finite-lag ARMA specification gives rise to an exponential decay in the

autocorrelation and thus it does not yield long memory. On the other hand, a generic infinite-lag ARMA specification does not make sense from an estimation perspective because of the excessive number of parameters to estimate. Therefore, in order to model long-memory, it becomes necessary to impose a parametric structure on an infinite-lag ARMA

Figure 8: Variance aggregation properties of swap rate changes



This can be achieved by generalizing the random walk (16) into a fractionally integrated process. First we twist the invariants as follows

$$\widetilde{\varepsilon}_t = (1-L)^d \varepsilon_t \quad (43)$$

where 0 ≤ d < 1/2; then we define

$$X_{t+1} = X_t + \widetilde{\varepsilon}_{t+1} \quad (44)$$

If d ≡ 0 in (44) we obtain the standard random walk. If 0 < d < 1/2 we obtain a highly structured integrated ARMA(0, ∞) series. Indeed, the following identity holds

$$(1-L)^d = \sum_{k=0}^{\infty} \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)} L^d \quad (45)$$

In particular, the autocorrelation of the shocks decays according to a power law

$$Cor\{\widetilde{\varepsilon}_t, \widetilde{\varepsilon}_{t-\tau}\} \approx \frac{\Gamma(1-d)}{\Gamma(d)} \tau^{2d-1} \quad (46)$$

see Baillie (1996)[51]. Therefore the fractional process (33) suitably describes the long memory phenomenon. Notice that the autocorrelation decay is reflected in a non-linear, power-law increase of the variance of $\boldsymbol{X}_{t+\tau} - \boldsymbol{X}_t$ as a function of τ .

For instance, the non-linear aggregation properties of the sample variance of the five-year swap rate changes displayed in Figure 6 are consistent with a value d ≈ 0, whereas for the one-year swap rate changes we have d ≈ 0.1.Further generalization of the fractional process (33)-(34) include adding several fractional lags to both the process and the invariant, similarly to (31).

3.4. Volatility clustering

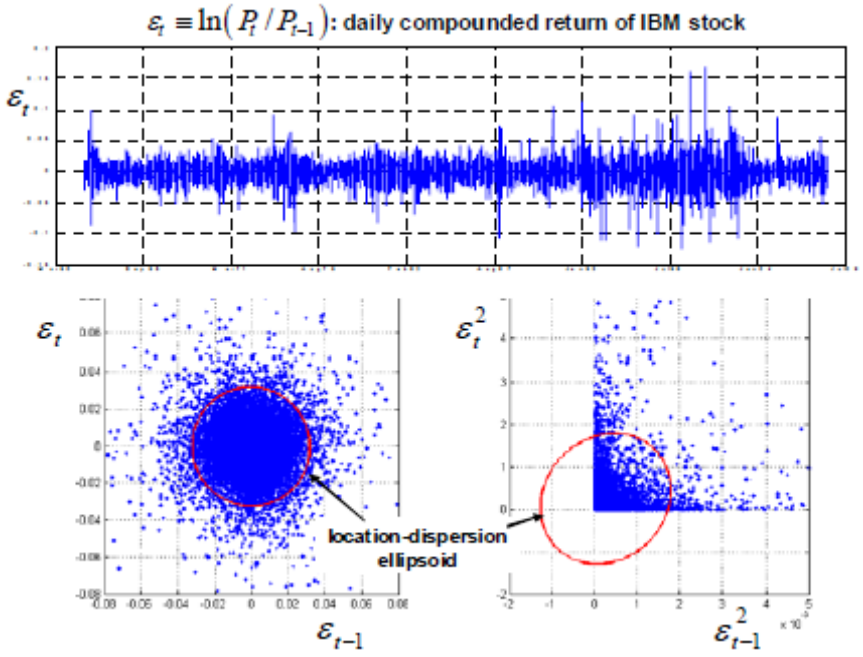
Volatility clustering is the phenomenon whereby the scatter of a financial variable displays evident autocorrelation, whereas the variable itself does not. This is the case among others for the daily returns of most stocks, see e.g. Cont (2005)[52].

For instance, in the left portion of Figure 9 we display the scatter plot of the daily returns of the IBM stock versus their lagged values one day before. The location-dispersion ellipsoid is a circle and thus the log-returns are not autocorrelated. On the other hand, in the right portion of Figure 9 we display the scatter plot of the squared daily returns of the IBM stock versus their lagged values one day before. The location-dispersion ellipsoid now is an ellipse tilted across the positive quadrant: the square log-returns, which proxy volatility, are positively autocorrelated.

The generalized representation (27) of the invariants suggests to model such volatility clustering in terms of stochastic volatility:

$$\varepsilon_t \overset{d}{=} \mu_t + \sigma_t Z_t \quad \textbf{(47)}$$

Figure 9: Autocorrelation properties of stocks log-returns



In this expression, the location invariant is typically constant $\mu_t \equiv \mu$ and $\boldsymbol{Z}_t$ is a generic standardized shape invariant. However, unlike in (27), the scatter term σ_t displays a non-invariant dynamics. The dynamics of σ_t can be specified in terms of an ARMA(p, q) process, a long memory process, or any other model that is not i.i.d. across time.

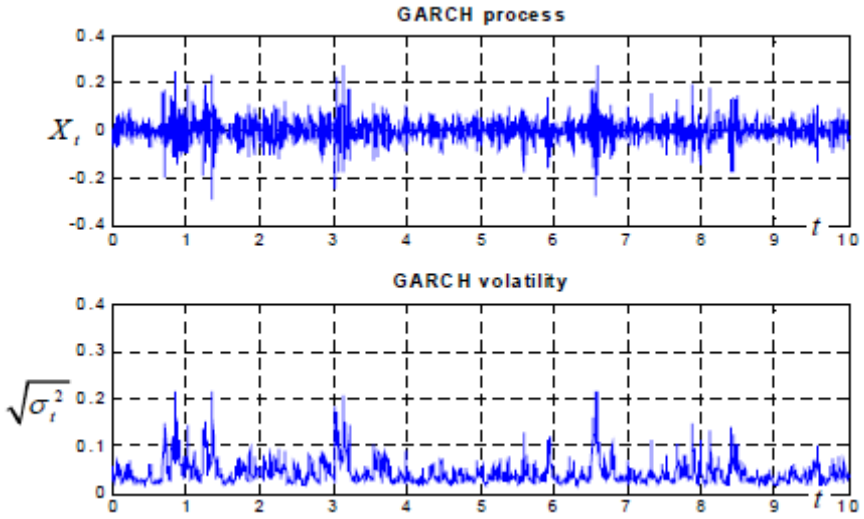
In particular, if the process for σ_t^2 is ARMA(p, q), with the same set of invariants $\boldsymbol{Z}_t$ as in (47), then the process ε_t, which is no longer an invariant, is called a generalized autoregressive conditional heteroskedastic, or GARCH(p, q), process. Among the most popular and parsimonious such specifications is the GARCH(1, 1) model:

$$\sigma_t^2 = \sigma^2 + a\sigma_{t-1}^2 + bZ_{t-1}^2 \quad (48)$$

which is stationary if 0 < a + b < 1.

In Figure 10 we display a sample path from the GARCH process, along with a sample path of the volatility that generates it is driven by sources of randomness other than the invariants $\boldsymbol{Z}_t$ that appear in (47) then we have the most common type of stochastic volatility models, where the variance is not conditionally deterministic.

Figure 10: GARCH model for volatility clustering



4. Part II Q: continuous-time processes

Here we discuss stochastic processes in continuous time: therefore we assume t □ R. In order to support intuition, each section mirrors its discrete-time counterpart in Part I.

4.1. Levy processes

A Levy process X_t is the continuous-time generalization of the random walk with drift. Here we highlights the main features; for more on this subject refer to Schoutens (2003)[53] and Cont and Tankov (2008)[54].

A Levy process is by definition a process such that its increments over any time interval are invariants, i.e. i.i.d. random variables. Therefore, any invariant discussed in Section 3.1 gives rise to a Levy process, as long as the distribution of the invariant is infinitely divisible. This technical condition is important because it ensures that each invariant can be represented in turns as the sum of invariants. More precisely, for an invariant is infinitely divisible if for any integer K the following holds

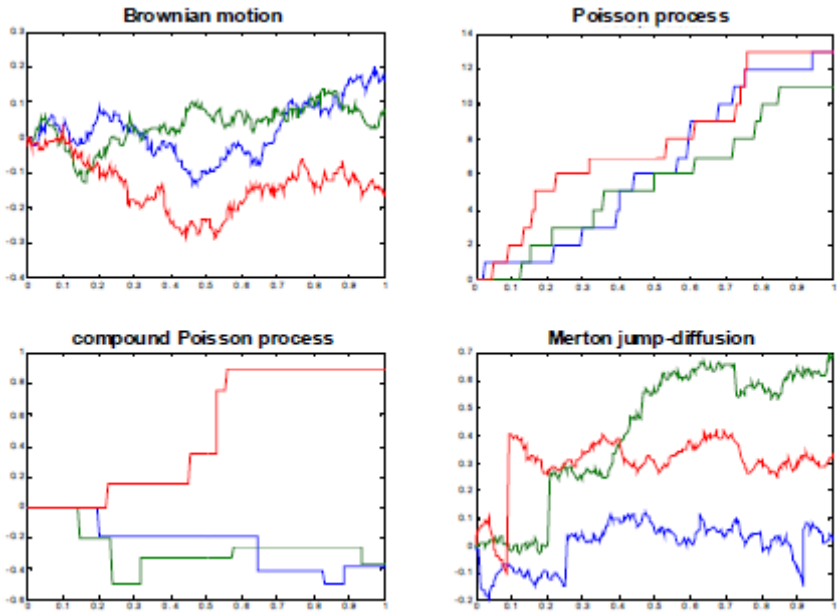
$$\varepsilon_t \equiv X_t - X_{t-1} = \left(X_t - X_{t-\frac{1}{K}} \right) + ... + \left(X_{t-1+\frac{1}{K}} - X_{t-1} \right) \stackrel{d}{=} Z_1 + ... + Z_K \quad (49)$$

where the Z_k's are i.i.d. This requirements is important to guarantee that we can consider time intervals of any size in the definition of the invariant. Levy processes are fundamentally of two kinds: Brownian diffusion and Poisson jumps. As it turns out, any Levy process can be generated with these fundamental bricks.

4.1.1 Diffusion

In order to construct a stochastic diffusion, we need its building blocks, i.e. the increments over any time interval, to be infinitely divisible and to span a continuum of potential values. In order for the model to be tractable, we need to be able to parametrize the distribution of these increments. Therefore, it is natural to turn to continuous stable distributions, and in particular to the one stable distribution that displays finite variance: the normal distribution

Figure 11: Sample paths from Levy processes



The Brownian motion with drift $\boldsymbol{B}_t^{\mu,\sigma^2}$ is a continuous, diffusive process whose invariants are normally distributed with expectation and variance proportional to the time lag between the increments:

$$\varepsilon_{t,\tau} \equiv B_t^{\mu,\sigma^2} - B_{t-\tau}^{\mu,\sigma^2} \sim N\left(\mu\tau, \sigma^2\tau\right) \quad (50)$$

This formulation generalizes (21). Since the sum of independent normal variables is normal, the increments of the Brownian motion are infinitely divisible as in (49), where each term $\boldsymbol{Z}_k$ is normally distributed. Therefore, the Brownian motion is properly defined and it is normally distributed at any time.

The benchmark process in the Q -measure quantitative finance is the Black- Scholes-Merton geometric Brownian motion, which is used to model, among many others financial variables, the price $\boldsymbol{P}_t$ of stocks. This is obtained by modelling the variable $\boldsymbol{X}_t \equiv \ln \boldsymbol{P}_t$ as a Brownian motion.

$$\ln P_t \overset{d}{\equiv} B_t^{\mu,\sigma^2} \quad (51)$$

Notice that the random walk (18) is much more general, in that the invariants, i.e. the compounded returns, need not be normally distributed. In Figure 9 we plot a few paths from the Brownian motion.

4.1.2 Jumps

In order to construct a stochastic jump, we need its building blocks, i.e. the increments over any time interval, to be infinitely divisible and to span a discrete set of potential values. In order for the model to be tractable, we need to be able to parametrize the distribution of these increments. From (26) it follows that the natural choice in this case is the Poisson distribution.

The Poisson process $\boldsymbol{P}_t^{\Delta,\lambda}$ jumps by positive multiples of a base (positive or negative) step Δ according to a Poisson distribution:

$$\varepsilon_{t,\tau} \equiv P_t^{\Delta,\lambda} - P_{t-\lambda}^{\Delta,\lambda} \sim P_o\left(\lambda\tau;\Delta\right) \text{(52)}$$

Since from (26) the sum of independent Poisson variables has a Poisson distribution, the increments of the Poisson process are infinitely divisible as in (49), where each term $\boldsymbol{Z}_k$ is Poisson distributed. Therefore, the Poisson process is properly defined and it is Poisson distributed at any non-integer time. In Figure 11 we plot a few paths from the standard Poisson process, which takes values on the integer grid.

As the time step decreases so does the probability of a jump. Indeed from (52) and (25) we obtain

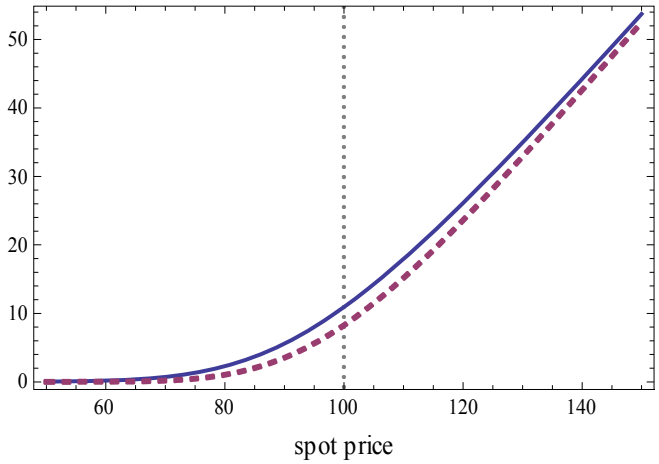
$$P\left(\varepsilon_{t,\tau} = 0\right) \approx 1 - \lambda\tau \text{ (53)}$$

$$P\left(\varepsilon_{t,\tau} = \Delta\right) \approx \lambda\tau \text{ (54)}$$

$$P\left(\varepsilon_{t,\tau} > \Delta\right) \approx 0 \text{ (55)}$$

Since $\boldsymbol{E}\{\varepsilon_t\} = \lambda\tau$, the parameter λ plays the role of the intensity of the jumps.

Figure 12: Option Prices in Merton's Jump Diffusion Model



The jump diffusion model, introduced in 1976 by Robert Merton[55], is a model for stock price behavior that incorporates small day-to-day "diffusive" movements together with larger, randomly occurring "jumps". The inclusion of jumps allows for more realistic "crash" scenarios and means that the standard dynamic replication hedging approach of the standard Black-Scholes model no

longer works. This causes option prices to increase compared to the Black-Scholes model and to depend on the risk aversion of investors. This figure how the price of European call options varies with the jump diffusion model parameters compared to black-scholes.

4.1.3 Generalized representations

The generic Levy process is a sum of the above processes. From (22) and (50), the weighted sum of independent Brownian motions is a Brownian motion:

$$\sum_{k=1}^{K} B_{t,k}^{\mu_k,\sigma^2} \overset{d}{=} B_t^{\mu_k,\sigma^2} \quad (56)$$

Where $\mu \equiv \sum_{k=1}^{K} \mu_k$ and $\sigma^2 \equiv \sum_{k=1}^{K} \sigma_k^2$. Also, from (26) and (52), the sum of independent Poisson processes with the same base step Δ is a Poisson process with that base step:

$$\sum_{k=1}^{K} P_t^{\Delta,\lambda_k} \overset{d}{=} P_t^{\Delta,\lambda} \quad (57)$$

where $\lambda \equiv \sum_{k=1}^{K} \lambda_k$. Therefore we can construct more general Levy processes from our building blocks by considering a continuum of terms in the sum as follows

$$X_t = B_t^{\mu,\sigma^2} + \int_{-\infty}^{\infty} P_t^{\Delta,\lambda_k} \, d\Delta \quad (58)$$

where the intensity $\lambda(\Delta)$ determines the relative weight of the discrete Poisson jumps on the grid ΔN. As it turns out, this representation is exhaustive: any Levy process, stemming from slicing as in (50) an arbitrarily distributed infinitely divisible invariant, can be represented as (58). As we show in Appendix 17.5, from (58) follows the Levy-Khintchine representation of Levy processes in terms of their characteristic function:

$$\ln\left(E\left\{e^{iwX_t}\right\}\right) = iw\mu t - \frac{1}{2}\sigma^2 w^2 t + t\int \left(e^{iw\Delta}\right)\lambda(\Delta)d\Delta \quad (59)$$

An alternative representation of the Levy processes follows from a result by Monroe (1978)[56], according to which Levy processes can be written as

$$X_t \overset{d}{=} B_T^{\mu_k,\sigma^2} \quad (60)$$

where Bt is a Brownian motion and Tt is another Levy process that never decreases, called subordinator. In practice, Tt is a stochastic time that indicates the activity of the financial markets. The specific case $T_t \equiv t$ recovers the Brownian motion.

4.1.4 Notable examples

In addition to (geometric) Brownian motion and Poisson processes, another important class of Levy processes are the α-stable processes by Mandelbrot (1963)[57], see also Samorodnitsky and Taqqu (1994)[58], Rachev (2003), whose distribution is closed under the sum. However, in order not to contradict the central limit theorem these processes have infinite variance.

Another important subclass are the compound Poisson processes, which generalize Poisson processes by allowing jumps to take on random values, instead of values in a fixed grid ΔN. The compound Poisson processes can be expressed as

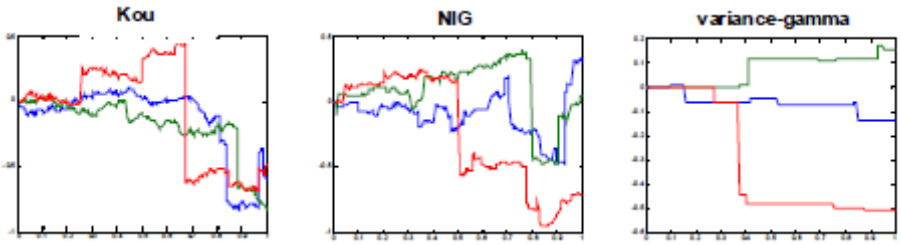
$$X_t = B_t^{\mu,\sigma^2} + \sum_{n=1}^{P_t^\lambda} Z_n \quad (61)$$

where P_t^λ is a Poisson process with values in the unit-step grid N and Zn are arbitrary i.i.d. variables. As we show in Appendix 17.6, the parameterization (58) for jump-diffusion models (61) is given by

$$\lambda(\Delta) \equiv \lambda f_Z(\Delta) \quad (62)$$

were f_Z is the pdf of Z_n.

Figure 13: Sample paths from Levy processes



A notable example in the class of compound Poisson processes is the thejump diffusion by Merton (1976), where Zn is normally distributed. We plot in Figure 13 a few paths sampled from this process.

Another tractable jump-diffusion process is the double-exponential by Kou (2002)[59]. Other notable parametric Levy processes include the Normal-Inverse- Gamma by Barndorff-Nielsen (1998)[60], the variance-gamma process by Madan and Milne (1991)[61] and the CGMY by Carr, Geman, Madan, and Yor (2003)[62], see Figure 9.

Figure 14: Difference NIG / BS

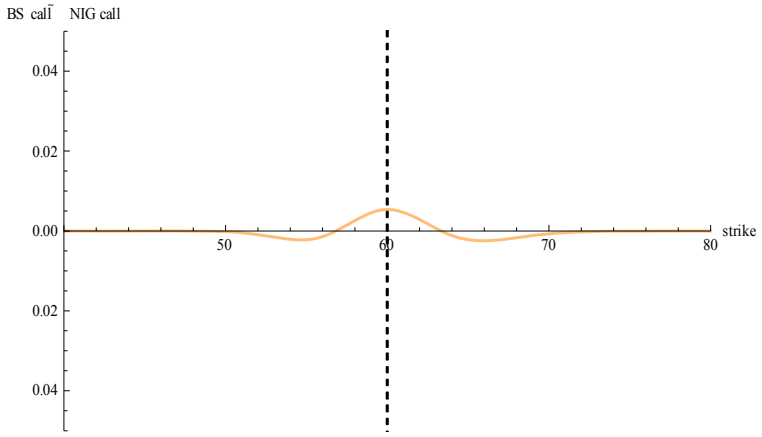


Figure 14 shows the difference between the price of a European call option on stock (no dividend is paid and the interest rate is 0) in the Black\Scholes model and the model (due to Barndorff\Nielson) based on a centered exponential Normal Inverse Gaussian (NIG) Lévy process, as a function of the strike price. There are three model parameters that control the NIG process: steepness, asymmetry, and scale (the fourth parameter, location, is set to 0). The other control parameter is the time to expiry of the option. Steepness parameter=100.7, assimetry=0.02, scale parameter= 0.502, 6 months expiry.

Figure 15: Option Prices in the Variance Gamma Model

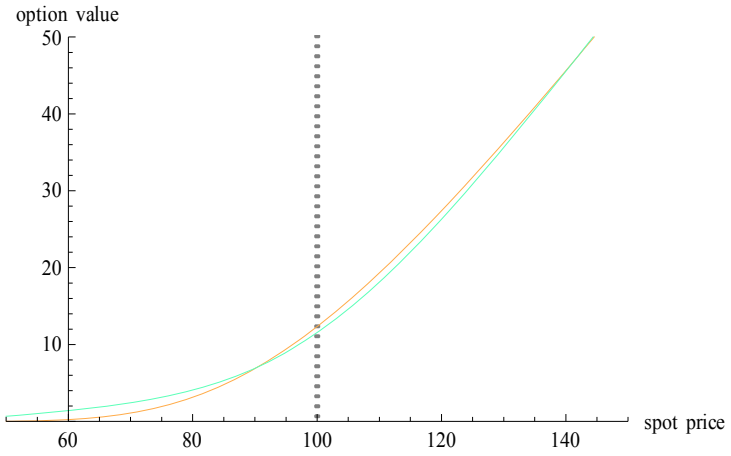


Figure 15 compares the values of the vanilla European Call option in the Black-Scholes model with the values of the same options in the Variance Gamma model. The strike price is fixed at 100. The control parameters volatility, risk-free interest rate and time to expiry are shared by both models while the parameters "drift" and "gamma variance" affect only the Variance Gamma model. Drift=0.04, gamma variance= 2, volatility: 0.25, risk free= 0.05, 1y expiry.

4.2. Autocorrelated processes

The continuous-time version of the AR(1) process (38) is the Ornstein-Uhlenbeck process. Its dynamics is defined by a stochastic differential equation

$$dX_t = -\theta(X_t - \mu)dt + dL_t \quad (63)$$

where μ and σ > 0 are constants and L is a Levy process. This process is defined for θ ≥ 0 and is stationary if θ > 0. This dynamics is easy to interpret: a change in the process is due to a deterministic component, which tends to pull the process back towards the equilibrium asymptote μ at an exponential rate, and a random shock dLt that perturbs this exponential convergence.
Of particular importance is the case where the Levy driver is a Brownian motion:

$$dX_t = -\theta(X_t - \mu)dt + \sigma dB_t \quad (64)$$

As we show in Appendix 17.3 this dynamics can be integrated explicitly:

$$\boldsymbol{X_t} = \left(1 - \boldsymbol{e}^{-\theta\tau}\right)\mu + \boldsymbol{e}^{-\theta\tau}\boldsymbol{X_{t-\tau}} + \eta_{t,\tau} \quad (65$$

where $\eta_{t,\tau}$ are normally distributed invariants

$$\eta_{t,\tau} \sim N\left(0, \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta\tau}\right)\right) \quad (66)$$

Then the long-term autocorrelation reads:

$$Cor\{X_t, X_{t+\tau}\} = e^{-\theta\tau} \quad (67)$$

A comparison of (38) with (65) yields the correspondence between discrete and continuous time parameters. Notice that for small time steps τ the process (64) behaves approximately as a Brownian motion. Indeed, a Taylor expansion of the terms in (65)-(66) yields

$$X_t \approx X_{t-\tau} + \varepsilon_{t,\tau} \quad (68)$$

where $\varepsilon_{t,\tau}$ is in the form of the normal invariant in (50). On the other hand, as the step τ increases, the distribution of the process stabilizes to its stationary behavior

$$X_t \sim N\left(\mu, \frac{\sigma^2}{2\theta}\right) \quad (69)$$

The AR(1) process for the five year swap rate in Figure 5 can be seen as the discrete-time sampling of an Ornstein-Uhlenbeck process, which in the context of interest rate modelling is known as the Vasicek model , see Vasicek (1977)[63].

Another useful autocorrelated process is obtained by applying Ito's rule to the square $\boldsymbol{Y_t \equiv X_t^2}$ of the process(63) for m ≡ 0. As we show in Appendix 17.4, this yields the CIR process by Cox, Ingersoll, and Ross (1985)[64]:

$$dY_t = -\tilde{\theta}\left(Y_t - \tilde{m}\right)dt + \tilde{\sigma}\sqrt{Y_t}\,dB_t \quad (70)$$

where the parameters $\tilde{\theta}_t$, $\tilde{\boldsymbol{m}}$ and $\tilde{\sigma}$ are simple functions of the parameters in (64). This process, is useful to model the evolution of positive random variables. For instance, the CIR dynamics provides an alternative to the Vasicek model for interest rates. Furthermore, the CIR process can model stochastic volatility, see Section 4.4.1.

The same way as by adding lags to the one-lag autoregression (38) we obtain the ARMA processes (41), so it is possible to add differentials to the Ornstein- Uhlenbeck process (63): the result are the so-called continuous autoregressive moving average, or CARMA, processes.

4.3. Long memory

As discussed in Section 4, some financial variables display long memory: the empirical autocorrelation displays a decay slower than the exponential pattern (67) prescribed by the Ornstein-Uhlenbeck or, in discrete time, by finite-lags ARMA processes.

The continuous-time version of the fractional process with normal invariants is the fractional Brownian motion. Similarly to (43)-(45), this process is defined as a structured average of past invariants

$$X_t \equiv \mu t + \sigma \int_{-\infty}^{t} \frac{(t-s)^d}{\Gamma(d+1)} dB_s \quad (71)$$

where 0 ≤ d < 1/2. This increments of this process

$$X_t = X_{t-\tau} + \tilde{\varepsilon}_{t,\tau} \quad (72)$$

are normally distributed:

$$\tilde{\varepsilon}_{t,\tau} \sim N\left(\mu\tau, \sigma^2\tau^{2H}\right) \quad (73)$$

where

$$H \equiv d + \frac{1}{2} \quad (74)$$

is the Hurst coefficient. For d ≡ 0 we recover the regular Brownian motion, where the increments are invariants as in (50). If 0 < d < 1/2, the increments are identically distributed and normal, but they are not independent, and therefore they do not represent invariants.

Figure 16: Sample paths from a fractional Brownian motion

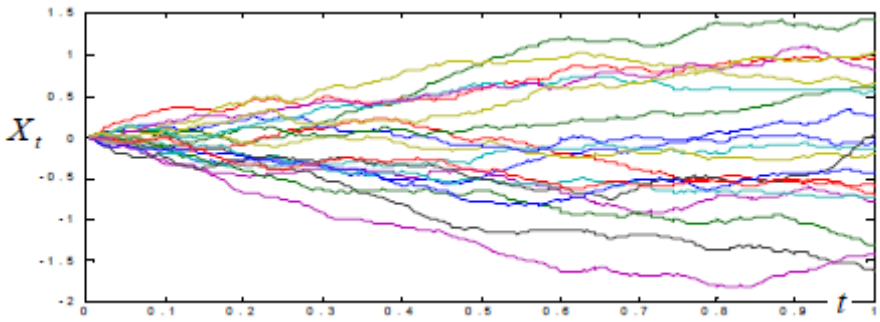
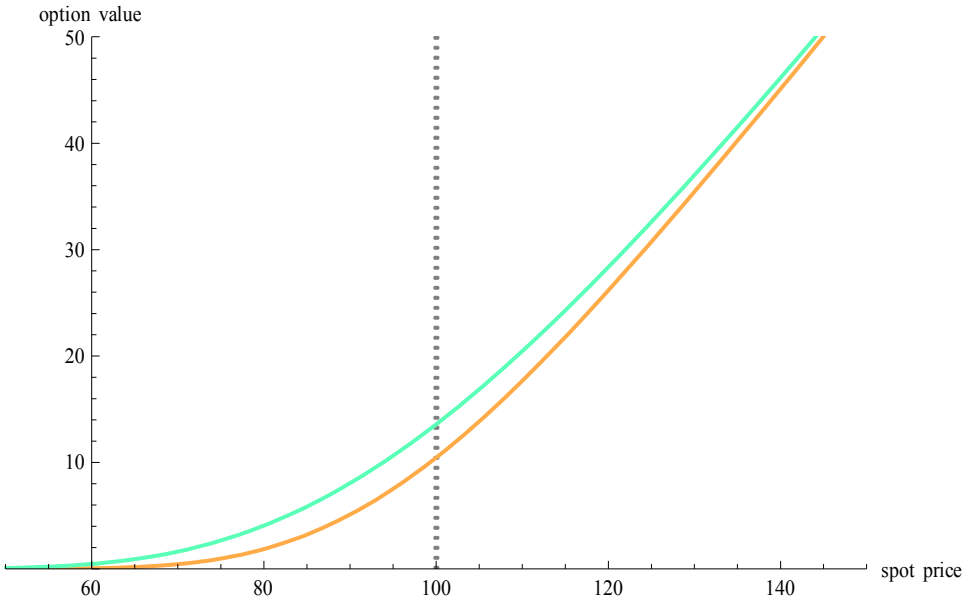


Figure 16 displays the persistent pattern of a few sample paths of fractional Brownian motion with Hurst coefficient H = 0.8. For a financial application of this process, refer to the one-year swap rate in Figure 5.

Figure 17: Option Prices under the Fractional Black-Scholes Model : Hurst exponent = 1



Unlike for the Ornstein-Uhlenbeck process (66), the differences of the fractional Brownian motion display a power law decay:

$$Cor\left(\Delta X_t, \Delta X_{t+\tau}\right) \approx d\left(2d+1\right)\tau^{2d-1} \quad (75)$$

see Comte and Renault (1996)[65], Baillie (1996).

4.4. Volatility clustering

To model volatility clustering in continuous time, we can proceed in two ways.

4.4.1 Stochastic volatility

To generate volatility clustering in discrete time we modified the stochastic volatility representation of the invariants (27) by imposing autocorrelation on the scatter as in (47).
Similarly, in continuous time first we represent the dynamics of the cumulative invariants, i.e. the generic Levy process (58) as follows:

$$X_t \equiv \mu t + \sigma Z_t \quad (76)$$

In this expression the first term is the location of $\boldsymbol{X}_t$ which must grow linearly; $\boldsymbol{Z}_t$ is a zero-location Levy process whose scatter at time t ≡ 1 is one; and σ is a positive constant. Then we modify the (trivial) Levy process for σ into a stationary process σt that displays autocorrelation: in this situation $\boldsymbol{X}_t$ is nolonger a Levy process and volatility clustering follows.
The most popular such model is the Heston model ,see Heston (1993)[66]: the normalized Levy process $\boldsymbol{Z}_t$ in (76) is a Brownian motion; the scatter parameter follows a CIR process(70) shocked by a different Brownian motion

$$d\sigma_t^2 = -k\left(\sigma_t^2 - \overline{\sigma}^2\right)dt + \lambda\sqrt{\sigma_t^2}\, dB_t \quad (77)$$

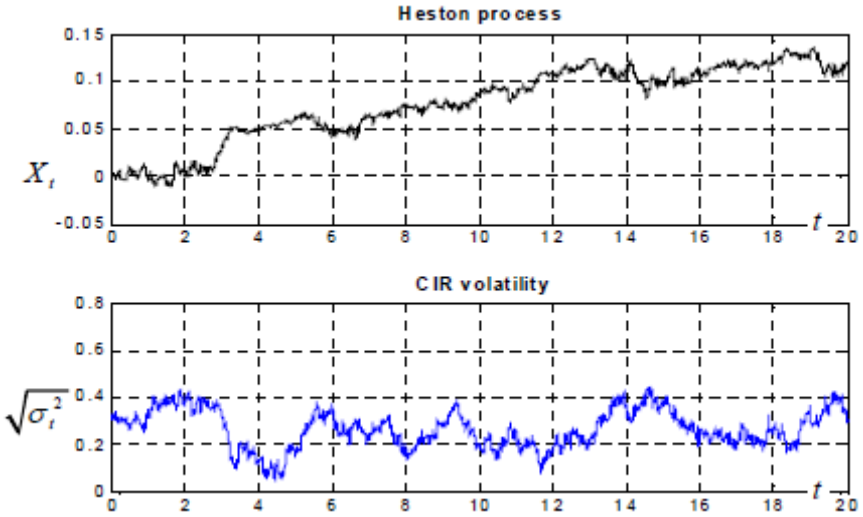
where

$$k\overline{\sigma}^2 > \lambda^2 \quad (78)$$

the copula between $\boldsymbol{B}_t$ and $\boldsymbol{Z}_t$ is normal at all times with constant correlation ρ; and ρ < 0 due to the leverage effect: lower returns tend to correspond to higher volatility. The Heston model is

heavily used by Q-finance quants to price options. Indeed, the original Black-Scholes-Merton model (51) is inconsistent with the observed smiles, skews and smirks in the implied volatility profile.

Figure 18: Heston stochastic volatility process



In Figure 18 we display a sample path from the Heston stochastic volatility process, along with a sample path of the volatility that generates it.

4.4.2 Subordination

An alternative way to model volatility clustering in continuous time is inspired by the self-similarity of the Brownian motion:

$$B_{\sigma^2 t} \overset{d}{=} \sigma B_t \quad (79)$$

Consider the subordination (60) of a Levy processes:

$$X_t \overset{d}{=} B_T^{\mu,\sigma^2} \quad (80)$$

If the subordinator, i.e. the stochastic time $\boldsymbol{T_t}$ increases faster than linearly, than the volatility increases: periods of large volatility correspond to periods where the pace of the market increases. In order to generate volatility clustering, i.e. autocorrelation in volatility, we simply relax the assumption that the pace of the market $\boldsymbol{T_t}$ is a Levy process.

Since Tt must be increasing, it can be defined as the integral of a process T˙s which is positive at all times

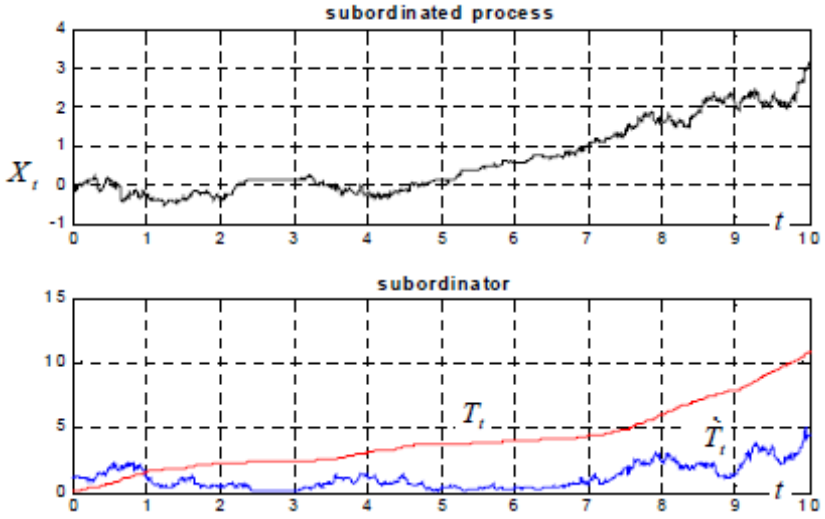
$$T_t \equiv \int_0^t \dot{T}_s ds \quad (81)$$

A natural choice for T˙s is the CIR process (70) and a natural choice for the equilibrium value for ˙Ts is one:

$$d\dot{T}_t = -\tilde{\theta}\left(\dot{T}_t - 1\right)dt + \tilde{\sigma}\sqrt{\dot{T}_t}\, dZ_t \quad (82)$$

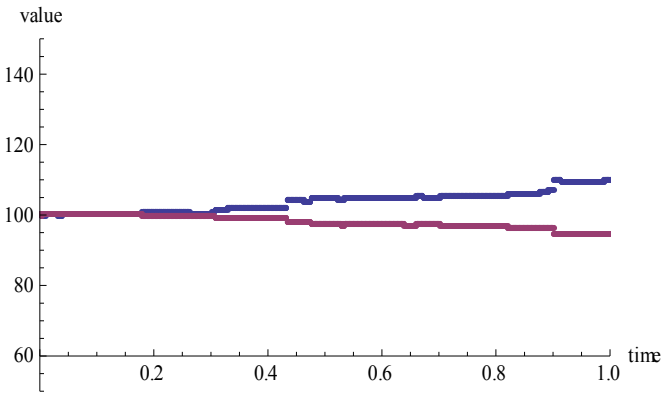
In order to further increase the flexibility of the subordination approach to stochastic volatility, one can subordinate a generic Levy process instead of the Brownian motion in (80).

Figure 19: CIR-driven subordinated Brownian motion



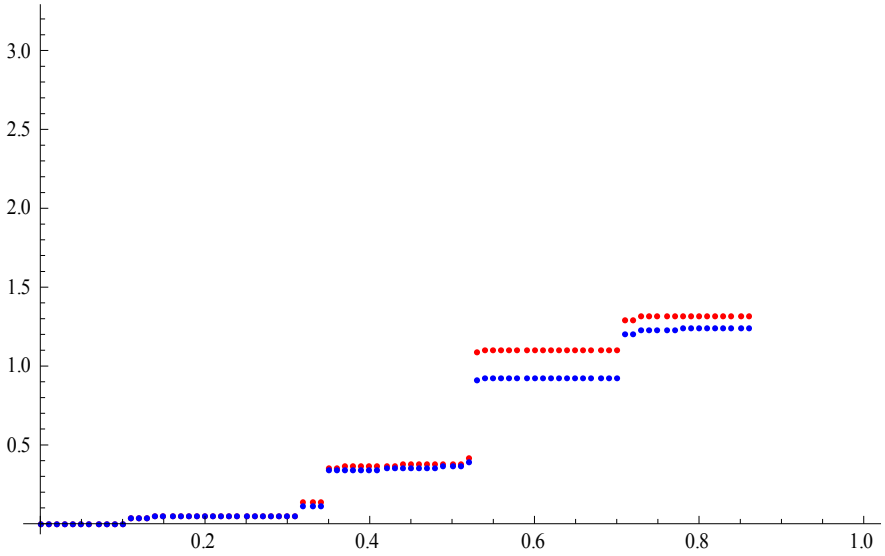
In Figure 19 we display a sample path from the subordinated Brownian motion, along with a sample path of the subordinator (81) that generates it, as well as the CIR process that generates the subordinator. Notice that periods of higher volatility correspond to period where the stochastic time elapses faster.

Figure 20: Correlated Gamma Variance Processes with Common Subordinator



This Figure shows the movements of the prices of two stocks given by exponential correlated Brownian motion that are time-changed with the same gamma process subordinator. In other words, the stock prices are given by two correlated exponential variance gamma Lévy processes whose large jumps tend to occur at the same time.

Figure 21: Correlated Lévy Processes via Lévy Copulas



This Figure shows a path of a two-dimensional subordinator (Lévy process with only positive jumps) whose components are 1/2-stable Lévy processes with dependence structure given by the Clayton\Lévy copula with parameter θ, which determines the degree of dependence between the marginal processes. Large values of θ correspond to strong dependence, small values to weak dependence.

4.5. Markov Switching Models

Markov switching models belong to a vast family of models that have found applications in many fields other than econometrics, such as genomics and speech recognition. The economic idea behind Markov switching models is that the economy undergoes discrete switches between economic states at random times. To each state corresponds a set of model parameters.

One of the first Markov switching models proposed is the Hamilton model[67]. The Hamilton model is based on two states, a state of “expansion” and a state of “recession.” Periods of recession are followed by periods of expansion and vice versa. The time of transition between states is governed by a two-state Markov chain. In each state, price processes follow a random walk model.

The Hamilton model can be extended to an arbitrary number of states and to more general VAR (Vector AutoRegressive) models. In a Markov switching context, a VAR model

$$\boldsymbol{x}_t = \left(\boldsymbol{A}_1\boldsymbol{L} + \boldsymbol{A}_2\boldsymbol{L}^2 + \ldots + \boldsymbol{A}_n\boldsymbol{L}^N\right)\boldsymbol{x}_t + \boldsymbol{m}(\boldsymbol{s}_t) + \varepsilon_t \quad (83)$$

has parameters that depend on a set of hidden states that are governed by a discrete-state, discrete-time Markov chain with transition probability matrix:

$$\begin{aligned} &\boldsymbol{p}_{i,j} = \boldsymbol{P}_r\left(\boldsymbol{s}_{t+1} = \boldsymbol{i} \middle| \boldsymbol{s}_t = \boldsymbol{i}\right) \\ &\sum_{j=1}^{M} \boldsymbol{p}_{i,j} = 1 \end{aligned} \quad (84)$$

Estimation of Markov switching VAR models can be done within a general maximum likelihood framework. The estimation procedure is rather complex as approximate iteration techniques are used. Hamilton made use of the Expectation Maximization (EM) algorithm (Arthur Dempster, Nan Laird, and Donald Rubin 1977)[68]. We will show am application afterwards.

Markov switching VAR models have been applied to macroeconomic problems, in particular to the explanation of business cycles. Applications to the modelling of large portfolios present significant problems of estimation given the large number of data necessary. Markov switching models are, in fact, typically estimated over long periods of time, say 20 or 30 years. If one wants to construct coherent data sets for broad aggregates such as the S&P 500, one rapidly runs into problems as many firms, over periods of that length, undergo significant change such as mergers and acquisitions or stock splits. As one cannot simply exclude these firms as doing so would introduce biases in the estimation process, ad hoc adjustment procedures are needed to handle change. Despite these difficulties, however, Markov switching models can be considered a promising technique for financial econometrics.

4.6. Fractals and multifractals in finance

4.6.1 Basic definitions

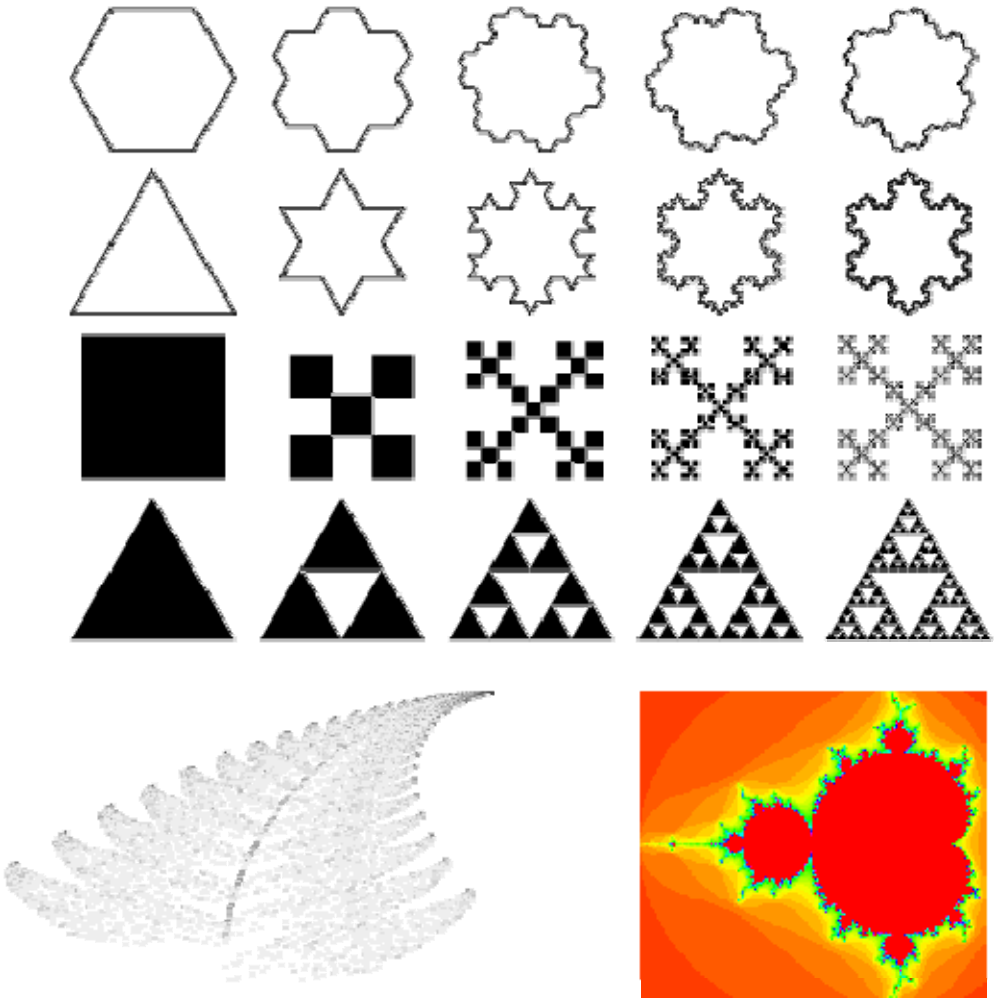
In short, fractals are the geometric objects that are constructed by repeating geometric patterns at a smaller and smaller scale. The fractal theory is a beautiful theory that describes beautiful objects.

Development of the fractal theory and its financial applications has been greatly influenced by Mandelbrot.[69]. In this chapter, a short introduction to the fractal theory relevant to financial applications is given. In this section, the basic definitions of the fractal theory are provided. Section 4.6.2 is devoted to the concept of multifractals that has been receiving a lot of attention in the recent research of the financial time series.

Self-similarity is the defining property of fractals. This property implies that the geometric patterns are isotropic, meaning shape transformations along all coordinate axes are the same. If the geometric patterns are not isotropic, say the object is contracted along the y-axis with a scale

different from that of along the x-axis, it is said that the object is self-affine. The difference between self-similarity and self-affinity is obvious for geometric objects. However, only selfaffinity is relevant for the graphs of financial time series. Indeed, since time and prices are measured with different units, their scaling factors cannot be compared. If the geometric pattern used in fractal design is deterministic, the resulting object is named a deterministic fractal. If the repetition or deletion is random then we have a random fractal. While the deterministic and stochastic fractals look quite different, they have the same fractal dimension.

Figure 22: Fractals



Consider a jagged line, such as a coastline. It is embedded into a plane. Thus, its dimension is lower than two. Yet, the more zigzagged the line is, the greater part of plane it covers. One may then expect that the dimension of a coastline is higher than one and it depends on a measure of jaggedness. Another widely used example is a crumpled paper ball. It is embedded in three-dimensional space. Yet, the volume of a paper ball depends on the sizes of its creases. Therefore, its dimension is expected to be in the range of two to three. Thus, we come to the notion of the fractal (non-integer) dimension for objects that cannot be accurately described within the framework of Euclidian geometry.

There are several technical definitions for the fractal dimension[70]. The most popular one is the box-counting dimension. It implies mapping the grid boxes of size h (e.g., squares and cubes for the two-dimensional and the three-dimensional spaces, respectively) onto the object of interest. The number of boxes that fill the object is $N(h) \sim h^{-D}$. The fractal dimension D is then the limit

$$D = \lim_{h \to 0} [\ln N(h) / \ln(1/h)] \quad (85)$$

The box-counting dimension has another equivalent definition with the fixed unit size of the grid box and varying object size L

$$D = \lim_{L \to \infty} [\ln N(L) / \ln(L)] \quad (86)$$

Random fractals exhibit self-similarity only in a statistical sense. Therefore, the scale invariance is a more appropriate concept for random fractals than self-similarity. The iterated function systems are commonly used for generating fractals. The two-dimensional iterated function algorithm for N fixed points can be presented as

$$\begin{aligned} X(k+1) &= \rho X(k) + (1-\rho) X_F(i) \\ Y(k+1) &= \rho Y(k) + (1-\rho) Y_F(i) \end{aligned} \quad (87)$$

In (87) ρ is the scaling parameter; XF(i) and YF(i) are the coordinates of the fixed point i; i = 1, 2, ... N. The fixed point i is selected at every iteration at random. A famous example with N = 3, the Sierpinski triangle, is shown in Figure 22.

Now, let us turn to the random processes relevant to financial time series. If a random process X(t) is self-affine, then it satisfies the scaling rule

$$X(ct) = c^H X(t) \quad (88)$$

The parameter H is named the Hurst exponent. Let us introduce the fractional Brownian motion BH(t). This random process satisfies the following conditions for all t and T [1]

$$\begin{aligned} E[B_H(t+T) - B_H(t)] &= 0 \\ E[B_H(t+T) - B_H(t)]^2 &= T^{2H} \end{aligned} \quad (89)$$

When H = 1⁄2 , the fractional Brownian motion is reduced to theregular Brownian motion. For the Brownian motion, the correlation between the past average

$$Correlation(E[B_H(t) - B_H(t+T)]/T, E[B_H(t+T) - B_H(t)]/T) = 2^{2H-.1} - 1 \quad (90)$$

Obviously, this correlation does not depend on T. If 1⁄2 < H < 1, then C > 0 and it is said that BH(t) is a persistent process. Namely, if BH(t) grew in the past, it will most likely grow in the immediate future.

Conversely, if BH(t) decreased in the past, it will most probably continue to fall. Thus, persistent processes maintain trend. In the opposite case (0 < H < 1⁄2, C < 0), the process is named anti-persistent. It is said also that anti-persistent processes are mean reverting; for example, if the current process innovation is positive, then the next one will most likely be negative, and vice

versa. There is a simple relationship between the box-counting fractal dimension and the Hurst exponent

$$D = 2 - H \quad (91)$$

The fractal dimension of a time series can be estimated using the Hurst's rescaled range (R/S) analysis . Consider the data set xi(i = 1, . . . N) with mean mN and the standard deviation σN. To define the rescaled range, the partial sums Sk must be calculated

$$\boldsymbol{S_k} = \sum_{\boldsymbol{i}=1}^{\boldsymbol{k}} \left(\boldsymbol{x_i} - \boldsymbol{m_N}\right),\ 1 \le \boldsymbol{k} \le \boldsymbol{N} \quad (92)$$

The rescaled range equals

$$\boldsymbol{R}/\boldsymbol{S} = \left[\max\left(\boldsymbol{S_k}\right) - \min\left(\boldsymbol{S_k}\right)\right]/\sigma_N,\ 1 \le \boldsymbol{k} \le \mathrm{N} \quad (93)$$

The value of R/S is always greater than zero since max (Sk) > 0 and min (Sk) < 0. For given R/S, the Hurst exponent can be estimated using the relation

$$\boldsymbol{R}/\boldsymbol{S} = \left(\boldsymbol{aN}\right)^{\boldsymbol{H}} \quad (94)$$

where a is a constant. The R/S analysis is superior to many other methods of determining long-range dependencies. But this approach has a noted shortcoming, namely, high sensitivity to the short-range memory[71].

4.6.2 Multifractals

Let us turn to the generic notion of multifractals[72]. Consider the map filled with a set of boxes that are used in the box-counting fractal dimension. What matters for the fractal concept is whether the given box belongs to fractal. The basic idea behind the notion of multifractals is that every box is assigned a measure m that characterizes some probability density (e.g., intensity of color between the white and black limits). The so-called multiplicative process (or cascade) defines the rule according to which measure is fragmented when the object is partitioned into smaller components. The fragmentation ratios that are used in this process are named multipliers. The multifractal measure is characterized with the Holder exponent α

$$\alpha = \lim_{\boldsymbol{h}\to 0}\left[\ln \mu\left(\boldsymbol{h}\right)/\ln\left(\boldsymbol{h}\right)\right] \quad (95)$$

where h is the box size. Let us denote the number of boxes with given h and a via Nh(α). The distribution of the Holder exponents in the limit h → 0 is sometimes called the multifractal spectrum

$$\boldsymbol{f}\left(\alpha\right) = -\lim_{\boldsymbol{h}\to 0}\left[\ln \boldsymbol{N_h}\left(\alpha\right)/\ln\left(\boldsymbol{h}\right)\right] \quad (96)$$

The distribution f(α) can be treated as a generalization of the fractal dimension for the multifractal processes.

Let us describe the simplest multifractal, namely the binomial measure m on the interval [0, 1]. In the binomial cascade, two positive multipliers, m0 and m1, are chosen so that m0 + m1= 1. At the step k = 0, the uniform probability measure for mass distribution, μ0 = 1, is used. At the next step (k =1), the measure m1 uniformly spreads mass in proportion m0=m1 on the intervals [0, 1⁄2] and

[1⁄2 , 1], respectively. Thus, μ1[0, 1⁄2] = m0 and μ1[1⁄2 , 1] = m1. In the next steps, every interval is again divided into two subintervals and the mass of the interval is distributed between subintervals in proportion m0/m1.

For example, at k = 2: μ 2[0, 1⁄4]= m0m0, μ 2[1⁄4, 1⁄2] = μ 2[1⁄2, 3⁄4] = m0m1, μ 2[3⁄4, 1] = m1m1 and so on. At the kth iteration, mass is partitioned into 2^k intervals of length 2^-k.

Let us introduce the notion of the binary expansion $0\beta_1\beta_2...\beta_k$ for the point $x = \beta_1 2^{-1} + \beta_2 2^{-2} + \beta_k 2^{-k}$ where 0 ≤ x ≤1 and 0 < Bk < 1. Then the measure for every dyadic interval $I_{0\beta 1\beta 2...\beta k}$ of length 2^-k equals

$$\mu_{0\beta 1\beta 2...\beta k} = \prod_{i=1}^{k} m\beta_i = m_0{}^n m_1{}^{k-n} \quad (97)$$

where n is the number of digits 0 in the address $0\dot{\beta}_1\beta_2...\beta_k$ of the interval's left end, and (k - n) is the number of digits 1. Since the subinterval mass is preserved at every step, the cascade is called conservative or microcanonical.

Figure 23: Binomial Measure Multifractal. Mo=0.6, 5 iterations

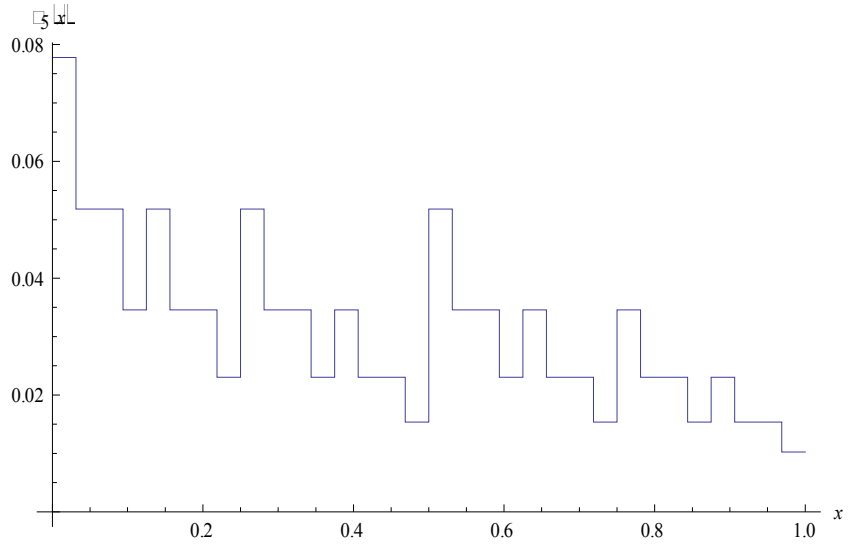


Figure 23 illustrates the simplest multifractal: Mandelbrot's binomial measure on the interval [0,1]. The iteration begins with a uniform distribution μ0[0,1]=1 (with 0<m0<1), subdivides it into a distribution with μ1[0,1/2]=m0 and μ2[1/2,1]=m1=1-m0 , further subdivides it into μ2[0,1/4]=m0m0, μ2[1/4,1/2]=m0m1, μ2[1/2,3/4]=m1m0, μ2[3/4,1]=m1m1 and so on. Additional iteration of this procedure gives a multiplicative cascade that generates an infinite sequence of measures; the limit of the measures is the binomial measure.

The multifractal spectrum of the binomial cascade equals

$$f(\alpha) = -\frac{\alpha_{\max} - \alpha}{\alpha_{\max} - \alpha_{\min}} \log_2\left(\frac{\alpha_{\max} - \alpha}{\alpha_{\max} - \alpha_{\min}}\right) - \frac{\alpha - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}} \log_2\left(\frac{\alpha - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}\right) \quad (98)$$

The distribution (98) is confined with the interval [αmin, αmax]. If m0 ≥ 0:5, then αmin = -log2 (m0) and αmax = -log2 (1 - m0). The binomial cascade can be generalized in two directions. First, one

can introduce a multinomial cascade by increasing the number of subintervals to N > 2. Note that the condition

$$\sum_{0}^{N-1} m_i = 1 \quad (99)$$

is needed for preserving the conservative character of the cascade. Secondly, the values of mi can be randomized rather than assigned fixed values. A cascade with randomized mi is called canonical. In this case, the condition (99) is satisfied only on average, that is

$$E\left[\sum_{0}^{N-1} m_i\right] = 1 \quad (100)$$

An example of the randomized cascade that has an explicit expression for the multifractal spectrum is the lognormal cascade.[73] In this process, the multiplier that distributes the mass of the interval, M, is determined with the lognormal distribution (i.e., log2 (M) is drawn from the Gaussian distribution). If the Gaussian mean and variance are l and s, respectively, then the conservative character of the cascade E[M] = 0.5 is preserved when

$$\sigma^2 = 2(\lambda - 1)/\ln(2) \quad (101)$$

The multifractal spectrum of the lognormal cascade that satisfies (101) equals

$$f(\alpha) = 1 - \frac{(\alpha - \lambda)^2}{4(\lambda - 1)} \quad (102)$$

Note that in contrast to the binomial cascade, the lognormal cascade may yield negative values of f(α), which requires interpretation of f (α) other than the fractal dimension.

Innovation of multifractal process, ΔX = X(t + Δt) - X(t),, is described with the scaling rule

$$\boldsymbol{E}\left[\left|\Delta \boldsymbol{X}\right|^{\boldsymbol{q}}\right] = \boldsymbol{c}(\boldsymbol{q})(\Delta \boldsymbol{t})^{\tau(\boldsymbol{q})+1} \quad (103)$$

Where c(q) and $\tau(\boldsymbol{q})$ (so-called scaling function) are deterministic functions of q. It can be shown that the scaling function $\tau(\boldsymbol{q})$ is always concave. Obviously $\tau(0) = -1$. A self-affine process (88) can be treated as a multifractal process with $\tau(\boldsymbol{q}) = \boldsymbol{H}\boldsymbol{q} - 1$. In particular, for the Wiener processes, H = 1⁄2 and $\tau_w(\boldsymbol{q}) = \boldsymbol{q}/2 - 1$. The scaling function of the binomial cascade can be expressed in terms of its multipliers

$$\tau(q) = \log_2\left(m_o^q + m_1^q\right) \quad (104)$$

The scaling function $\tau(\boldsymbol{q})$ is related to the multifractal spectrum f(α) via the Legendre transformation

$$\tau(q) = \min_{\alpha}\left[q\alpha - f(\alpha)\right] \quad (105)$$

which is equivalent to

$$f(\alpha) = \arg\min_{q}\left[q\alpha - \tau(q)\right] \quad (106)$$

Note that $f(\alpha) = q(\alpha - H) + 1$ for the self-affine processes. In practice, the scaling function of a multifractal process X(t) can be calculated using so-called partition function

$$S_q(T, \Delta t) = \sum_0^{N-1} |X(t + \Delta t) - X(t)|^q \quad (107)$$

where the sample X(t) has N points within the interval [0, T] with the mesh size Δt. It follows from (103) that

$$\log\{E[S_q(T, \Delta t)]\} = \tau(q)\log(\Delta t) + c(q)\log T \quad (108)$$

Thus, plotting log {E[Sq(T, Δt)]} against log (Δt) for different values of q reveals the character of the scaling function $\tau(q)$.

4.6.3 Multifractal model of asset returns

According to MMAR the logarithmic price P(t) is assumed to follow a compound process consisting of a fractional Brownian motion B_H and $\theta(t)$:

$$P(t) = B_H(\theta(t)) \quad (109)$$

Here BH represents a monofractal process which is a sum of random variables sampled by c.d.f. of a multifractal measure. Both B_H and $\theta(t)$ are independent. A crucial role in the considered process plays the virtual trading time which can be interpreted as a deformation of the homogeneous clock-time or as a local volatility corresponding to faster or slower trading.

The linear correlation of P(t) depends on the Hurst exponent H fully characterizing the Brownian motion, whereas the multifractal properties are generated by a multiplicative cascade. It has to be noted that in the original formalism of Mandelbrot & Calvet[74] the whole cascade is generated globally at the same moment for each level k. However, for the sake of prediction we need an iterative procedure that is able to differentiate past and future events. This is the rationale behind the application of a multiplicative measure proposed by Lux.[75]

Instead of θ it is better to consider its increments θ′(t) expressed by

$$\theta'(t) = 2^k \prod_{i=1}^{k} m_t(t) \quad (110)$$

where 2k is a normalizing factor and mi is a random multiplier taken from the log-normal distribution in accordance with the formula

$$m_{t+1}^i = \begin{cases} \exp(N(-\lambda \ln 2, 2(\lambda - 1)\ln 2)) \\ m_t^{(i)} \end{cases} \quad (111)$$

where i = 1, ..., k and N(μ, σ2) denotes the conventional normal distribution.

The upper option is taken either with the probability $2^{-(k-i)}$ or if for any preceding i this option has already been chosen. Otherwise the multiplier remains the same as for previous t. We can imitate in this way the structure of a binary cascade and, on average, preserve its essential features.

Based on this construction we see that in order to describe the multifractal properties of such a cascade we need only one parameter λ. Theoretical multifractal spectrum is then given by

$$f(\alpha) = 1 - \frac{(\alpha - \lambda)^2}{4(\lambda - 1)} \quad (112)$$

This formula implies that f(α) is symmetric and has a maximum localized at α = λ. Under the above formalism a return can be viewed as a compositionof local volatility and white noise:

$$x(t) = \sqrt{\theta'(t)}\sigma N(0,1) = \sqrt{2^k \prod_i^k m_t(t)}\sigma N(0,1) \quad (113)$$

4.6.4 Markov switching multifractal

In financial econometrics, the Markov-switching multifractal (MSM)[76] is a model of asset returns that incorporates stochastic volatility components of heterogeneous durations. MSM captures the outliers, log-memory-like volatility persistence and power variation of financial returns. In currency and equity series, MSM compares favorably with standard volatility models such as GARCH(1,1) and FIGARCH both in- and out-of-sample. MSM is used by practitioners in the financial industry to forecast volatility, compute value-at-risk, and price derivatives.
In continuous time the price process follows the diffusion:

$$\frac{dP_t}{P_t} = \mu dt + \sigma(M_t) dW_t \quad (114)$$

where $\sigma(M_t) = \bar{\sigma}(M_{1,t} \ldots\ldots M_{\bar{k},t})^{\frac{1}{2}}$, W_t is a standard Brownian motion, and μ and $\bar{\sigma}$ are constants. Each component follows the dynamics: $M_{k,t}$ drawn from distribution M with probability $\gamma_k dt$ and $M_{k,t+dt} = M_{k,t}$ with probability $1 - \gamma_k dt$.

The intensities vary geometrically with k:

$$\gamma_k = \gamma_1 b^{k-1} \quad (115)$$

When the number of components $\bar{k}$ goes to infinity, continuous-time MSM converges to a multifractal diffusion, whose sample paths take a continuum of local Hölder exponents on any finite time interval.

MSM often provides better volatility forecasts than some of the best traditional models both in and out of sample. Calvet and Fisher report considerable gains in exchange rate volatility forecasts at horizons of 10 to 50 days as compared with GARCH(1,1), Markov-Switching GARCH, and Fractionally Integrated GARCH . Lux obtains similar results using linear predictions.

4.7. Quantum Finance

Quantum theory can be used to model secondary financial markets. Contrary to stochastic descriptions, the formalism emphasizes the importance of trading in determining the value of a security. All possible realizations of investors holding securities and cash are taken as the basis of the Hilbert space of market states. The temporal evolution of an isolated market is unitary in this space. Linear operators representing basic financial transactions such as cash transfer and the buying or selling of securities are constructed and simple model Hamiltonians that generate the temporal evolution due to cash flows and the trading of securities are proposed.
The Hamiltonian describing financial transactions becomes local when the profit/loss from trading is small compared to the turnover. This approximation may describe a highly liquid and efficient stock market. The lognormal probability distribution for the price of a stock with a variance that is proportional to the elapsed time is reproduced for an equilibrium market. The asymptotic volatility of a stock in this case is related to the long-term probability that it is traded. The author of this thesis is researching with Martin Schaden on this exciting new field.

4.8. State Space representation

There is another representation of time series called state-space models. As we will see in this section, state-space models are equivalent to ARMA models. While the latter are typical of econometrics, state-space models originated in the domain of engineering and system analysis. Consider a system defined for t ≥ 0 and described by the following set of linear difference equations:

$$\begin{cases} \mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t + \mathbf{B}\mathbf{u}_t \\ \mathbf{x}_t = \mathbf{C}\mathbf{z}_t + \mathbf{D}\mathbf{u}_t + \mathbf{E}\mathbf{s}_t \end{cases} \quad (116)$$

Where:

$\boldsymbol{x}_t$ = and n dimensional vector

z_t = a k dimensional vector

$\boldsymbol{u}_t$ = an m-dimensional vector

$\boldsymbol{s}_t$ = a k-dimensional vector

A = a k×k matrix

B = a k×m matrix

C = an n×k matrix

D = an n×m matrix

E = an n×k matrix

In the language of system theory, the variables $\boldsymbol{u}_t$ are called the inputs of the system, the variables $\boldsymbol{z}_t$ are called the state variables of the system, and the variables $\boldsymbol{x}_t$ are called the observations or outputs of the system, and $\boldsymbol{s}_t$ are deterministic terms that describe the deterministic components if they exist.

The system is formed by two equations. The first equation is a purely autoregressive AR(1) process that describes the dynamics of the state variables. The second equation is a static regression of the observations over the state variables, with inputs as innovations. Note that in this state-space representation the inputs ut are the same in both equations. It is possible to reformulate state space models with different, independent inputs for the states, and the observables. The two representations are equivalent.

The fact that the first equation is a first order equation is not restrictive as any AR(p) system can be transformed into a first-order AR(1) system by adding variables. The new variables are defined as the lagged values of the old variables. This can be illustrated in the case of a single second-order autoregressive equation:

$$\boldsymbol{X}_{t+1} = \alpha_0 \boldsymbol{X}_t + \alpha_1 \boldsymbol{X}_{t-1} + \varepsilon_{t+1} \quad (117)$$

Define $\boldsymbol{Y}_t = \boldsymbol{X}_{t+1}$.

The previous equation is then equivalent to the firstorder system:

$$\begin{aligned} \boldsymbol{X}_{t+1} &= \alpha_0 \boldsymbol{X}_t + \alpha_1 \boldsymbol{X}_{t-1} + \varepsilon_{t+1} \\ \boldsymbol{Y}_t &= \boldsymbol{X}_{t+1} \end{aligned} \quad (118)$$

This transformation can be applied to systems of any order and with any number of equations.

Note that this state-space representation is not restricted to white noise inputs. A state-space representation is a mapping of inputs into outputs. Given a realization of the inputs ut and an initial state z_0, the realization of the outputs xt is fixed. The state-space representation can be seen as a black-box, characterized by A, B, C, D, and z_0 that maps any m-dimensional input sequence into an n-dimensional output sequence. The mapping S = S(A,B,C,D, z_0) of u → x is called a black-box representation in system theory. State-space representations are not unique. Given a state-space representation, there are infinite other state-space representations that implement the same mapping u → x. In fact, given any non-singular (invertible) matrix Q, it can be easily verified that

$$\boldsymbol{S}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},z_0) = \boldsymbol{S}\left(\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^{-1},\boldsymbol{Q}\boldsymbol{B},\boldsymbol{C}\boldsymbol{Q}^{-1},\boldsymbol{D},\boldsymbol{Q}z_0\right) \quad (119)$$

Any two representations that satisfy the above condition are called equivalent. The minimal size of a system that admits a state-space representation is the minimum possible size k of the state vector. A representation is called minimal if its state vector has size k. We can now establish the connection between state-space and infinite moving-average representations and the equivalence of ARMA and state-space representations. Consider a n-dimensional process $\boldsymbol{x}_t$, which admits an infinite moving-average representation

$$\boldsymbol{x}_t = \sum_{i=0}^{\infty} \boldsymbol{H}_i \boldsymbol{\varepsilon}_{t-i} \quad (120)$$

where $\boldsymbol{\varepsilon}_t$ is an n-dimensional, zero-mean, white noise process with non-singular variance-covariance matrix Ω and $\boldsymbol{H}_0 = \boldsymbol{I}$, or a linear moving average model

$$\boldsymbol{x}_t = \sum_{i=0}^{\infty} \boldsymbol{H}_i \boldsymbol{\varepsilon}_{t-i} + \boldsymbol{h}(\boldsymbol{t})\boldsymbol{z} \quad (121)$$

It can be demonstrated that this system admits the state-space representation:

$$\begin{cases} \mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t + \mathbf{B}\boldsymbol{\varepsilon}_t \\ \mathbf{x}_t = \mathbf{C}\mathbf{z}_t + \mathbf{D}\boldsymbol{\varepsilon}_t \end{cases} \quad (122)$$

if and only if its Hankel matrix is of finite rank. In other words, a time series which admits an infinite moving-average representation and has a Hankel matrix of finite rank can be generated by a state-space system where the inputs are the noise. Conversely, a state-space system with white-noise as inputs generates a series that can be represented as an infinite moving-average with a Hankel matrix of finite rank. This conclusion is valid for both stationary and non-stationary processes.

We have seen in the previously that a time series which admits an infinite moving-average representation can also be represented as an ARMA process if and only if its Hankel matrix is of finite rank. Therefore we can conclude that a time series admits an ARMA representation if and only if it admits a state-space representation. ARMA and state space representations are equivalent. The standard method for state space estimation is the Kalman Filter.

Kalman filters are based on linear dynamical systems discretized in the time domain. They are modelled on a Markov chain built on linear operators perturbed by Gaussian noise. The state of the system is represented as a vector of real numbers. At each discrete time increment, a linear operator is applied to the state to generate the new state, with some noise mixed in, and optionally some information from the controls on the system if they are known.

Then, another linear operator mixed with more noise generates the observed outputs from the true ("hidden") state. The Kalman filter may be regarded as analogous to the hidden Markov model, with the key difference that the hidden state variables take values in a continuous space (as opposed to a discrete state space as in the hidden Markov model). Additionally, the hidden Markov model can represent an arbitrary distribution for the next value of the state variables, in contrast to the Gaussian noise model that is used for the Kalman filter. There is a strong duality between the equations of the Kalman Filter and those of the hidden Markov model. A review of this and other models is given in Roweis and Ghahramani (1999).[77]

In order to use the Kalman filter to estimate the internal state of a process given only a sequence of noisy observations, one must model the process in accordance with the framework of the Kalman filter. This means specifying the following matrices: $\boldsymbol{F}_k$, the state-transition model; $\boldsymbol{H}_k$, the observation model; $\boldsymbol{Q}_k$, the covariance of the process noise; $\boldsymbol{R}_k$, the covariance of the

observation noise; and sometimes $\boldsymbol{B}_k$, the control-input model for each time-step, k, as described below.

The Kalman filter model assumes the true state at time k is evolved from the state at (k − 1) according to:

$$\boldsymbol{x}_k = \boldsymbol{F}_k \boldsymbol{x}_{k-1} + \boldsymbol{B}_k \boldsymbol{u}_k + \boldsymbol{w}_k \quad (123)$$

Where

$\boldsymbol{F}_k$ is the state transition model which is applied to the previous state $\boldsymbol{x}_{k-1}$

$\boldsymbol{B}_k$ is the control-input model which is applied to the control vector $\boldsymbol{u}_k$

$\boldsymbol{w}_k$ is the process noise which is assumed to be drawn from a zero mean multivariate normal distribution with covariance $\boldsymbol{Q}_k$. $\boldsymbol{w}_k \sim N(0, \boldsymbol{Q}_k)$.

At time k an observation (or measurement) $\boldsymbol{z}_k$ of the true state $\boldsymbol{x}_k$ is made according to

$$\boldsymbol{z}_k = \boldsymbol{H}_k \boldsymbol{x}_k + \boldsymbol{v}_k \quad (124)$$

where $\boldsymbol{H}_k$ is the observation model which maps the true state space into the observed space and $\boldsymbol{v}_k$ is the observation noise which is assumed to be zero mean Gaussian white noise with covariance $\boldsymbol{R}_k$ $\boldsymbol{v}_k \sim N(0, \boldsymbol{R}_k)$

The initial state, $\{\boldsymbol{x}_0, \boldsymbol{w}_1, ...\boldsymbol{w}_k, \boldsymbol{v}_1, ...\boldsymbol{v}_k\}$ and the noise vectors at each step are all assumed to be mutually independent.

Many real dynamical systems do not exactly fit this model. In fact, unmodelled dynamics can seriously degrade the filter performance, even when it was supposed to work with unknown stochastic signals as inputs. The reason for this is that the effect of unmodelled dynamics depends on the input, and, therefore, can bring the estimation algorithm to instability (it diverges). On the other hand, independent white noise signals will not make the algorithm diverge. The problem of separating between measurement noise and unmodelled dynamics is a difficult one and is treated in control theory under the framework of robust control.

4.9. Bayesian statistics

Statistical analysis is employed from the vantage point of either of the two main statistical philosophical traditions—‘‘frequentist’’ and ‘‘Bayesian.’’ An important difference between the two, lies with the interpretation of the concept of probability. As the name suggests, advocates of frequentist statistics adopt a frequentist interpretation: The probability of an event is the limit of its long-run relative frequency (i.e., the frequency with which it occurs as the amount of data increases without bound). Strict adherence to this interpretation is not always possible in practice. When studying rare events, for instance, large samples of data may not be available and in such cases proponents of frequentist statistics resort to theoretical results. The Bayesian view of the

world is based on the subjectivist interpretation of probability: Probability is subjective, a degree of belief that is updated as information or data is acquired. Closely related to the concept of probability is that of uncertainty.

Proponents of the frequentist approach consider the source of uncertainty to be the randomness inherent in realizations of a random variable. The probability distributions of variables are not subject to uncertainty. In contrast, Bayesian statistics treats probability distributions as uncertain and subject to modification as new information becomes available. Uncertainty is implicitly incorporated by probability updating. The probability beliefs based on the existing knowledge base take the form of the prior probability.

The posterior probability represents the updated beliefs. Since the beginning of last century, when quantitative methods and models became a mainstream tool to aid in understanding financial markets and formulating investment strategies, the framework applied in finance has been the frequentist approach. The term ‘‘frequentist’’ usually refers to the Fisherian philosophical approach named after Sir Ronald Fisher.

Strictly speaking, ‘‘Fisherian’’ has a broader meaning as it includes not only frequentist statistical concepts such as unbiased estimators, hypothesis tests, and confidence intervals, but also the maximum likelihood estimation framework pioneered by Fisher. Only in the last two decades has Bayesian statistics started to gain greater acceptance in financial modelling, despite its introduction about 250 years ago by Thomas Bayes, a British minister and mathematician. It has been the advancements of computing power and the development of new computational methods that has fostered the growing use of Bayesian statistics in finance.

The concept of subjective probability is derived from arguments for rationality of the preferences of agents. It originated in the 1930s with the (independent) works of Bruno de Finetti and Frank Ramsey, and was further developed by Leonard Savage and Dennis Lindley. The subjective probability interpretation can be traced back to the Scottish philosopher and economist David Hume, who also had philosophical influence over Harry Markowitz (by Markowitz’s own words in his autobiography.

The three steps of Bayesian decision making are:

1. Formulating the prior probabilities to reflect existing information.
2. Constructing the quantitative model, taking care to incorporate the uncertainty intrinsic in model assumptions.
3. Selecting and evaluating a utility function describing how uncertainty affects alternative model decisions.

Bernardo and Smith (1994)[78] offer the following definition of bayesian statistics:

“Bayesian statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterizing how an individual shoud act in order to avoid certin kinds of undesirable behavioural inconsistencies. The theory establishes that expected utility maximization provides the basis for rational decision making and that Bayes’ Theorem provides the key to the ways in which beliefs should fit together in the light of changing evidence”

4.9.1 The likelihood function

Suppose we are interested in analyzing the returns on a given stock and have available a historical record of returns. Any analysis of these returns, beyond a very basic one, would require that we make an educated guess about (propose) a process that might have generated these return data. Assume that we have decided on some statistical distribution and denote it by

$$p(y|\theta) \quad (125)$$

where y is a realization of the random variable Y (stock return) and θ is a parameter specific to the distribution, p. Assuming that the distribution we proposed is the one that generated the observed data, we draw a conclusion about the value of θ. Obviously, central to that goal is our ability to summarize the information contained in the data. The likelihood function is a statistical construct with this precise role. Denote the n observed stock returns by $y_1, y_2, ...y_n$. The joint density function of Y, for a given value of θ, is

$$f(y_1, y_2, ...y_n|\theta) \quad (126)$$

We can observe that the function above can also be treated as a function of the unknown parameter, θ, given the observed stock returns. That function of θ is called the likelihood function. We write it as

$$L(\theta|y_1, y_2, ...y_n) = f(y_1, y_2, ...y_n|\theta) \quad (127)$$

Suppose we have determined from the data two competing values of θ, θ1 and θ2, and want to determine which one is more likely to be the true value (at least, which one is closer to the true value). The likelihood function helps us make that decision. Assuming that our data were indeed generated by the distribution in (126), θ1 is more likely than θ2 to be the true parameter value whenever $L(y_1, y_2, ...y_n|\theta_1) > L(y_1, y_2, ...y_n|\theta_2)$. This observation provides the intuition behind the method most often employed in ‘‘classical’’ statistical inference to estimate θ from the data alone—the method of maximum likelihood. The value of θ most likely to have yielded the observed sample of stock return data, y1, y2, . . . , yn, is the maximum likelihood estimate, $\hat{\theta}$, obtained from maximizing the likelihood function in (127).

To illustrate the concept of a likelihood function, we briefly discussed an example based on the Poisson distribution (a discrete distribution).

4.9.2 The Poisson Distribution Likelihood Function

The Poisson distribution is often used to describe the random number of events occurring within a certain period of time. It has a single parameter, θ, indicating the rate of occurrence of the random event, that is, how many events happen on average per unit of time. The probability distribution of a Poisson random variable, X, is described by the following expression:

$$p(X = k) = \frac{\theta^k}{k!} e^{-\theta}, \quad k = 1,2,... \quad (128)$$

Suppose we are interested to examine the annual number of defaults of North American corporate bond issuers and we have gathered a sample of data for the period from 1986 through 2005. Assume that these corporate defaults occur according to a Poisson distribution. Denoting the 20 observations by $x_1, x_2, ...x_{20}$, we write the likelihood function for the Poisson parameter θ (the average rate of defaults) as

$$L(\theta|y_1, y_2, ...x_{20}) = \prod_{i=1}^{20} p(X = x_i|\theta) = \prod_{i=1}^{20} \frac{\theta^{x_i}}{x_i!} e^{-\theta} = \\ = \frac{\theta^{\sum_{i=1}^{20} x_i}}{\prod_{i=1}^{20} x_i!} e^{-20\theta} \quad (129)$$

It is often customary to retain in the expressions for the likelihood function and the probability distributions only the terms that contain the unknown parameter(s); that is, we get rid of the terms that are constant with respect to the parameter(s). Thus, (130) could be written as

$$L(\theta|y_1, y_2, ...x_{20}) \propto \theta^{\sum_{i=1}^{20} x_i} e^{-20\theta} \quad (130)$$

where α denotes ‘‘proportional to.’’ Clearly, for a given sample of data, the expressions in (129) and (130) are proportional to each other and therefore contain the same information about θ. Maximizing either of them with respect to θ, we obtain that the maximum likelihood estimator of the Poisson parameter, θ, is the sample mean, $\bar{x}$:

$$\hat{\theta} = \bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} \quad (131)$$

For the 20 observations of annual corporate defaults, we get a sample mean of 51.6. The Poisson probability distribution function (evaluated at θ equal to its maximum-likelihood estimate, $\hat{\theta} = 51.6$.

4.9.3 Bayes theorem

Consider a probability space $[S, \tilde{A}, P(\cdot)]$ and a collection $B_n \in \tilde{A}\ (n = 1,2,..., N)$ of mutually disjoint events such that $P(B_n) > 0\ (n = 1,2,...., N)$ and $B_1 \cup B_2 \cup ... \cup B_N = S$. Then

$$P(B_n|A) = \frac{P(A|B_n)P(B_n)}{\sum_{j=1}^{N} P(A|B_j)P(B_j)}\ (n = 1,2,...., N) \quad (132)$$

for every $A \in \tilde{A}$ such that $P(A) > 0$.

Bayes’ Theorem and Model Selection

The usual approach to modelling of a financial phenomenon is to specify the analytical and distributional properties of a process that one thinks generated the observed data and treat this process as if it were the true one. Clearly, in doing so, one introduces a certain amount of error

into the estimation process. Accounting for model risk might be no less important than accounting for (within-model) parameter uncertainty, although it seems to preoccupy researchers less often.
One usually entertains a small number of models as plausible ones. The idea of applying the Bayes' theorem to model selection is to combine the information derived from the data with the prior beliefs one has about the degree of model validity. One can then select the single ''best'' model with the highest posterior probability and rely on the inference provided byit or one can weigh the inference of each model by its posterior probability and obtain an ''averaged-out'' conclusion.

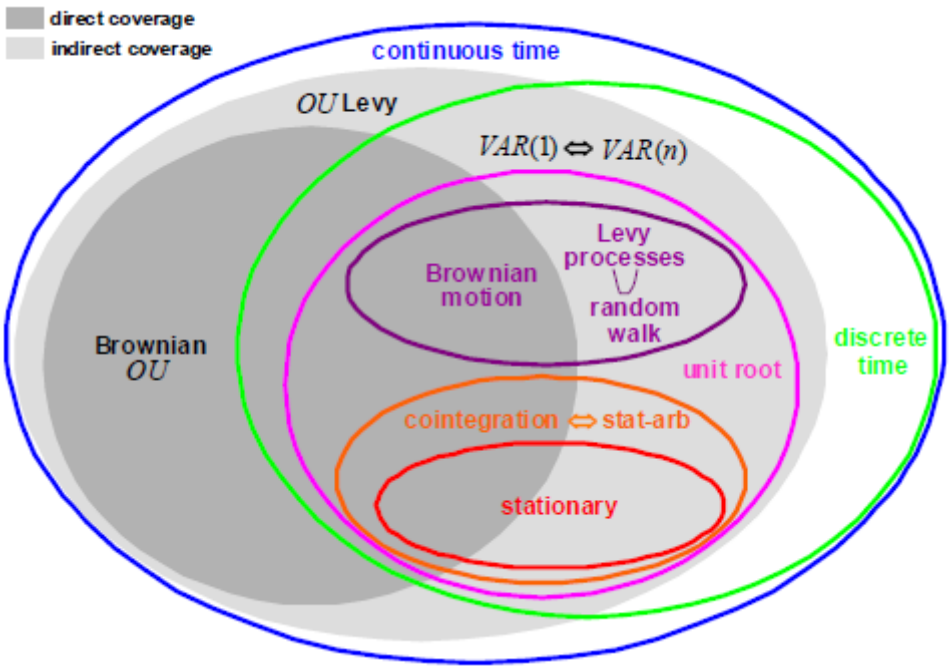
Bayes' Theorem and Classification

Classification refers to assigning an object, based on its characteristics, into one out of several categories. It is most often applied in the area of credit and insurance risk, when a creditor (an insurer) attempts to determine the creditworthiness (riskiness) of a potential borrower (policyholder). Classification is a statistical problem because of the existence of information asymmetry—the creditor's (insurer's) aim is to determine with very high probability the unknown status of the borrower (policyholder).

5. Statistical Arbitrage, Cointegration, and Multivariate Ornstein-Uhlenbeck

We introduce the multivariate Ornstein-Uhlenbeck process and discuss how it generalizes a vast class of continuous-time and discrete-time multivariate processes. Relying on the simple geometrical interpretation of the dynamics of the Ornstein-Uhlenbeck process we introduce cointegration and its relationship to statistical arbitrage. We illustrate an application to swap contract strategies.
The multivariate Ornstein-Uhlenbeck process is arguably the model most utilized by academics and practitioners alike to describe the multivariate dynamics of financial variables. Indeed, the Ornstein-Uhlenbeck process is parsimonious, and yet general enough to cover a broad range of processes. Therefore, by studying the multivariate Ornstein-Uhlenbeck process we gain insight into the properties of the main multivariate features used daily by econometricians. continuous time discrete

Figure 24: Multivariate processes and coverage by OU



Indeed, the following relationships hold, refer to Figure 24 and see below for a proof. The Ornstein-Uhlenbeck is a continuous time process. When sampled in discrete time, the Ornstein-Uhlenbeck process gives rise to a vector autoregression of order one, commonly denoted by VAR(1). More general VAR(n) processes can be represented in VAR(1) format, and therefore they are also covered by the Ornstein-Uhlenbeck process. VAR(n) processes include unit root processes, which in turn include the random walk, the discrete-time counterpart of Levy processes. VAR(n) processes also include cointegrated dynamics, which are the foundation of statistical arbitrage. Finally, stationary processes are a special case of cointegration.

In Section 5.1 we derive the analytical solution of the Ornstein-Uhlenbeck process. In Section 5.2 we discuss the geometrical interpretation of the solution.

Building on the solution and its geometrical interpretation, in Section 5.3 we introduce naturally the concept of cointegration and we study its properties.

In Section 5.4 we discuss a simple model-independent estimation technique for cointegration and we apply this technique to the detection of mean-reverting trades, which is the foundation of statistical arbitrage.

5.1. The multivariate Ornstein-Uhlenbeck process

The multivariate Ornstein-Uhlenbeck process is defined by the following stochastic differential equation

$$d\boldsymbol{X}_t = -\boldsymbol{\Theta}\left(\boldsymbol{X}_t - \pi\right)dt + \boldsymbol{S}d\boldsymbol{B}_t \quad (133)$$

In this expression $\boldsymbol{\Theta}$ is the transition matrix, namely a fully generic square matrix that defines the deterministic portion of the evolution of the process; μ is a fully generic vector, which represents the unconditional expectation when this is defined, see below; S is the scatter generator, namely a fully generic square matrix that induces the dispersion of the process; $\boldsymbol{B}_t$ is typically assumed to be a vector of independent Brownian motions, although more in general it is a vector of independent Levy processes, see Barndorff-Nielsen and Shephard (2001)[79] and refer to Figure 24.

To integrate the process (133) we introduce the integrator

$$\boldsymbol{Y}_t \equiv \boldsymbol{e}^{\Theta t}\left(\boldsymbol{X}_t - \pi\right) \quad (134)$$

Using Ito's lemma we obtain

$$\boldsymbol{dY}_t \equiv \boldsymbol{e}^{\Theta t}\boldsymbol{SdB}_t \quad (135)$$

Integrating both sides and substituting the definition (134) we obtain

$$\boldsymbol{X}_{t+\tau} = \left(\boldsymbol{I} - \boldsymbol{e}^{-\Theta\tau}\right)\pi + \boldsymbol{e}^{-\Theta\tau}\boldsymbol{X}_t + \varepsilon_{t,\tau} \quad (136)$$

where the invariants are mixed integrals of the Brownian motion and are thus normally distributed

$$\varepsilon_{t,\tau} \equiv \int_t^{t+\tau} \boldsymbol{e}^{\Theta(u-\tau)}\boldsymbol{SdB}_u \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma\tau}\right) \quad (137)$$

The solution (136) is a vector autoregressive process of order one VAR(1), which reads

$$X_{t+\tau} = c + CX_t + \varepsilon_{t,\tau} \quad (138)$$

for a conformable vector and matrix c and C respectively. A comparison between the integral solution (136) and the generic VAR(1) formulation (138) provides the relationship between the continuous-time coefficients and their discrete-time counterparts.

The conditional distribution of the Ornstein-Uhlenbeck process (136) is normal at all times

$$X_{t+\tau} \sim N\left(x_{t+\tau}, \Sigma_\tau\right) \quad (139)$$

The deterministic drift reads

$$x_{t+\tau} \equiv \left(I - e^{-\theta\tau}\right)\mu + e^{-\theta\tau} x_t \quad (140)$$

and the covariance can be expressed as in Van der Werf (2007)[80] in terms of the stack operator vec and the Kronecker sum $\oplus$ as

$$vec(\Sigma_\tau) \equiv (\Theta \oplus \Theta)^{-1}\left(I - e^{-(\Theta\oplus\Theta)\tau}\right)vec(\Sigma) \quad (141)$$

where $\Sigma \equiv \mathbf{S}\mathbf{S}'$, see the proof in Appendix 17.9. Formulas (140)-(141) describe the propagation law of risk associated with the Ornstein-Uhlenbeck process: the location-dispersion ellipsoid defined by these parameters provides an indication of the uncertainly in the realization of the next step in the process.

Notice that (140) and (141) are defined for any specification of the input parameters $\boldsymbol{\Theta}$, $\boldsymbol{\mu}$, and $\boldsymbol{S}$ in (98). For small values of τ, a Taylor expansion of these formulas shows that

$$X_{t+\tau} \approx X_t + \varepsilon_{t,\tau} \quad (142)$$

where $\varepsilon_{t,\tau}$ is a normal invariant:

$$\varepsilon_{t,\tau} \sim N(\tau\Theta\mu, \tau\Sigma) \quad (143)$$

In other words, for small values of the time step τ the Ornstein-Uhlenbeck process is indistinguishable from a Brownian motion, where the risk of the invariants (143), as represented by the standard deviation of any linear combination of its entries, displays the classical "square-root of τ "propagation law.

As the step horizon τ grows to infinity, so do the expectation (140) and the covariance (141), unless all the eigenvalues of $\boldsymbol{\Theta}$ have positive real part. In that case the distribution of the process stabilizes to a normal whose unconditional expectation and covariance read

$$x_\infty = \mu \quad (144)$$

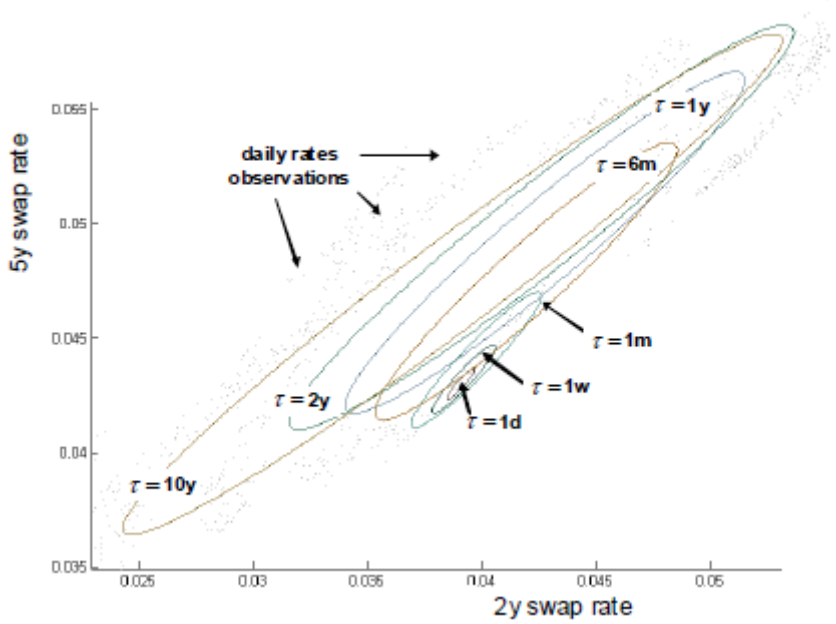
$$vec(\Sigma_\infty) \equiv (\Theta \oplus \Theta)^{-1} vec(\Sigma) \quad (145)$$

To illustrate, we consider the bivariate case of the two-year and the ten-year par swap rates. The benchmark assumption among buy-side practitioners is that par swap rates evolve as the random walk (142).

However, rates cannot diffuse indefinitely. Therefore, they cannot evolve as a random walk for any size of the time step τ : for steps of the order of a month or larger, mean-reverting effects must become apparent.

The Ornstein-Uhlenbeck process is suited to model this behavior. We fit this process for different values of the time step τ and we display in Figure 25 the location-dispersion ellipsoid defined by the expectation (140) and the covariance (141).

Figure 25: Propagation law of risk for OU process fitted to swap rates



For values of τ of the order of a few days, the drift is linear in the step and the size increases as the square root of the step, as in (143). As the step increases and mean-reversion kicks in, the ellipsoid stabilizes to its unconditional values (144)-(145).

5.2. The geometry of the Ornstein-Uhlenbeck dynamics

The integral (136) contains all the information on the joint dynamics of the Ornstein-Uhlenbeck process (133). However, that solution does not provide any intuition on the dynamics of this process. In order to understand this dynamics we need to observe the Ornstein-Uhlenbeck process in a different set of coordinates.

Consider the eigenvalues of the transition matrix $\boldsymbol{\Theta}$: since this matrix has real entries, its eigenvalues are either real or complex conjugate: we denote them respectively by $(\lambda_1,...,\lambda_K)$ and $((\gamma_1 \boldsymbol{i}\boldsymbol{w}_1),...,(\gamma_J \boldsymbol{i}\boldsymbol{w}_J))$ where K + 2J = N . Now consider the matrix **B** whose columns are the respective, possibly complex, eigenvectors and define the real matrix $\mathbf{A} \equiv \mathbf{Re}(\mathbf{B}) - \mathbf{Im}(\mathbf{B})$. Then the transition matrix can be decomposed in terms of eigenvalues and eigenvectors as follows

$$\mathbf{\Theta} \equiv \mathbf{A}\mathbf{\Gamma}A^{-1} \quad (146)$$

where Γ is a block-diagonal matrix

$$\mathbf{\Gamma} \equiv diag\left(\lambda_1,....\lambda_K, \mathbf{\Gamma}_1,..., \mathbf{\Gamma}_J\right) \quad (147)$$

and the generic j-th matrix $\mathbf{\Gamma}_J$ is defined as

$$\mathbf{\Gamma} \equiv \begin{pmatrix} \gamma_j & w_j \\ -w_j & \gamma_j \end{pmatrix} \quad (148)$$

With the eigenvector matrix A we can introduce a new set of coordinates

$$\mathbf{z} \equiv \mathbf{A}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right) \quad (149)$$

The original Ornstein-Uhlenbeck process (133) in these coordinates follows from Ito's lemma and reads

$$\mathbf{dZ_t} = -\mathbf{\Gamma Z_t dt} + \mathbf{V dB_t} \quad (150)$$

where $\boldsymbol{V} \equiv \boldsymbol{A}^{-1}\boldsymbol{S}$. Since this is another Ornstein-Uhlenbeck process, its solution is normal

$$\mathbf{Z_t} \sim \mathbf{N}\left(\mathbf{z_t}, \mathbf{\Phi_t}\right) \quad (151)$$

for a suitable deterministic drift $\mathbf{z_t}$ and covariance $\mathbf{\Phi}_t$. The deterministic drift zt is the solution of the ordinary differential equation

$$\mathbf{dz_t} = -\mathbf{\Gamma z_t dt} \quad (152)$$

Given the block-diagonal structure of (147), the deterministic drift splits into separate sub-problems. Indeed, let us partition the N -dimensional vector $\mathbf{z_t}$ into K entries which correspond to the real eigenvalues in (147), and J pairs of entries which correspond to the complex-conjugate eigenvalues summarized by (148):

$$z_t \equiv \left(z_{1,t},....,z_{K,t},z_{1,t}^{(1)},z_{1,t}^{(2)}....,z_{J,t}^{(2)},z_{J,t}^{(2)}\right)' \text{ (153)}$$

For the variables corresponding to the real eigenvalues, (153) simplifies to:

$$dz_{kt} = -\lambda_k z_{k,t} dt \text{ (154)}$$

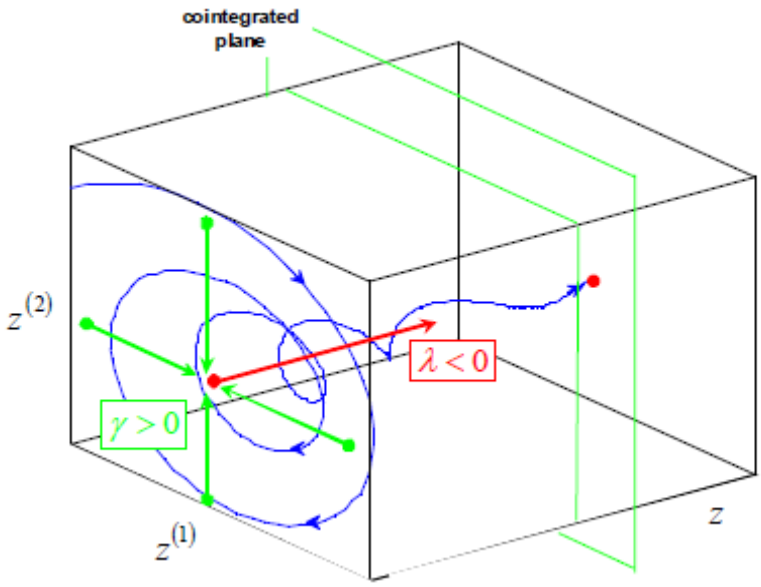
For each real eigenvalue indexed by k = 1, . . . , K the solution reads

$$z_{k;t} = e^{-\lambda_k t} z_{k,0} \text{ (155)}$$

This is an exponential shrinkage at the rate λ_k. Note that (155) is properly defined also for negative values of λ_k, in which case the trajectory is an exponential explosion. If $\lambda_k > 0$ we can compute the half-life of the deterministic trend, namely the time required for the trajectory (155) to progress half way toward the long term expectation, which is zero:

$$\tilde{t} \equiv \frac{\ln 2}{\lambda_k} \text{ (156)}$$

Figure 26: Deterministic drift of OU process



As for the variables among (153) corresponding to the complex eigenvalue pairs (152) simplifies to

$$\mathbf{dz_{j,t}} = -\mathbf{\Gamma_j z_{j,t} dt}, \quad \boldsymbol{j} = 1,...,\boldsymbol{J} \text{ (157)}$$

For each complex eigenvalue, the solution reads formally

$$\mathbf{z_{j;t}} \equiv \mathbf{e}^{-\mathbf{\Gamma_k t}} \mathbf{z_{j,0}} \text{ (158)}$$

This formal bivariate solution can be made more explicit component-wise. First, we write the matrix (148) as follows

$$\mathbf{\Gamma}_j = \gamma_j \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + w_j \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{(159)}$$

As we show in Appendix A.1, the identity matrix on the right hand side generates an exponential explosion in the solution (158), where the scalar γ_j determines the rate of the explosion. On the other hand, the second matrix on the right hand side in (159) generates a clockwise rotation, where the scalar $\boldsymbol{w}_j$ determines the frequency of the rotation. Given the minus sign in the exponential in (158) we obtain an exponential shrinkage at the rate γ_j, coupled with a counterclockwise rotation with frequency $\boldsymbol{w}_j$

$$z_{j;t}^{(1)} \equiv e^{-\gamma_j t}\left(z_{j;0}^{(1)} \cos w_j t - z_{j;0}^{(2)} \sin w_j t\right) \text{(160)}$$

$$z_{j;t}^{(2)} \equiv e^{-\gamma_j t}\left(z_{j;0}^{(1)} \cos w_j t - z_{j;0}^{(2)} \sin w_j t\right) \text{(161)}$$

Again, (160)-(161) are properly defined also for negative values of γ_j, in which case the trajectory is an exponential explosion.

To illustrate, we consider a tri-variate case with one real eigenvalue λ and two conjugate eigenvalues $\gamma + \boldsymbol{iw}$ and $\gamma - \boldsymbol{iw}$, where λ < 0 and γ > 0. In Figure 26 we display the ensuing dynamics for the deterministic drift (153), which in this context reads $\left(z_t, z_t^{(1)}, z_t^2\right)$: the motion (155) escapes toward infinity at an exponential rate, whereas the motion (140)-(141) whirls toward zero, shrinking at an exponential rate.

Once the deterministic drift in the special coordinates $\mathbf{z}_t$ is known, the deterministic drift of the original Ornstein-Uhlenbeck process (136) is obtained by inverting (149). This is an affine transformation, which maps lines in lines and planes in planes

$$x_t \equiv \mu + Az_t \quad (162)$$

Therefore the qualitative behavior of the solutions (153) and (160)-(161) sketched In Figure 26 is preserved.

Although the deterministic part of the process in diagonal coordinates (149) splits into separate dynamics within each eigenspace, these dynamics are not independent. In Appendix 17.9 we derive the quite lengthy explicit formulas for the evolution of all the entries of the covariance $\mathbf{\Phi}_t$ in (151). For instance, the covariances between entries relative to two real eigenvalues reads:

$$\Phi_{k;\tilde{k};t} = \frac{\Phi_{k;\tilde{k}}}{\lambda_k + \lambda_{\tilde{k}}}\left(1 - e^{-(\lambda_k + \lambda_{\tilde{k}})t}\right) \quad (163)$$

where $\mathbf{\Phi} \equiv \mathbf{V}\mathbf{V}'$. More in general, the following observations apply. First, as time evolves, the relative volatilities change: this is due both to the different speed of divergence/shrinkage induced by the real parts λ's and γ's of the eigenvalues, and to different speed of rotation, induced by the imaginary parts ω's of the eigenvalues. Second, the correlations only vary if rotations occur: if the imaginary parts ω's are null, the correlations remain unvaried.

Once the covariance $\mathbf{\Phi}_t$ in the diagonal coordinates is known, the covariance of the original Ornstein-Uhlenbeck process (136) is obtained by inverting (149) and using the affine equivariance of the covariance, which leads to $\mathbf{\Sigma}_t \equiv \mathbf{A}\mathbf{\Phi}_t\mathbf{A}'$.

5.3. Cointegration

The solution (136) of the Ornstein-Uhlenbeck dynamics (133) holds for any choice of the input parameters $\boldsymbol{\mu}$, $\mathbf{\Theta}$ and $\mathbf{S}$. However, from the formulas for the covariances (163), and similar formulas in Appendix 17.9, we verify that if some $\lambda_k \leq 0$, or some $\gamma_k \leq 0$, i.e. if any eigenvalues of the transition matrix $\mathbf{\Theta}$ are null or negative, then the overall covariance of the Ornstein-Uhlenbeck $\boldsymbol{X}_t$ does not converge: therefore $\boldsymbol{X}_t$ is stationary only if the real parts of all the eigenvalues of $\mathbf{\Theta}$ are strictly positive.

Nevertheless, as long as some eigenvalues have strictly positive real part, the covariances of the respective entries in the transformed process (149) stabilizes in the long run. Therefore these processes and any linear combination thereof are stationary. Such combinations are called cointegrated: since from (149) they span a hyperplane, that hyperplane is called the cointegrated space, see Figure 26.

To better discuss cointegration, we write the Ornstein-Uhlenbeck process (136) as

$$\Delta_\tau \mathbf{X}_t = \boldsymbol{\Psi}_\tau \left(\mathbf{X}_t - \boldsymbol{\mu} \right) + \boldsymbol{\varepsilon}_{t,\tau} \quad (164)$$

where Δτ is the difference operator $\Delta_\tau X_t \equiv X_{t+\tau} - X_t$ and $\boldsymbol{\Psi}_\tau$ is the transition matrix

$$\boldsymbol{\Psi}_\tau \equiv \mathbf{e}^{-\boldsymbol{\Theta}\tau} - \mathbf{I} \quad (165)$$

If some eigenvalues in $\boldsymbol{\Theta}$ are null, the matrix $e^{-\boldsymbol{\Theta}\tau}$ has unit eigenvalues: this follows from

$$\mathbf{e}^{-\boldsymbol{\Theta}\tau} = \mathbf{A} \mathbf{e}^{-\boldsymbol{\Gamma}\tau} \mathbf{A}^{-1} \quad (166)$$

which in turn follows from (146). Processes with this characteristic are known as unit-root processes. A very special case arises when all the entries in $\boldsymbol{\Theta}$ are null. In this circumstance, $e^{-\boldsymbol{\Theta}\tau}$ is the identity matrix, the transition matrix $\boldsymbol{\Psi}_\tau$ is null and the Ornstein-Uhlenbeck process becomes a multivariate random walk.

More in general, suppose that L eigenvalues are null. Then $\boldsymbol{\Psi}_\tau$ has rank N − L and therefore it can be expressed as

$$\boldsymbol{\Psi}_\tau \equiv \boldsymbol{\Phi}'_\tau \mathbf{Y}_\tau \quad (167)$$

where both matrices $\boldsymbol{\Phi}_\tau$ and $\mathbf{Y}_\tau$ are full-rank and have dimension (N − L)×N . The representation (167), known as the error correction representation of the process (164), is not unique: indeed any pair $\widetilde{\boldsymbol{\Phi}}_\tau \equiv \mathbf{P}' \boldsymbol{\Phi}_\tau$ and $\widetilde{\mathbf{Y}}_\tau \equiv \mathbf{P}^{-1} \mathbf{Y}_\tau$ gives rise to the same transition matrix $\boldsymbol{\Psi}_\tau$ for fully arbitrary invertible matrices P.

The L-dimensional hyperplane spanned by the rows of $\mathbf{Y}_\tau$ does not depend on the horizon τ . This follows from (165) and (166) and the fact that since $\boldsymbol{\Gamma}$ generates rotations (imaginary part of the eigenvalues) and/or contractions (real part of the eigenvalues), the matrix $e^{-\boldsymbol{\Theta}\tau}$ maps the eigenspaces of $\boldsymbol{\Theta}$ into themselves. In particular, any eigenspace of Θ relative to a null eigenvalue is mapped into itself at any horizon.

Assuming that the non-null eigenvalues of $\boldsymbol{\Theta}$ have positive real part, the rows of $\mathbf{Y}_\tau$, or of any alternative representation, span the contraction hyperplanes and the process $\mathbf{Y}_\tau X_t$ is stationary and would converge exponentially fast to the the unconditional expectation $\mathbf{Y}_\tau \boldsymbol{\mu}$, if it were not for the shocks $\varepsilon_{t,\tau}$ in (164).

5.4. Vector autoregressive model

Vector autoregression (VAR) is an econometric model used to capture the evolution and the interdependencies between multiple time series, generalizing the univariate AR models. All the variables in a VAR are treated symmetrically by including for each variable an equation explaining its evolution based on its own lags and the lags of all the other variables in the model.

5.4.1 Definition

VAR model is a multivariate AR(n) model. In a VAR model the current value of each variable is a linear function of the past values of all variables plus random disturbances. In full generality, a VAR model can be written as follows

$$\boldsymbol{x}_t = \boldsymbol{A}_1 \boldsymbol{x}_{t-1} + \boldsymbol{A}_2 \boldsymbol{x}_{t-2} + \ldots + \boldsymbol{A}_p \boldsymbol{x}_{t-p} + \boldsymbol{D}_{s_t} + \varepsilon_t \quad (168)$$

where xt = (x1, t, …, xn, t) is a multivariate stochastic time series in vector notation, Ai, i = 1,2,...,p, and D are deterministic n×n matrices, εt = ε1, t, …, εn, t is a multivariate white noise with variance-covariance matrix Ω = { σij } and st = s1, t, …, sn, t is a vector of deterministic terms. Using the lag-operator L notation, a VAR model can be written in the following form:

$$\boldsymbol{x}_t = \left(\boldsymbol{A}_1 \boldsymbol{L} + \boldsymbol{A}_2 \boldsymbol{L}^2 + \ldots + \boldsymbol{A}_n \boldsymbol{L}^N\right)\boldsymbol{x}_t + \boldsymbol{D}_{s_t} + \varepsilon_t \quad (169)$$

VAR models can be written in equivalent forms that will be useful in the next section. In particular, a VAR model can be written in terms of the differences Δxt in the following error-correction form:

$$\Delta \boldsymbol{x}_t = \left(\Phi_1 \boldsymbol{L} + \Phi_2 \boldsymbol{L}^2 + \ldots + \Phi_{n-1} \boldsymbol{L}^{N-1}\right)\Delta \boldsymbol{x}_t + \prod \boldsymbol{L}^{N-1} + \boldsymbol{D}_{s_t} + \varepsilon_t \quad (170)$$

where the first n – 1 terms are in first differences and the last term is in levels. The multivariate random walk model of log prices is the simplest VAR model:

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{x}_t + \boldsymbol{m} + \varepsilon_t \\ \Delta \boldsymbol{x}_t &= \boldsymbol{m} + \varepsilon_t \end{aligned} \quad (171)$$

Note that in this model log prices are autoregressive while returns (that is, the first differences) are simply correlated multivariate white noise plus a constant term.

As we know from our discussion on ARMA models, the stationarity and stability properties of a VAR model depend on the roots of the polynomial matrix

$$\boldsymbol{A}_1 z + \boldsymbol{A}_2 z^2 + \ldots + \boldsymbol{A}_n z^N \quad (172)$$

In particular, if all the roots of the above polynomial are strictly outside the unit circle, then the VAR process is stationary. In this case, the VAR process can be inverted and rewritten as an infinite moving average of a white-noise process. If all the roots are outside the unit circle with the exception of some root which is on the unit circle, then the VAR process is integrated. In this case it cannot be inverted as an infinite moving average. If some of the roots are inside the unit circle, then the process is explosive. If the VAR process starts at some initial point characterized by initial values or distributions, then the process cannot be stationary.

However, if all the roots are outside the unit circle, the process is asymptotically stationary. If some root is equal to 1, then the process can be differentiated to obtain an asymptotically stationary process.

5.5. Principal Components Analysis Statistical Arbitrage

Cointegration, along with its geometric interpretation, was introduced above building on the multivariate Ornstein-Uhlenbeck dynamics. However, cointegration is a model-independent concept. Consider a generic multivariate process $\boldsymbol{X}_t$. This process is cointegrated if there exists a linear combination of its entries which is stationary. Let us denote such combination as follows

$$\mathbf{Y}_{\mathbf{t}}^{\mathbf{W}} \equiv \mathbf{X}_{\mathbf{t}}'\mathbf{w} \text{ (173)}$$

where w is normalized to have unit length for future convenience. If w belongs to the cointegration space, i.e. $\mathbf{Y}_{\mathbf{t}}^{\mathbf{W}}$ is cointegrated, its variancestabilizes as $\boldsymbol{t} \to \infty$. Otherwise, its variance diverges to infinity. Therefore, the combination that minimizes the conditional variance among all the possible combinations is the best candidate for cointegration:

$$\widetilde{w} \equiv \arg\min_{\|w\|=1}\left[Var\left\{Y_{\infty}^{w}\middle|x_0\right\}\right] \text{(174)}$$

Based on this intuition, we consider formally the conditional covariance of the process

$$\Sigma_{\infty} \equiv Cov\left\{X_{\infty}\middle|x_0\right\} \text{(175)}$$

although we understand that this might not be defined. Then we consider the formal principal component factorization of the covariance

$$\mathbf{\Sigma}_{\infty} \equiv \mathbf{E}\mathbf{\Lambda}\boldsymbol{E} \text{ (176)}$$

where **E** is the orthogonal matrix whose columns are the eigenvectors

$$E \equiv \left(e^{(1)},.....,e^{(N)}\right) \text{(177)}$$

and **Λ** is the diagonal matrix of the respective eigenvalues, sorted in decreasing order

$$\mathbf{\Lambda} \equiv \boldsymbol{diag}\left(\lambda^{(1)},.....,\lambda^{(N)}\right) \quad (178)$$

Note that some eigenvalues might be infinite. The formal solution to (174) is $\widetilde{\mathbf{w}} \equiv \mathbf{e}^{(\mathbf{N})}$, the eigenvector relative to the smallest eigenvalue $\lambda^{(N)}$. If $\mathbf{e}^{(\mathbf{N})}$ gives rise to cointegration, the process $\boldsymbol{Y}_t^{\boldsymbol{e}^{(N)}}$ is stationary and therefore the eigenvalue $\lambda^{(N)}$ is not infinite, but rather it represents the unconditional variance of $\boldsymbol{Y}_t^{\boldsymbol{e}^{(N)}}$.

If cointegration is found with $\mathbf{e}^{(\mathbf{N})}$ the next natural candidate for another possible cointegrated relationship is $\mathbf{e}^{(\mathbf{N}-1)}$. Again, if $\mathbf{e}^{(\mathbf{N}-1)}$ gives rise to cointegration, the eigenvalue $\lambda_t^{(N-1)}$ converges to the unconditional variance of $\boldsymbol{Y}_t^{\boldsymbol{e}^{(N-1)}}$.

In other words, the PCA decomposition (176) partitions the space into two portions: the directions of infinite variance, namely the eigenvectors relative to the infinite eigenvalues, which are not cointegrated, and the directions of finite variance, namely the eigenvectors relative to the finite eigenvalues, which are cointegrated.

The above approach assumes knowledge of the true covariance (175), which in reality is not known. However, the sample covariance of the process $\boldsymbol{X}_t$ along the cointegrated directions approximates the true asymptotic covariance.

Therefore, the above approach can be implemented in practice by replacing the true, unknown covariance (175) with its sample counterpart.

To summarize, the above rationale yields a practical routine to detect the cointegrated relationships in a vector autoregressive process (1). Without analyzing the eigenvalues of the transition matrix fitted to an autoregressive dynamics, we consider the sample counterpart of the covariance (175); then we extract the eigenvectors (177); finally we explore the stationarity of the combinations $\boldsymbol{Y}_t^{\boldsymbol{e}^{(N)}}$ for n = N, . . . , 1.

To illustrate, we consider a trading strategy with swap contracts. First, we note that the p&l generated by a swap contract is faithfully approximated by the change in the respective swap rate times a constant, known among practitioners as "dv01". Therefore, we analyze linear combinations of swap rates, which map into portfolios p&l's, hoping to detect cointegrated patterns.

In particular, we consider the time series of the 1y, 2y, 5y, 7y, 10y, 15y, and 30y rates. We compute the sample covariance and we perform its PCA decomposition. In Figure 4 we plot the time series corresponding with the first, second, fourth and seventh eigenvectors.

In particular, to test for the stationarity of the potentially cointegrated series it is convenient to fit to each of them a AR(1) process, i.e. the univariate version of (136), to the cointegrated combinations:

$$y_{t+\tau} \equiv \left(1 - e^{-\theta\tau}\right)\mu + e^{-\theta\tau} y_t + \varepsilon_{t,\tau} \quad (179)$$

In the univariate case the transition matrix Θ becomes the scalar θ. Consistently with (158) cointegration corresponds to the condition that the mean-reversion parameter θ be larger than zero.

By specializing (144) and (145) to the one-dimensional case, we can compute the expected long-term gain

$$\alpha \equiv \left| y_t - E\{y_\infty\} \right| = \left| y_t - \mu \right| \quad (180)$$

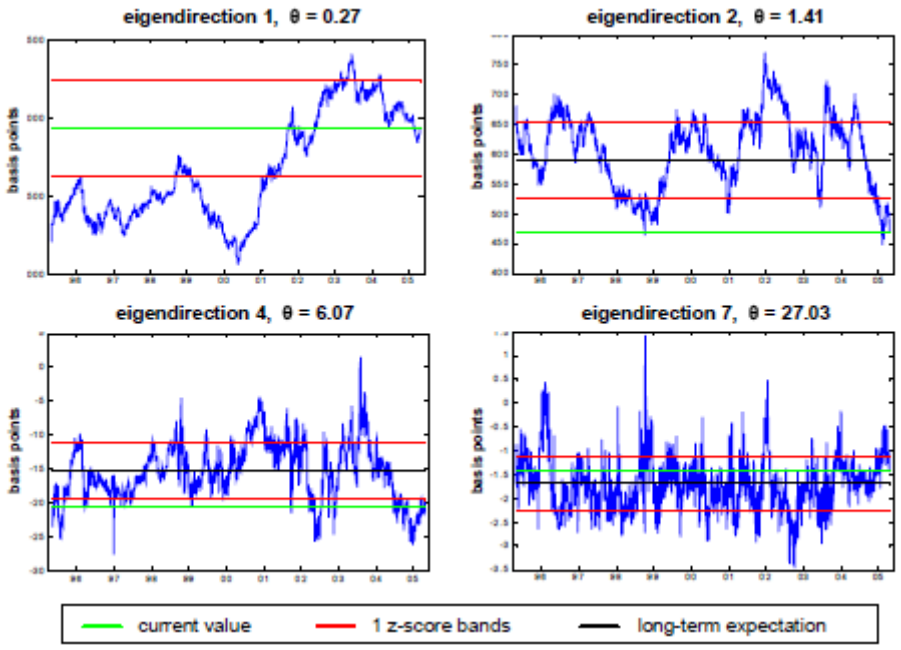
the z-score

$$Z_t \equiv \frac{\left| y_t - E\{y_\infty\} \right|}{Sd\{y_\infty\}} = \frac{\left| y_t - \mu \right|}{\sqrt{\sigma^2 / 2\theta}} \quad (181)$$

and the half-life (156) of the deterministic trend

$$\tilde{\tau}\alpha \frac{1}{\theta} \quad (182)$$

Figure 27: Cointegration search among swap rates



The expected gain (181) is also known as the "alpha" of the trade. The z-score represents the ex-ante Sharpe ratio of the trade and can be used to generate signals: when the z-score is large we

enter the trade; when the z-score has narrowed we cash a profit; and if the z-score has widened further we take a loss. The half-life represents the order of magnitude of the time required before we can hope to cash in any profit: the higher the mean-reversion θ, i.e. the more cointegrated the series, the lesser the wait.

A caveat is due at this point: based on the above recipe, one would be tempted to set up trades that react to signals from the most mean-reverting combinations. However, such combinations face two major problems. First, insample cointegration does not necessarily correspond to out-of-sample results: as a matter of fact, the eigenseries relative to the smallest eigenvalues, i.e. those that allow to make quicker profits, are the least robust out-of-sample.

Second, the "alpha" (181) of a trade has the same order of magnitude as its volatility. In the case of cointegrated eigenseries, the volatility is the square root of the respective eigenvalue (178): this implies that the most mean-reverting series correspond to much a much lesser potential return, which is easily offset by the transaction costs.

In the example in Figure 27 the AR(1) fit (179) confirms that cointegration increases with the order of the eigenseries. In the first eigenseries the meanreversion parameter θ ≈ 0.27 is close to zero: indeed, its pattern is very similar to a random walk. On the other hand, the last eigenseries displays a very high mean-reversion parameter θ ≈ 27.

The current signal on the second eigenseries appears quite strong: one would be tempted to set up a dv01-weighted trade that mimics this series and buy it. However, the expected wait before cashing in on this trade is of the order of $\tilde{\tau}\alpha$ 1/1.41 ≈ 0.7 years.

The current signal on the seventh eigenseries is not strong, but the meanreversion is very high, therefore, soon enough the series should hit the 1-z-score bands: if the series first hits the lower 1-z-score band one should buy the series, or sell it if the series first hits the upper 1-z-score band, hoping to cash in in $\tilde{\tau}\alpha$ 252/27 ≈ 9 days. However, the "alpha" (181) on this trade would be minuscule, of the order of the basis point: such "alpha" would not justify the transaction costs incurred by setting up and unwinding a trade that involves long-short positions in seven contracts.

The current signal on the fourth eigenseries appears strong enough to buy it and the expected wait before cashing in is of the order of $\tilde{\tau}\alpha$ 12/6.07 ≈ 2 months. The "alpha" is of the order of five basis points, too low for seven contracts. However, the dv01-adjusted presence of the 15y contract is almost null and the 5y, 7y, and 10y contracts appear with the same sign and can be replicated with the 7y contract only without affecting the qualitative behaviour of the eigenseries. Consequently the trader might want to consider setting up this trade.

6. Statistical Model Estimation

6.1. Statistical Estimation and Testing

Most statistical models have parameters that must be estimated. Statistical estimation is a set of criteria and methodologies for determining the best estimates of parameters. Testing is

complementary to estimation. Critical parameters are often tested before the estimation process starts in earnest, although some tests of the adequacy of models can be performed after estimation. In general terms, statistics is a way to make inferences from a sample to the entire population from which the sample is taken. In financial econometrics, the sample is typically an empirical time series. Data may be returns, prices, rates, company specific financial data, or macroeconomic data. The objective of estimation techniques is to estimate the parameters of models that describe the empirical data. The key concept in estimation is that of estimators. An estimator is a function of sample data whose value is close to the true value of a parameter in a distribution. For example, the empirical average (i.e., the sum of the samples divided by the number of samples) is an estimator of the mean; that is, it is a function of the empirical data that approximates the true mean. Estimators can be simple algebraic expressions; they can also be the result of complex calculations.

Estimators must satisfy a number of properties. In particular, estimators:

• should get progressively closer to the true value of the parameter to be estimated as the sample size becomes larger,

• should not carry any systematic error, and

• should approach the true values of the parameter to be estimated as rapidly as possible.

Being a function of sample data, an estimator is a random (i.e., stochastic) variable. Therefore, the estimator has a probability distribution referred to as the sampling distribution. In general, the probability distribution of an estimator is difficult to compute accurately from small samples but is simpler for large samples. The sampling distribution is important because certain decisions, such as determining whether a process is integrated, must often be made on the basis of estimators. Because estimators are random variables, decisions are based on comparing empirical estimators with critical values computed from the sampling distribution.

A critical value is a number that allows one to discriminate between accepting or rejecting a hypothesis. For example, suppose we need to know whether a process is integrated. Integration means that the autoregression parameter is 1. Even if a process is integrated, however, every estimate will give results different from 1 because of purely statistical fluctuations. But sampling theory of regressions allows us to determine critical values so that we can reject the hypothesis that the process is integrated if the autoregression coefficient is smaller or larger than the upper/lower critical values.

6.2. Estimation Methods

Because estimation methods involve criteria that cannot be justified by themselves, they are subject to some arbitrariness. The crucial point is that, whereas an estimation process must “fit” a distribution to empirical data, any distribution can, with a few restrictions, be fitted to any empirical data. The choice of distributions thus includes an element of arbitrariness. Suppose we want to determine the probability distribution of the faces of a tossed coin, and in 1,000 experiments, heads comes out 950 times. We probably would conclude that the coin is highly biased and that

heads has a 95 percent probability of coming up. We have no objective way, however, to rule out the possibility that the coin is fair and that we are experiencing an unlikely event. Ultimately, whatever conclusion we draw is arbitrary. Three estimation methods are commonly used in financial econometrics: the least-squares, maximum-likelihood, and Bayesian methods.

6.2.1 The Least-Squares Estimation Method.

The least-squares (LS) estimation method is a best-fit technique adapted to a statistical environment. Suppose a set of points is given and we want to find the straight line that best approximates these points. In financial modelling, a point may represent, for example, a return at a given time. A sensible criterion in this case is to compute the distance of each point from a generic straight line, form the sum of the squares of these distances, and choose the line that minimizes this sum—in short, the ordinary least-squares (OLS) method.

The least-squares method can be adapted to any set of points and to different functional forms (straight lines, polynomial functions, and so on). It can be used, for example, to regress the returns of a stock on a financial ratio.

6.2.2 The Maximum-Likelihood Estimation Method.

The maximum-likelihood (ML) estimation method involves maximizing the likelihood of the sample given an assumption of the underlying distribution (for example, that it is a normal distribution or a uniform distribution). Likelihood is the distribution computed for the sample. For example, suppose a coin is biased so that heads has a 30 percent probability of coming up and tails, a 70 percent probability. What is the likelihood in a random independent sample of 3 heads and 2 tails coming up? It is $0.3 \times 0.3 \times 0.3 \times 0.7 \times 0.7$. The ML method would choose these parameters because they maximize the probability (likelihood) of the sample being observed.

As just noted, the ML method implies that one knows the form of the distribution; otherwise, one cannot compute the likelihood. ML methods can be used, for example, to estimate the long-run relationships (cointegration) between various rates of return.

6.2.3 Bayesian Estimation Methods.

Bayesian estimation methods are based on an interpretation of statistics that is different from that of the OLS or ML methods. Bayesian statistics explicitly assume a subjective element in probability. This subjective element is expressed by the so-called prior distribution, which is the distribution that represents all available knowledge prior to data collection. Bayesian statistics use a specific rule, Bayes' theorem, to update prior probabilities as a function of the arrival of new data to form posterior distributions. Bayes' theorem simply states that the posterior distribution is the prior distribution multiplied by the likelihood. Thus, Bayesian estimates have three ingredients: a prior distribution, a likelihood, and an updating rule.

Note that to write the likelihood, one needs to know the form of the distribution—for example, that it is a Gaussian distribution. The prior distribution will typically be expressed as a distribution of the parameters of the likelihood. In practice, the Bayesian estimate of a model implies that one has an idea of a typical model and that the estimated model is a "perturbation" of the typical model. Bayesian methods are frequently used to allow a portfolio manager to plug his or her own views into a model—that is, to subjectively influence or "perturb" the model.

6.2.4 Robust Estimation

With the widespread use of large predictive models having many parameters, the techniques of robust estimation (that is, estimation that is relatively insensitive to (1) a violation of one or more assumptions and/or (2) estimation errors in the inputs) have gained importance; they are now a key component of estimation technology. For example, robust estimation takes the uncertainty in the estimates into account in portfolio optimization. The motivation for robust estimation is that, because of the size of available samples, estimates are typically noisy when large models are being estimated. In addition, data may contain mistakes. Therefore, extracting the maximum amount of meaningful information from a noisy process is important.

To understand the need for robust estimation, consider the estimation of a correlation matrix. The correlation between two random variables is a number that assumes values between –1 and +1. The value 0 indicates absence of correlation. As discussed in the previous section, the empirical estimator of the correlation parameter is a random variable. Consider two normally distributed variables that are independent (consequently, the true, or population, correlation is zero). For n samples of these variables, it is known in statistics that the correlation parameter will, with 99 percent probability (99 percent of the time), be in the range of plus/minus three times the reciprocal of the square root of the number of samples. If we have 1,000 samples, the correlation parameter will fall (approximately) in the range between –0.1 and +0.1, with 99 percent probability. That is, with 1,000 samples, if the absolute value of their estimated correlation exceeds 0.1, we can conclude at a 99 percent confidence level that the two variables are correlated. Now, consider the correlations of stock returns in a sample of 1,000 trading days (that is, four years) of an aggregate stock index. In the case of the S&P 500 Index, because of symmetry, the correlation matrix has approximately 125,000 entries. Returns on individual stocks are known to be strongly correlated with one another.

In fact, in any four-year period, the empirical average correlation well exceeds the 10 percent level. If we try to evaluate individual correlations (i.e., to discriminate between the correlation levels of various pairs), however, we glean little information. In fact, the distribution of empirical correlation coefficients in the entire correlation matrix is similar to random fluctuations around some mean correlation. If we were to feed this correlation matrix to a mean–variance optimizer, we would obtain meaningless (and dangerous) results—the so-called corner portfolios—because the optimizer would treat low or negative correlations appearing essentially at random, as though they represented actual investment opportunities.

Separating information from noise is a difficult problem. For example, in the case of a correlation matrix, we have to extract a meaningful correlation structure from a correlation matrix whose entries are corrupted by noise. The problem of separating a signal (useful information) from noise by maximizing the “signal-to-noise ratio” is well known in engineering and high energy physics. Communications technology and speech and image recognition, to name a few, are areas where the minimization of noise is a critical component. The following sections outline the techniques of robust estimation used for each class of financial models.

6.3. Estimation of Matrices

Consider the task of estimating a variance–covariance matrix. Suppose we have two random variables. Assume first that they have zero means. The variance of the two variables is defined as the expectation of their square, and the covariance, as the expectation of their product; the correlation is covariance scaled by dividing it by the square root of the individual variances (i.e., the volatilities). Suppose we have a sample formed by extracting n pairs of the two variables from a population. In this case, the empirical variance of each variable is the sum of the squares of the samples divided by the number of samples (i.e., the empirical average of the square of the variables). The empirical variance is a measure of the dispersion of the variables around zero. The empirical covariance between the two variables is a measure of how the two variables move together. It is defined as the sum of the products of the samples of two variables divided by the number of samples. In other words, the empirical covariance is the empirical average of the products of the two variables. Empirical correlation is empirical covariance normalized with (i.e., divided by) the square root of the individual empirical variances.

If the two variables have nonzero means, we simply subtract the mean. For example, we define the variance as the expectation of the variable minus the mean and the covariance as the expectation of the product of the differences of each variable minus the respective means. The empirical variances and covariances are formed by subtracting the empirical mean, defined as the sum of the samples divided by the number of samples.

If k variables are given, we can form a k × k matrix whose entries are the variances and covariances of each pair of the given variables. We can also form a matrix whose entries are the correlations of each pair of variables. The empirical variances, covariances, and correlations as defined here are estimators of the true variances, covariances, and correlations.

The empirical variance–covariance and empirical correlation matrices are noisy, but a number of techniques can be used to make estimates more robust. One such technique, a method based on principal-component analysis (PCA).

6.4. Estimation of Regression Models

6.4.1 Linear regression is the workhorse of equity modelling.

Estimation of regression models is typically performed by using OLS methods. OLS produces estimators that are algebraic expressions of the data. In the two dimensional xy plane, the OLS method can easily be understood: For a set of xy pairs, one can calculate the straight line in the xy plane that minimizes the sum of the squares of the distance from the line to each pair. To illustrate the OLS method, we randomly generated 500 points and, using the OLS method, fitted the best straight line.

It has been proven that the estimators of the regression parameters determined in this way are optimal linear estimators. Under the assumption that the residuals are normally distributed, the OLS estimators of regression parameters coincide with the ML estimators. The ML estimators are obtained by first computing the residuals with respect to a generic regression and then evaluating the likelihood. The likelihood is obtained by computing the value of a normal distribution on the residuals. The likelihood is then minimized.

Now, suppose we want to estimate the linear regression of one dependent time series on one or more independent time series. At each time step, we observe a sample of the linear regression to be estimated. However, there may be one complication: Residuals might be autocorrelated. The autocorrelation of residuals does not invalidate the standard OLS estimators, but it makes them less efficient and thus not optimal for small samples. Corrections that take into account the autocorrelation have been suggested and can be easily applied—provided one knows how to determine the autocorrelations of residuals. The asymptotic sampling distribution of regression parameters (i.e., the distribution of regression parameters estimated on large samples) can be easily determined. In large samples, regression parameters are normally distributed, with mean and variance that are simple algebraic functions of data.

The estimates of a regression can be made robust. Robustness can be achieved by replacing the standard OLS estimators with estimators that are less sensitive to outliers (that is, to sample values much larger than the bulk of sample data). Although linear regressions are simple statistical constructs, the analysis and eventual improvement of their performance is delicate. The achievement of robustness and performance of a linear regression hinges on our ability to (1) identify a set of optimal regressors and (2) partition the samples to improve performance.

Consider first the identification of a set of optimal regressors. Simply increasing the number of regressors is not a good strategy because by adding regressors, we increase the number of parameters that must be estimated. Adding regressors also augments the noise of all estimated parameters. Therefore, each additional regressor must be understood and its contribution must be evaluated. We can determine the importance of a regressor by calculating the ratio of the variance explained by that regressor to total variance. A regressor that explains only a small fraction of the variance has little explanatory power and can be omitted. To gauge the total effect of adding or removing a regressor, one can use a penalty function that grows with the number of regressors. In this function, the eventual contribution of an additional regressor is penalized to take into account the overall negative effect of estimating more parameters. This type of analysis is performed by most statistical software packages.

Clustering the sample data achieves different objectives. For example, the clustering of sample data corresponds to the need to make estimates more robust by averaging regression parameters estimated on different clusters. This approach is the basic idea behind the techniques of shrinkage and random coefficient models. Alternatively, to improve performance, regressions might be made contextual. That is, for example, a given predictor of returns might be particularly effective in a specific context, such as a particular market segment or in particular market conditions (Sorensen, Hua, and Qian 2005)[81].

Clearly, despite the intrinsic simplicity of the model, designing and estimating linear regressions is a delicate statistical (and, ultimately, economic) problem. It entails one of the critical issues in testing and modelling—the ever-present tradeoffs among model complexity, model risk, and model performance. These tradeoffs are a major theme throughout this monograph. Increasing the dimensionality of the model (for example, by adding regressors) makes it more powerful but also makes the model noisier and thus "riskier."

6.5. Estimation of Vector Autoregressive Models

In principle, VAR models are kinds of regression models, so estimating VAR models is similar to regression estimation. Some VAR models are subject to restrictions, however, that require the use of special techniques. The simplest case is estimating unrestricted stable VAR models. An unrestricted model is a model in which the parameters are allowed to take any value that results from the estimation process. A model is restricted if its parameters can assume values only in specified ranges.

A VAR model is considered to be stable if its solutions are stationary—that is, if the mean, variance, and covariances of its solutions do not change over time. Stability conditions of a VAR model are expressed through conditions that must be satisfied by its parameters—that is, coefficients of every stable model satisfy certain conditions. In particular, stability conditions require solutions to be exponentials with exponent less than 1 in modulus. Stable VAR models can be estimated by using standard LS and ML methods. In fact, a VAR model is a linear regression of its variables over their own lagged values plus error terms. Clearly, all such variables can be grouped together and residuals can be expressed in terms of data and model parameters. If the residuals are uncorrelated, we can then use multivariate LS methods to minimize the sum of squared residuals as a function of the model parameters. As a result, if we arrange the sample data in appropriate vectors and matrices, we can express the estimators of the model parameters as algebraic functions that involve solely matrix operations, such as inversion and multiplication. These estimators are implemented in commercial software programs. If we make specific assumptions about the distribution of residuals, we can also use ML methods. In particular, if the model residuals are normally distributed, the ML model estimators coincide with the LS estimators. If the VAR model is not stable, unrestricted LS methods might still be usable.

In dealing with an unstable VAR model, however, one is generally interested in testing and estimating cointegrating relationships. Recall that a cointegrating relationship is a stationary linear combination of the process variables. Taking into account cointegrating relationships in estimating a VAR model cannot be done with standard ML regression methods. The cointegrating relationships impose complicated restrictions on the likelihood function that must be maximized. State-of-the-art ML-based estimation methods for cointegrated systems use a complicated procedure to eliminate constraints from the likelihood function (see Johansen 1991; Banerjee and Hendry 1992)[82]. Other methodologies have been proposed, including one based on PCA that is applicable to large data sets.

Bayesian VARs (BVARs) are VAR models estimated by Bayesian methods. When applied to VAR models, Bayesian estimates start from a priori distribution of the model parameters. In practice, this distribution embodies a formulation of an idealized model. The a priori distribution is then multiplied by the likelihood function, computed as usual by using ML methods. The resulting so-called a posteriori likelihood is maximized.

Perhaps the best known BVAR is the model proposed by Litterman (1986)[83]. The essence of the Litterman model is that any financial VAR model is a perturbation of a multivariate random walk. The Litterman model determines the a priori distribution so that the average of this distribution is simply a random walk. The likelihood function updates the a priori distribution, and the result is maximized.

Because it requires that the solutions of estimated VAR models do not deviate much from a random walk, the Litterman model is robust. Extending Bayesian estimates to cointegrated VAR models is not straightforward. The problem is that one has to impose a cointegration structure as an a priori distribution. A number of solutions to this problem have been proposed, but none of them has obtained the general acceptance enjoyed by BVARs.

6.6. Estimation of Linear Hidden-Variable Models

Hidden-variable models include linear state-space models in various formulations and nonlinear models—in particular, Markov switching–VAR (MS–VAR) models.

Linear State-Space Models. Because they include variables that are not observed but must be estimated, state-space models cannot be estimated by using standard regression techniques. A crucial component in estimating state-space models is a tool to filter data known as the “Kalman filter.” It is a recursive computational algorithm that, assuming that the model is known, estimates the states from the data.

It was conceived in the 1960s to solve the problem of estimating true data—in particular, the position of an aircraft or a missile—from noisy measurements. Estimating state-space models is done through two general methodologies: ML-based methods and subspace methods. ML-based methods compute the likelihood function of the state-space model that includes hidden variables. Hidden variables are then estimated from the data by using the Kalman filter and the assumption

of an unknown generic model. The result is a likelihood function that is expressed as a function of only the unknown model parameters. Maximizing this likelihood yields the estimators. Subspace methods are technical. They estimate the states by using the Kalman filter, divide the sample data into two sections (conventionally called the “past” and the “future”), and then perform a regression of the future on the past.Dynamic factor models are a version of state-space models. Several other estimation methods have been proposed, including estimating the equivalent statespace model and the use of PCA-based methods.

Robust Estimation Methods for Linear Models.

These estimation methods for VAR models are not intrinsically robust and do not scale well to large systems that are common in finance. Litterman’s BVAR is a robust model but can be applied only to small systems (e.g., systems made up of indices). Making VAR estimates robust in the case of a large system requires reducing the dimensionality of the model, which calls for factor models and, in particular, dynamic factor models of prices or returns.

6.7. Estimation of Nonlinear Hidden-Variable Models

In discussing the estimation of nonlinear hidden-variable models, we focus on the MS–VAR models because the generalized autoregressive conditional heteroscedasticity (GARCH) family of nonlinear hidden-variable models is rarely used in equity modelling. MS–VAR models, however, are being adopted to model regime changes.
Because nonlinear MS–VAR models have hidden variables, their estimation presents the same difficulties as does the estimation of linear state-space models. And for nonlinear MS–VAR models, no equivalent of the Kalman filter exists. Estimation techniques for MS–VAR models typically use the expectation-maximization algorithm (often referred to as the “EM” algorithm) used by Hamilton (1996) in his regime-shift model.

7. Nonlinear Dynamical Systems

7.1. Motivation

It is well known that many nonlinear dynamical systems, including seemingly simple cases, can exhibit chaotic behavior. In short, the presence of chaos implies that very small changes in the initial conditions or parameters of a system can lead to drastic changes in its behavior. In the chaotic regime, the system solutions stay within the phase space region named strange attractor. These solutions never repeat themselves; they are not periodic and they never intersect. Thus, in the chaotic regime, the system becomes unpredictable. The chaos theory is an exciting and

complex topic. In this thesis, I only outline the main concepts that may be relevant to quantitative finance.

The first reason to turn to chaotic dynamics is a better understanding of possible causes of price randomness. Obviously, new information coming to the market moves prices. Whether it is a company's performance report, a financial analyst's comments, or a macroeconomic event, the company's stock and option prices may change, thus reflecting the news. Since news usually comes unexpectedly, prices change in unpredictable ways. But is new information the only source reason for price randomness? One may doubt this while observing the price fluctuations at times when no relevant news is released. A tempting proposition is that the price dynamics can be attributed in part to the complexity of financial markets. The possibility that the deterministic processes modulate the price variations has a very important practical implication: even though these processes can have the chaotic regimes, their deterministic nature means that prices may be partly forecastable. Therefore, research of chaos in finance and economics is accompanied with discussion of limited predictability of the processes under investigation.[84]

There have been several attempts to find possible strange attractors in the financial and economic time series.[85] Discerning the deterministic chaotic dynamics from a ''pure'' stochastic process is always a non-trivial task. This problem is even more complicated for financial markets whose parameters may have non-stationary components.[86] So far, there has been little (if any) evidence found of low-dimensional chaos in financial and economic time series. Still, the search of chaotic regimes remains an interesting aspect of empirical research.

There is also another reason for paying attention to the chaotic dynamics. One may introduce chaos inadvertently while modelling financial or economic processes with some nonlinear system. This problem is particularly relevant in agent-based modelling of financial markets where variables generally are not observable. Nonlinear continuous systems exhibit possible chaos if their dimension exceeds two. However, nonlinear discrete systems (maps) can become chaotic even in the one-dimensional case. Note that the autoregressive models being widely used in analysis of financial time series are maps in terms of the dynamical systems theory. Thus, a simple nonlinear expansion of a univariate autoregressive map may lead to chaos, while the continuous analog of this model is perfectly predictable. Hence, understanding of nonlinear dynamical effects is important not only for examining empirical time series but also for analyzing possible artifacts of the theoretical modelling.

This chapter continues with a widely popular one-dimensional discrete model, the logistic map, which illustrates the major concepts in the chaos theory. Furthermore, the framework for the continuous systems is introduced after that. Then the three dimensional Lorenz model, being the classical example of the low-dimensional continuous chaotic system, is described (Section 7.4).

Finally, the main pathways to chaos and the chaos measures are outlined in Section 7.5 and Section 7.6, respectively.

7.2. Discrete systems: the logistic map

The logistic map is a simple discrete model that was originally used to describe the dynamics of biological populations[87]. Let us consider a variable number of individuals in a population, N. Its value at the k-th time interval is described with the following equation:

$$N_k = rN_{k-1} - BN_{k-1}^2 \quad (183)$$

Parameter r characterizes the population growth that is determined by such factors as food supply, climate, etc. Obviously, the population grows only if r > 1. If there are no restrictive factors (i.e., when B = 0), the growth is exponential, which never happens in nature for long. Finite food supply, predators, and other causes of mortality restrict the population growth, which is reflected in factor B. The maximum value of Nk equals Nmax= r/B. It is convenient to introduce the dimensionless variable Xk=Nk/Nmax. Then 0≤Xk≤1, and equation (183) has the form:

$$X_k = rX_{k-1}(1 - X_{k-1}) \quad (184)$$

A generic discrete equation in the form

$$X_k = f(X_{k-1}) \quad (185)$$

is called an (iterated) map, and the function f(Xk-1) is called the iteration function. The map (184) is named the logistic map. The sequence of values Xk that are generated by the iteration procedure is called a trajectory. Trajectories depend not only on the iteration function but also on the initial value X0. Some initial points turn out to be the map solution at all iterations. The value X* that satisfies the equation

$$X^* = f(X^*) \quad (186)$$

is named the fixed point of the map. There are two fixed points for the logistic map (184):

$$X_1^* = 0, \text{ and } X_2^* = (r-1)/r \quad (187)$$

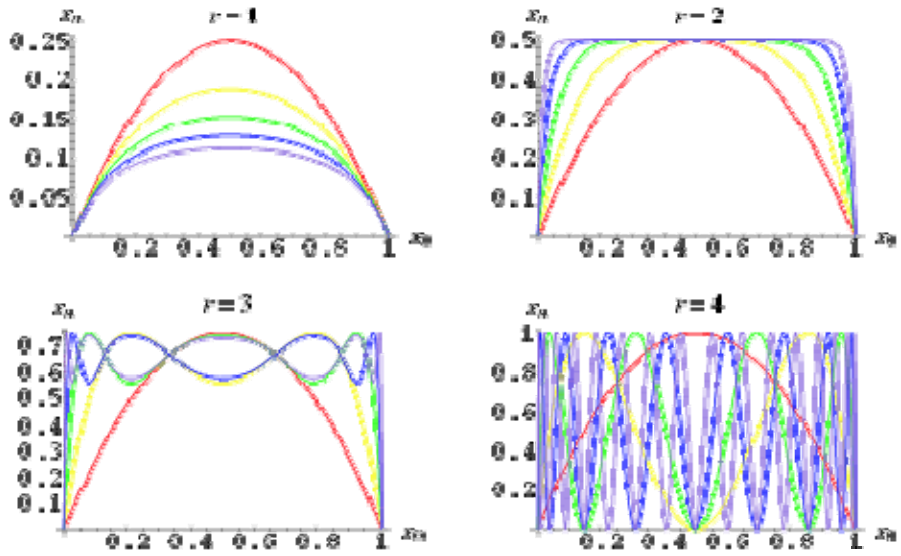
If r ≤1, the logistic map trajectory approaches the fixed point X* 1 from any initial value 0 ≤X0 ≤1. The set of points that the trajectories tend to approach is called the attractor. Generally, nonlinear dynamical systems can have several attractors. The set of initial values from which the trajectories approach a particular attractor are called the basin of attraction.

For the logistic map with r < 1, the attractor is X*1= 0, and its basin is the interval 0 ≤X0 ≤1. If 1 < r < 3, the logistic map trajectories have the attractor X* 2 = (r - 1)/r and its basin is also 0 ≤X0 ≤1. In the mean time, the point X* 1=0 is the repellent fixed point, which implies that any trajectory that starts near X* 1 tends to move away from it.

A new type of solutions to the logistic map appears at r > 3. Consider the case with r =3:1: the trajectory does not have a single attractor but rather oscillates between two values, X ≈ 0:558 and X ≈ 0:764. In the biological context, this implies that the growing population overexerts its survival capacity at X ≈ 0:764. Then the population shrinks ‘‘too much’’ (i.e., to X ≈ 0:558), which yields capacity for further growth, and so on. This regime is called period- 2. The parameter value at which solution changes qualitatively is named the bifurcation point. Hence, it is said that the period-doubling bifurcation occurs at r = 3. With a further increase of r, the oscillation amplitude

grows until r approaches the value of about 3.45. At higher values of r, another period-doubling bifurcation occurs (period-4). This implies that the population oscillates among four states with different capacities for further growth. Period doubling continues with rising r until its value approaches 3.57. Typical trajectories for different r's are given in Figure 28. With further growth of r, the number of periods becomes infinite, and the system becomes chaotic. Note that the solution to the logistic map at r > 4 is unbounded.

Figure 28: Logistic map



Specifics of the solutions for the logistic map are often illustrated with the bifurcation diagram in which all possible values of X are plotted against r. Interestingly, it seems that there is some order in this diagram even in the chaotic region at r > 3:6. This order points to the fractal nature of the chaotic attractor, which will be discussed later on.

Another manifestation of universality that may be present in chaotic processes is the Feigenbaum's observation of the limiting rate at which the period-doubling bifurcations occur. Namely, if rn is the value of r at which the period-2n occurs, then the ratio:

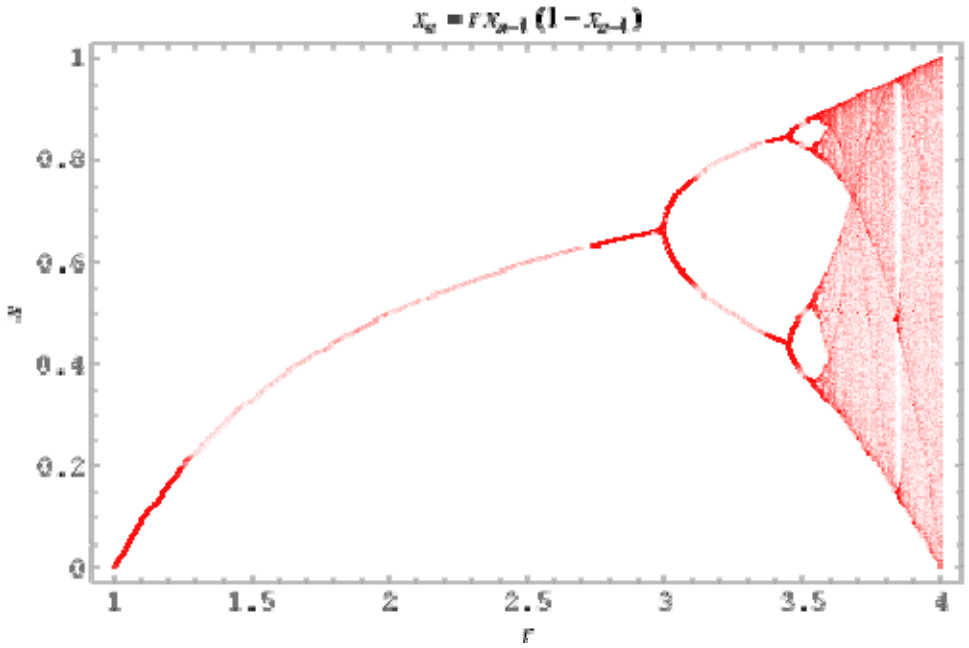
$$\delta_n = (r_n - r_{n-1})/(r_{n+1} - r_n) \quad (188)$$

has the limit

$$\lim_{n\to\infty} \delta_n = 4.669... \quad (189)$$

It turns out that the limit (189) is valid for the entire family of maps with the parabolic iteration functions A very important feature of the chaotic regime is extreme sensitivity of trajectories to the initial conditions. Thus, the logistic map provides an illuminating example of complexity and universality generated by interplay of nonlinearity and discreteness.

Figure 29: Bifurcation diagram



7.3. Continuous systems

While the discrete time series are the convenient framework for financial data analysis, financial processes are often described using continuous presentation[88]. Hence, we need understanding of the chaos specifics in continuous systems[89].. First, let us introduce several important notions with a simple model of a damped oscillator. Its equation of motion in terms of the angle of deviation from equilibrium,θ , is:

$$\frac{d^2\theta}{dt^2} + \gamma\frac{d\theta}{dt} + \omega^2\theta = 0 \quad (190)$$

In (190), γ is the damping coefficient and v is the angular frequency. Dynamical systems are often described with flows, sets of coupled differential equations of the first order. These equations in the vector notations have the following form:

$$\frac{dX}{dt} = F(X(t)), X = (X_1, X_2, ...X_N)' \quad (191)$$

We shall consider so-called autonomous systems for which the function F in the right-hand side of (191) does not depend explicitly on time. A non-autonomous system can be transformed into an autonomous one by treating time in the function F(X, t) as an additional variable, XN+1 = t, and adding another equation to the flow

$$\frac{dX_{N+1}}{dt} = 1 \quad (192)$$

As a result, the dimension of the phase space increases by one. The notion of the fixed point in continuous systems differs from that of discrete systems (186). Namely, the fixed points for the flow (191) are the points X* at which all derivatives in its left-hand side equal zero. For the obvious reason, these points are also named the equilibrium (or stationary) points: If the system reaches one of these points, it stays there forever.

Equations with derivatives of order greater than one can be also transformed into flows by introducing additional variables. For example, equation (190) can be transformed into the system

$$\frac{d\theta}{dt} = \varphi, \ \frac{d\varphi}{dt} = -\gamma\varphi + \omega^2\theta \quad (193)$$

Hence, the damped oscillator may be described in the two-dimensional phase space (φ, θ)

The energy of the damped oscillator, E,

$$E = 0.5\left(\varphi^2 + \omega^2\theta^2\right) \quad (194)$$

evolves with time according to the equation

$$\frac{dE}{dt} = -\gamma\varphi^2 \quad (195)$$

It follows from (195) that the dumped oscillator dissipates energy (i.e., is a dissipative system).

Chaos is usually associated with dissipative systems. Systems without energy dissipation are named conservative or Hamiltonian systems. Some conservative systems may have the chaotic regimes, too (so-called non-integrable systems), but this case will not be discussed here. One can easily identify the sources of dissipation in real physical processes, such as friction, heat radiation, and so on. In general, flow (191) is dissipative if the condition

$$div(F) \equiv \sum_{i=1}^{N} \frac{\partial F}{\partial X_i} < 0 \quad (196)$$

is valid on average within the phase space. Besides the point attractor, systems with two or more dimensions may have an attractor named the limit cycle. An example of such an attractor is the solution of the Van der Pol equation. This equation describes an oscillator with a variable damping coefficient

$$\frac{d^2\theta}{dt^2} + \gamma\left[\left(\theta/\theta_0\right)^2 - 1\right]\frac{d\theta}{dt} + \omega^2\theta = 0 \quad (197)$$

Since the solution to the Van der Pol equation changes qualitatively from the point attractor to the limit cycle at e=0, this point is a bifurcation. Those bifurcations that lead to the limit cycle are named the Hopf bifurcations.

In three-dimensional dissipative systems, two new types of attractors appear. First, there are quasi-periodic attractors. These trajectories are associated with two different frequencies and are located on the surface of a torus. Another type of attractor that appears in three-dimensional systems is the strange attractor. It will be introduced using the famous Lorenz model in the next section.

7.4. Lorenz model

The Lorenz model describes the convective dynamics of a fluid layer with three dimensionless variables:

$$\frac{dX}{dt} = p(Y - X)$$
$$\frac{dY}{dt} = -XZ + rX - Y \quad (198)$$
$$\frac{dZ}{dt} = XY - bZ$$

In (198), the variable X characterizes the fluid velocity distribution, and the variables Y and Z describe the fluid temperature distribution. The dimensionless parameters p, r, and b characterize the thermohydrodynamic and geometric properties of the fluid layer. The Lorenz model, being independent of the space coordinates, is a result of significant simplifications of the physical process under consideration. Yet, this model exhibits very complex behavior.

Figure 30: Trajectories Lorenz model

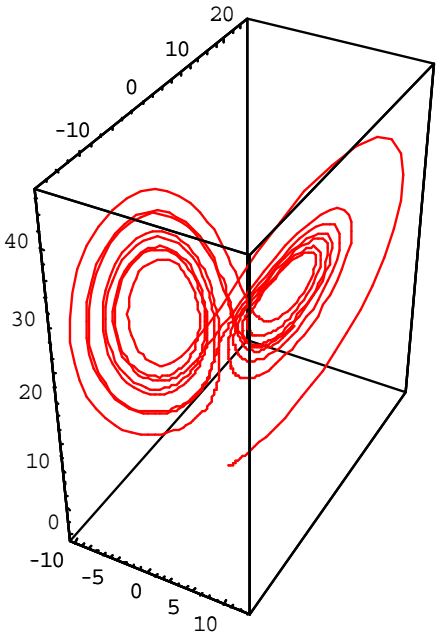


Figure 31: 2D and 3D Packard-Takens Autocorrelation Plots of Sinusoidal Functions

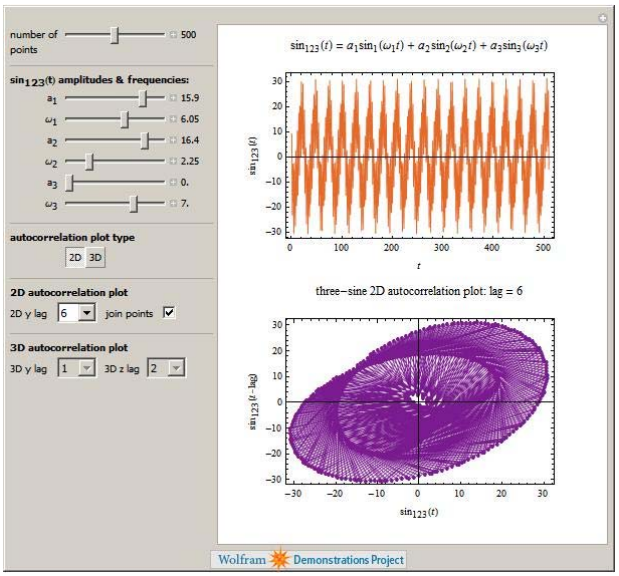


Figure 31 shows the autocorrelations of one- to three-term sinusoidal functions plotted in two and three dimensions. You can select the function's underlying amplitudes and frequencies as well as the lags and the number of generated points. You can also choose whether or not to connect the points in the two-dimensional display only.

7.5. Pathways to chaos

A number of general pathways to chaos in nonlinear dissipative systems have been described in the literature. [90]. All transitions to chaos can be divided into two major groups: local bifurcations and global bifurcations. Local bifurcations occur in some parameter range, but the trajectories become chaotic when the system control parameter reaches the critical value. Three types of local bifurcations are discerned: period-doubling, quasi-periodicity, and intermittency. Period-doubling starts with a limit cycle at some value of the system control parameter. With further change of this parameter, the trajectory period doubles and doubles until it becomes infinite. This process was proposed by Landau as the main turbulence mechanism. Namely, laminar flow develops oscillations at some sufficiently high velocity. As velocity increases, another (imcommensurate) frequency appears in the flow, and so on. Finally, the frequency spectrum has the form of a practically continuous band. An alternative mechanism of turbulence (quasi-periodicity) was proposed by Ruelle and Takens[91]. They have shown that the quasi-periodic trajectories confined on the torus surface can become chaotic due to high sensitivity to the input parameters. Intermittency is a broad category itself. Its pathway to chaos consists of a sequence of periodic and chaotic regions. With changing the control parameter, chaotic regions become larger and larger and eventually fill the entire space. In the global bifurcations, the trajectories approach simple attractors within some control parameter range. With further change of the control parameter, these trajectories become increasingly complicated and in the end, exhibit

chaotic motion. Global bifurcations are partitioned into crises and chaotic transients. Crises include sudden changes in the size of chaotic attractors, sudden appearances of the chaotic attractors, and sudden destructions of chaotic attractors and their basins. In chaotic transients, typical trajectories initially behave in an apparently chaotic manner for some time, but then move to some other region of the phase space. This movement may asymptotically approach a non-chaotic attractor. Unfortunately, there is no simple rule for determining the conditions at which chaos appears in a given flow. Moreover, the same system may become chaotic in different ways depending on its parameters. Hence, attentive analysis is needed for every particular system.

7.6. Measuring chaos

As it was noticed in in Section 7.1, it is important to understand whether randomness of an empirical time series is caused by noise or by the chaotic nature of the underlying deterministic process. To address this problem, let us introduce the Lyapunov exponent. The major property of a chaotic attractor is exponential divergence of its nearby trajectories. Namely, if two nearby trajectories are separated by distance d0 at t =0, the separation evolves as

$$\boldsymbol{d}(\boldsymbol{t}) = \boldsymbol{d}_0 \exp(\lambda \boldsymbol{t}) \quad (199)$$

The parameter λ in (199) is called the Lyapunov exponent. For the rigorous definition, consider two points in the phase space, Xo and Xo + ∆xo, that generate two trajectories with some flow (157). If the function ∆x(Xo, t) defines evolution of the distance between these points, then

$$\lambda = \lim \frac{1}{\boldsymbol{t}} \ln \frac{|\Delta \boldsymbol{x}(\boldsymbol{X}_0 \boldsymbol{t})|}{\Delta \boldsymbol{x}_0}, \ \boldsymbol{t} \to \infty, \Delta \boldsymbol{x} \to 0 \quad (200)$$

When λ < 0, the system is asymptotically stable. If λ =0, the system is conservative. Finally, the case with λ > 0 indicates chaos since the system trajectories diverge exponentially.

The practical receipt for calculating the Lyapunov exponent is as follows. Consider n observations of a time series $\boldsymbol{x}(\boldsymbol{t}) : \boldsymbol{x}(\boldsymbol{t}_k) = \boldsymbol{x}_k, \boldsymbol{k} = 1,...,\boldsymbol{n}$ First, select a point xi and another point xj close to xi. Then calculate the distances

$$\boldsymbol{d}_0 = \left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|, \boldsymbol{d}_1 = \left|\boldsymbol{x}_{i+1} - \boldsymbol{x}_{j+1}\right|, ..., \boldsymbol{d}_n = \left|\boldsymbol{x}_{i+n} - \boldsymbol{x}_{j+n}\right| \quad (201)$$

If the distance between Xi+n and Xj+n evolves with n accordingly with (199), then

$$\lambda(\boldsymbol{x}_i) = \frac{1}{\boldsymbol{n}} \ln \frac{\boldsymbol{d}_n}{\boldsymbol{d}_0} \quad (202)$$

The value of the Lyapunov exponent λ (xi) in (202) is expected to be sensitive to the choice of the initial point xi. Therefore, the average value over a large number of trials N of λ (xi) is used in practice

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} \lambda(x_i) \quad (203)$$

Due to the finite size of empirical data samples, there are limitations on the values of n and N, which affects the accuracy of calculating the Lyapunov exponent. More details about this problem, as well as other chaos quantifiers, such as the Kolmogorov-Sinai entropy.

The generic characteristic of the strange attractor is its fractal dimension. In fact, the non-integer (i.e., fractal) dimension of an attractor can be used as the definition of a strange attractor..A computationally simpler alternative, so-called correlation dimension, is often used in nonlinear dynamics. Consider a sample with N trajectory points within an attractor. To define the correlation dimension, first the relative number of points located within the distance R from the point i must be calculated

$$p_i(R) = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \theta\left(R - \left|x_j - x_i\right|\right) \quad (204)$$

In (204) , the Heaviside step function θ equals

$$\theta = \begin{cases} 0, & \boldsymbol{x} < 0 \\ 0, & \boldsymbol{x} \geq 0 \end{cases} \quad (205)$$

Then the correlation sum that characterizes the probability of finding two trajectory points within the distance R is computed

$$\boldsymbol{C}(\boldsymbol{R}) = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{p}_{\boldsymbol{i}}(\boldsymbol{R}) \quad (206)$$

It is assumed that $\boldsymbol{C}(\boldsymbol{R}) \approx \boldsymbol{R}^{\boldsymbol{Dc}}$. Hence, the correlation dimension Dc equals

$$\boldsymbol{D}_{\boldsymbol{c}} = \lim_{\boldsymbol{R}\to 0} \left[\ln \boldsymbol{C}(\boldsymbol{R}) / \ln \boldsymbol{R}\right] \quad (207)$$

There is an obvious problem of finding the limit (207) for data samples on a finite grid. Yet, plotting ln[C(R)] versus ln(R) (which is expected to yield a linear graph) provides an estimate of the correlation dimension. An interesting question is whether a strange attractor is always chaotic, in other words, if it always has a positive Lyapunov exponent. It turns out there are rare situations when an attractor may be strange but not chaotic. One such example is the logistic map at the period-doubling points: Its Lyapunov exponent equals zero while the fractal dimension is about 0.5. Current opinion, however, holds that the strange deterministic attractors may appear in discrete maps rather than in continuous systems.

8. Technical analysis

In finance, technical analysis is a security analysis discipline for forecasting the future direction of prices studying past market data, primarily price, return and volume and other transformations.

8.1. History

The principles of technical analysis derive from the observation of financial markets over hundreds of years.The oldest known hints of technical analysis appear in Joseph de la Vega's

accounts of the Dutch markets in the 17th century. In Asia, the oldest example of technical analysis is thought to be a method developed by Homma Munehisa during early 18th century which evolved into the use of candlestick techniques, and is today a main charting tool Dow Theory is based on the collected writings of Dow Jones co-founder and editor Charles Dow, and inspired the use and development of modern technical analysis from the end of the 19th century. Other pioneers of analysis techniques include Ralph Nelson Elliott[92] and William Delbert Gann who developed their respective techniques in the early 20th century.
Many more technical tools and theories have been developed and enhanced in recent decades, with an increasing emphasis on computer-assisted techniques.

8.2. General description

Technical analysts seek to identify price patterns and trends in financial markets and attempt to exploit those patterns.[93] While technicians use various methods and tools, the study of price charts is primary. Technicians especially search for archetypal patterns, such as the well-known head and shoulders or double top reversal patterns, study indicators such as moving averages, and look for forms such as lines of support, resistance, channels, and more obscure formations such as flags, pennants or balance days.
Technical analysts also extensively use indicators, which are typically mathematical transformations of price or volume. These indicators are used to help determine whether an asset is trending, and if it is, its price direction.
Technicians also look for relationships between price, volume and, in the case of futures, open interest. Examples include the relative strength index, and MACD. Other avenues of study include correlations between changes in options (implied volatility) and put/call ratios with price. Other technicians include sentiment indicators, such as Put/Call ratios and Implied Volatility in their analysis.
Technicians seek to forecast price movements such that large gains from successful trades exceed more numerous but smaller losing trades, producing positive returns in the long run through proper risk control and money management.
There are several schools of technical analysis. Adherents of different schools (for example, candlestick charting, Dow Theory, and Elliott wave theory) may ignore the other approaches, yet many traders combine elements from more than one school. Some technical analysts use subjective judgment to decide which pattern a particular instrument reflects at a given time, and what the interpretation of that pattern should be. Some technical analysts also employ a strictly mechanical or systematic approach to pattern identification and interpretation.
Technical analysis is frequently contrasted with fundamental analysis, the study of economic factors that influence prices in financial markets. Technical analysis holds that prices already reflect all such influences before investors are aware of them, hence the study of price action alone. Some traders use technical or fundamental analysis exclusively, while others use both

types to make trading decisions. Users of technical analysis are most often called technicians or market technicians. Some prefer the term technical market analyst or simply market analyst. An older term, chartist, is sometimes used, but as the discipline has expanded and modernized the use of the term chartist has become less popular.

8.3. Characteristics

Technical analysis employs models and trading rules based on price and volume transformations, such as the relative strength index, moving averages, regressions, inter-market and intra-market price correlations, cycles or, classically, through recognition of chart patterns.

Technical analysis stands in contrast to the fundamental analysis approach to security and stock analysis. Technical analysis "ignores" the actual nature of the company, market, currency or commodity and is based solely on "the charts," that is to say price and volume information, whereas fundamental analysis does look at the actual facts of the company, market, currency or commodity. For example, any large brokerage, trading group, or financial institution will typically have both a technical analysis and fundamental analysis team.

Technical analysis is widely used among traders and financial professionals, and is very often used by active day traders, market makers, and pit traders. In the 1960s and 1970s it was widely dismissed by academics. In a recent review, Irwin and Park reported that 56 of 95 modern studies found it produces positive results, but noted that many of the positive results were rendered dubious by issues such as data snooping so that the evidence in support of technical analysis was inconclusive; it is still considered by many academics to be pseudoscience. Academics such as Eugene Fama say the evidence for technical analysis is sparse and is inconsistent with the weak form of the efficient market hypothesis. Users hold that even if technical analysis cannot predict the future, it helps to identify trading opportunities.

In the foreign exchange markets, its use may be more widespread than fundamental analysis. While some isolated studies have indicated that technical trading rules might lead to consistent returns in the period prior to 1987, most academic work has focused on the nature of the anomalous position of the foreign exchange market.It is speculated that this anomaly is due to central bank intervention. Recent research suggests that combining various trading signals into a Combined Signal Approach may be able to increase profitability and reduce dependence on any single rule.

Principles

Technical analysts say that a market's price reflects all relevant information, so their analysis looks at the history of a security's trading pattern rather than external drivers such as economic, fundamental and news events. Price action also tends to repeat itself because investors collectively tend toward patterned behavior – hence technicians' focus on identifiable trends and conditions.

Market action discounts everything

Based on the premise that all relevant information is already reflected by prices, pure technical analysts believe it is redundant to do fundamental analysis – they say news and news events do not significantly influence price, and cite supporting research such as the study by Cutler, Poterba, and Summers titled "What Moves Stock Prices?" On most of the sizable return days [large market moves]...the information that the press cites as the cause of the market move is not particularly important. Press reports on adjacent days also fail to reveal any convincing accounts of why future profits or discount rates might have changed. Our inability to identify the fundamental shocks that accounted for these significant market moves is difficult to reconcile with the view that such shocks account for most of the variation in stock returns.

8.4. Prices move in trends ?

Technical analysts believe that prices trend directionally, i.e., up, down, or sideways (flat) or some combination. The basic definition of a price trend was originally put forward by Dow Theory
History tends to repeat itself
Technical analysts believe that investors collectively repeat the behavior of the investors that preceded them. "Everyone wants in on the next Microsoft," "If this stock ever gets to $50 again, I will buy it," "This company's technology will revolutionize its industry, therefore this stock will skyrocket" – these are all examples of investor sentiment repeating itself. To a technician, the emotions in the market may be irrational, but they exist. Because investor behavior repeats itself so often, technicians believe that recognizable (and predictable) price patterns will develop on a chart.
Technical analysis is not limited to charting, but it always considers price trends. For example, many technicians monitor surveys of investor sentiment. These surveys gauge the attitude of market participants, specifically whether they are bearish or bullish. Technicians use these surveys to help determine whether a trend will continue or if a reversal could develop; they are most likely to anticipate a change when the surveys report extreme investor sentiment. Surveys that show overwhelming bullishness, for example, are evidence that an uptrend may reverse – the premise being that if most investors are bullish they have already bought the market (anticipating higher prices). And because most investors are bullish and invested, one assumes that few buyers remain. This leaves more potential sellers than buyers, despite the bullish sentiment. This suggests that prices will trend down, and is an example of contrarian trading.

8.5. Rule-based trading

Rule-based trading is an approach intended to create trading plans using strict and clear-cut rules. Unlike some other technical methods and the approach of fundamental analysis, it defines a set of rules that determine all trades, leaving minimal discretion. The theory behind this approach is that by following a distinct set of trading rules you will reduce the number of poor

decisions, which are often emotion based. For instance, a trader might make a set of rules stating that he will take a long position whenever the price of a particular instrument closes above its 50-day moving average, and shorting it.

8.6. Indicators

We see in table 3 a list of the technical indicators by categories

Table 4: Technical indicators by categories

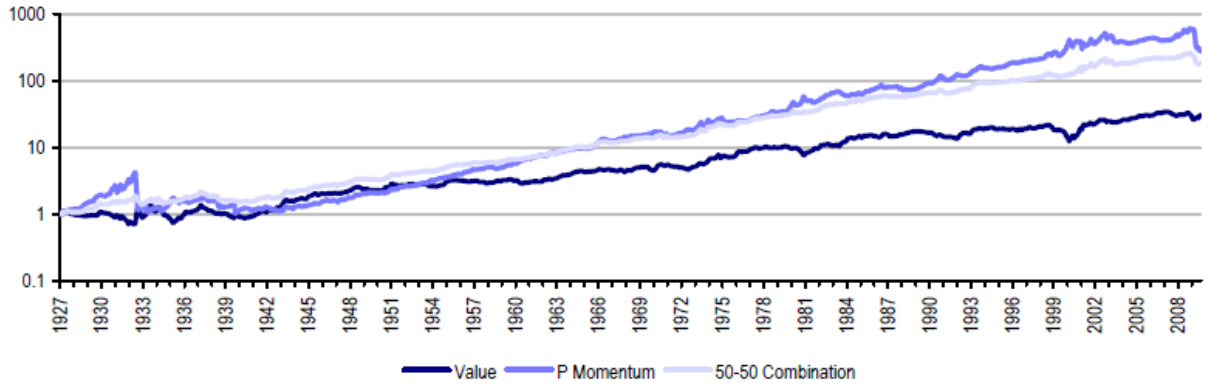
Volatility Indicators	Momentum Indicators	Cycle Indicators
Average True Range	Accumulation Swing Index	Cycle Lines
Bollinger Bands	Chande Momentum Oscillator	Detrended Price Oscillator
Commodity Channel Index	Commodity Channel Index	Fibonacci
Moving Average (Variable)	Dynamic Momentum Index	Fourier Transform
ODDS Probability Cones	Intraday Momentum Index	MESA Sine Wave Indicator
Relative Volatility Index	Linear Regression Slope	
Standard Deviation	MACD	**Support and Resistance Indicators**
Standard Error Bands	Mass Index	Andrew's Pitchfork
Volatility, Chalkin's	Momentum Indicator	Envelope
	Price Oscillator	Fibonacci Arcs, Fans, Retracements
Market Strength Indicators	Price Rate-Of-Change	Gann Lines, Fans, Grids
Accumulation-Distribution	Random Walk Index	Projection Bands
Chaikin Money Flow	Range Indicator	Projection Oscillator
Chaikin A/D Oscillator	Relative Momentum Index	Quadrant Lines
Demand Index	Relative Strength Index	Speed Resistance Lines
Ease of Movement	Stochastic Momentum Index	Tirone Levels
Herrick Payoff Index	Stochastic Oscillator	Trendlines
Klingler Oscillator	Swing Index	
Money Flow Index	Williams' %R	
Moving Average (Volume Adjusted)	Williams' Accum-Distribu	
Negative Volume Index	**Trend Indicators**	**Reversal**
On Balance Volume	Commodity Selection Index	Moving averages crossovers
Open Interest	DEMA	
Positive Volume Index	Directional Movement	
Price Volume Trend	Forecast Oscillator	
Trade Volume Index	Linear Regression Indicator	
Volume	Linear Regression Slope	
Volume Oscillator	Linear Regression Trendline	
Volume Rate-Of-Change	MACD	
	Moving Averages (all methods)	
	Parabolic SAR	
	Performance	
	Polarized Fractal Efficiency	
	Price Oscillator	
	Qstick Indicator	
	r-squared	
	Raff Regression Channel	
	Standard Deviation Channel	
	Standard Error	
	Standard Error Bands	
	Standard Error Channel	
	Vertical Horizontal Filter	

We show in this thesis the indicators and not the chart techniques, the reason is simple while the indicators have a mathematical definition and can be backtested to find if there are useful in a given context and situation, chart patterns are hardly backtestable. By definition we would avoid techniques that cannot be analyzed in a statistical rigorous way.

9. Statistical Arbitrage Applications – Momentum and value analysis

We will analyse in depth now the value and momentum factors in the equities US context.

Figure 32: Value and momentum in the US, January 1927-August 2009.



Source: Kenneth French website, author calculations. The chart shows a total return index calculated for a long short strategy on a log scale. Momentum is based on a formation period of 12 months and a lag of two months. For the value strategy stocks are sorted by book to price with annual rebalancing.

Figure 32 shows that, in the wake of the credit crisis of 2008, value is back on trend, while momentum is affected significantly but not any more than in the early 1930s. A naive combined strategy, obtained by giving equal weight to momentum and value, shows just a modest drop in performance. The chart is reminiscent of the results obtained by Asness (2008), who analysed the performance of a simple combination of value and momentum using data as of April 2008.[94]

It is worth noting that the data displayed in Figure 32 illustrates very basic implementations of the two strategies. Performance can be improved upon by:

(1) measuring value better

(2) rotating between value and momentum

(3) using sector-adjusted strategies[95]

The initial look at the available empirical evidence suggests that the recent underperformance is in line with history. In the first part of this research report we will present the results of a thorough statistical analysis which indicate that no support can be found in the US data for the idea that basic quant signals have stopped working.

From the point of view of a portfolio manager, the empirical issue that we are investigating in this PhD thesis has far-reaching consequences on the investing process. A permanent deterioration of the performance of price momentum, for example, is consistent with the idea that momentum is an anomaly, generated by behavioural effects. Rationalists have argued that any anomaly is short lived, because the economic agents will identify it and profit from it, until it is completely eliminated. This position is expressed difficult to give such an interpretation to the recent debacle of value strategies, since many studies have interpreted the value premium as a reward for risk

(see Fama and French, 1995[96], or Petkova, 2006[97]). Why would such a risk premium disappear in the long term?

If we go back to the late 1920s, the current underperformace no longer seems unprecedented There is considerable scope for improvement of the performance of simple quant strategies. The theme of the future of Quant has received considerable attention recently, for example in Asness (2008) and Sorensen (2009)[98]. While the former emphasises the importance of diversification in any quantitative investment process, the latter predicts the success of new approaches to quant investing that can incorporate more fundamental inputs and exploit new proprietary datasets.

In this PHD thesis, we start with a simple characterisation of Quant. Quantitative processes should provide long short strategies that: (1) generate significant alpha and (2) have a low exposure to market risk. In our work we investigate whether (1) and (2) have been true over the last ten years or so. Also, we look at the returns obtained by hedge funds which market themselves as market neutral equity long short.

The main questions that we address are:

- Is it true that the market is so crowded with quant funds that profits have evaporated in recent times?
- Can we identify a structural break over the last ten years or so, implying that (1) or (2) have broken down?

Before presenting the analysis, we shall give a brief overview of the main conclusions. First, we find no evidence in the data that value and price momentum strategies have stopped working over the last ten years or so. The recent episode of underperformance is in line with the historical behaviour of quant signals. In addition, we detect significant changes in the exposure of simple quantitative strategies to market risk. However, we argue that a combination of value and momentum signals, coupled with a simple mechanism to adjust leverage dynamically, is enough to obtain both a stable alpha and market neutrality simultaneously.

9.1. The historical performance of value and price momentum

9.1.1 The big picture: Value and price momentum in the US

Throughout this note we focus on US data for two reasons. First, the performance of simple quantitative strategies in the US has been studied more than in any other market. In addition, reliable time series of US stock returns are available over a longer sample period compared to other markets.

Before we describe our methodology, a brief digression on the definition of price momentum is useful. Each momentum strategy is typically identified by a triplet (I, J, K), where I is the length of the formation period, K is the length of the holding period and J is the length of the gap between the two periods.

Here price momentum is defined as (12, 2, 1): past 12 months, skipping the last one and for 1 month holding period, rather than (6, 1, 1). It is worth noting that historically a (12, 2, 1) strategy has tended to outperform a (6, 1, 1) one in most markets (see e.g. Sefton and Scowcroft, 2005[99]). If the holding period is just one month long, it pays to select a longer formation period. Jegadeesh and Titman (1993)[100] chose a holding period of six months instead in their seminal work.

Our composite value indicator, as usual, combines four standardised scores based on book to price, PE, the ratio of EBIT to enterprise value and dividend yield. We used it alongside the traditional indicator based simply on price to book.

We also produce sector neutral strategies by ranking stocks within each of the ten GICS industries (and then aggregating the resulting High/Low portfolios). The 50-50 combinations simply assume that at the end of each month we rebalance the portfolio by allocating 50% to a value (High/Low price to book) strategy and 50% to price momentum (12, 2, 1). The High (Low) portfolio is made up of the top (bottom) third of names in each ranking. Fama and French (1992)[101] used the top and bottom 30% instead. Moreover, we mitigate the size effect by sorting on market cap first. The return to each style portfolio6 is obtained as the simple average of the return to a large cap (Russell 1000) and a small cap (Russell 2000) portfolio. This does not seem to affect the results significantly.

Table 5 shows the gross Sharpe ratios for several alternative value, momentum and combined strategies. As usual we assume for each long short strategy that the portfolio has three components, each of same size: cash, long and short. The cash component earns the riskfree rate. The main conclusions are in line with our previous results:

- Imposing sector neutrality improves the Sharpe ratio for all strategies
- The simple combination of value and price momentum outperforms the individual components
- (12, 2, 1) price momentum performs far better than the (6, 1, 1) version
- Composite value outperforms Book to Price.

Table 5: Sharpe ratios of simple quant strategies in the US 1/1980-6/2009

	Sharpe ratio
Price momentum 6.1.1	0.02
Price momentum 6.1.1 (SN)	0.07
FF SMB	0.14
HML Book to price	0.24
HML Book to price (SN)	0.35
Market minus riskfree	0.36
Composite value	0.37
FF HML Book to price	0.38
FF price momentum 12.2.1	0.50
Price momentum 12.2.1	0.51
Composite value (SN)	0.51
Price momentum 12.2.1 (SN)	0.66
FF 50-50 value plus momentum	0.68
50-50 value plus momentum	0.90
50-50 value plus momentum (SN)	1.31

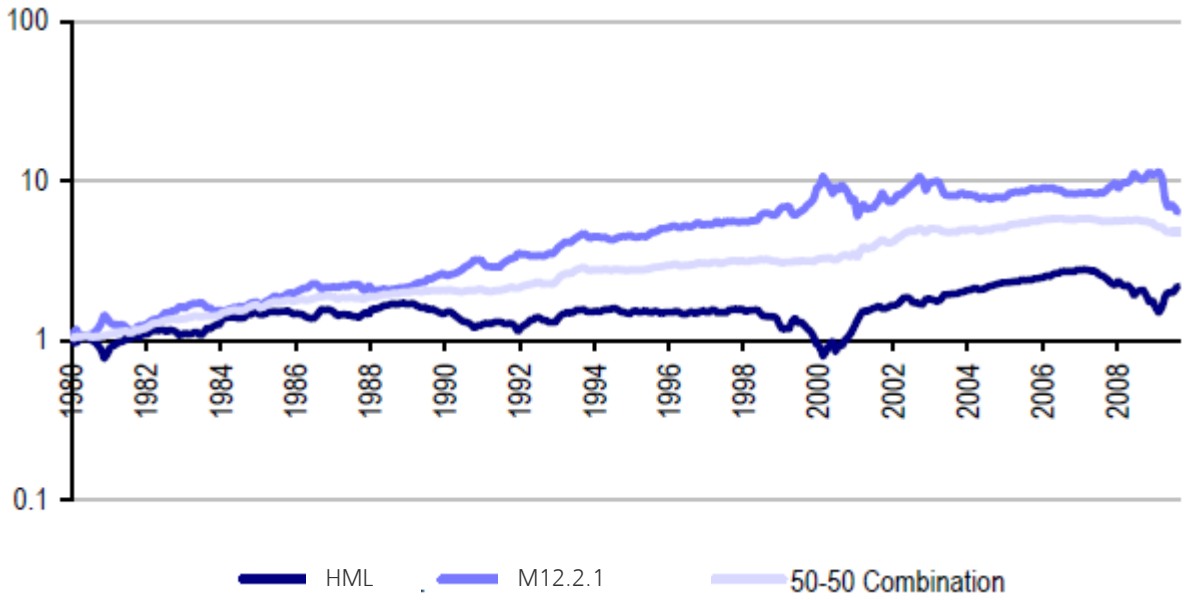
Source: Own estimates, Kenneth French website, CRSP. The universe is Russell 3000 for the styles, CRSP for the Fama-French portfolios. SN stands for 'sector neutral.'

9.1.2 A closer look at historical performance Portfolio returns

Figure 32 suggests that price momentum and particularly value are below trend but no structural break seems to have occurred over the past decade. A similar picture emerges for the period 1980-2009 if we use a more investable universe like Russell 3000 (Figure 33). Asness (2008) showed a very similar chart, although his sample period ended in April 2008.

The visual impression is that the trend has not broken down...

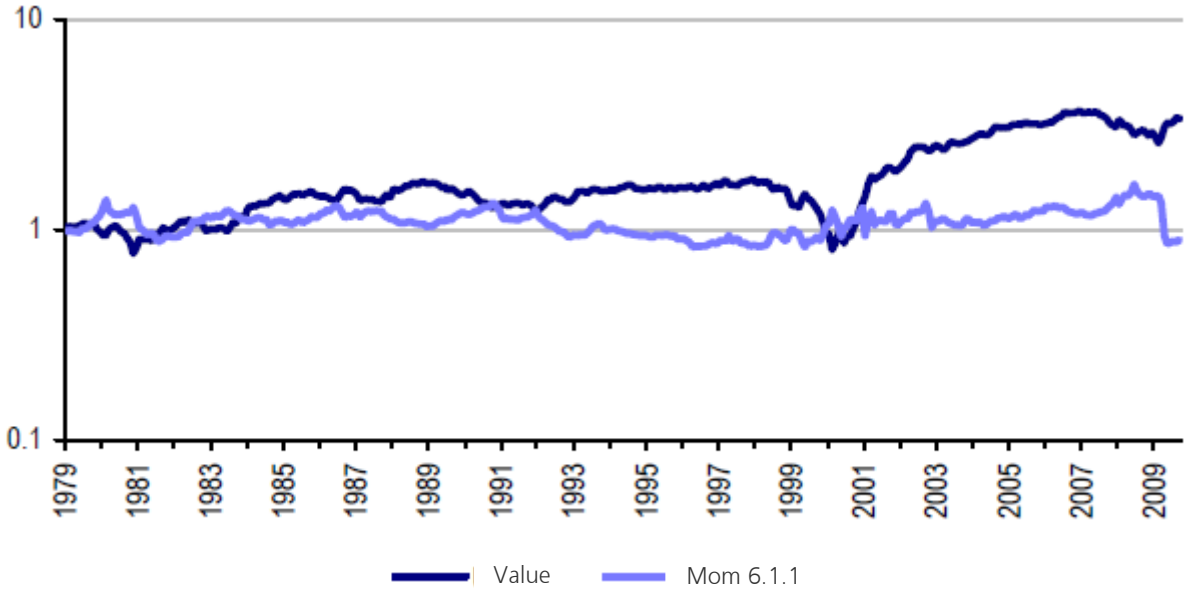
Figure 33: Performance of value and price momentum, Russell 3000, 1/1980 – 8/2009



Source: Author estimates. HML is a long short strategy obtained by sorting stocks on book to price. Momentum is defined as in Figure 32. The universe is Russell 3000.

Figure 34 displays the performance of two alternative style indices that we regularly monitor, i.e. our medium term (6, 1, 1) price momentum index and our composite value index. The performance of this particular implementation of momentum was poor in the early 1990s but it has been in line with the one in Figure 33 since 1999. Overall, using a different definition of value and momentum confirms the initial impression that a breakdown in the main quant strategies does not seem to be warranted by the data. In particular, the idea that the profitability of traditional quant long short strategies has been arbitraged away does not seem to be supported by the available evidence, at least in the US.

Figure 34 Performance of value and price momentum, Russell 3000, 1/1980 – 8/2009



Source: Author estimates. Here value is our composite value style, while momentum uses a 6-month formation period with a one month lag.

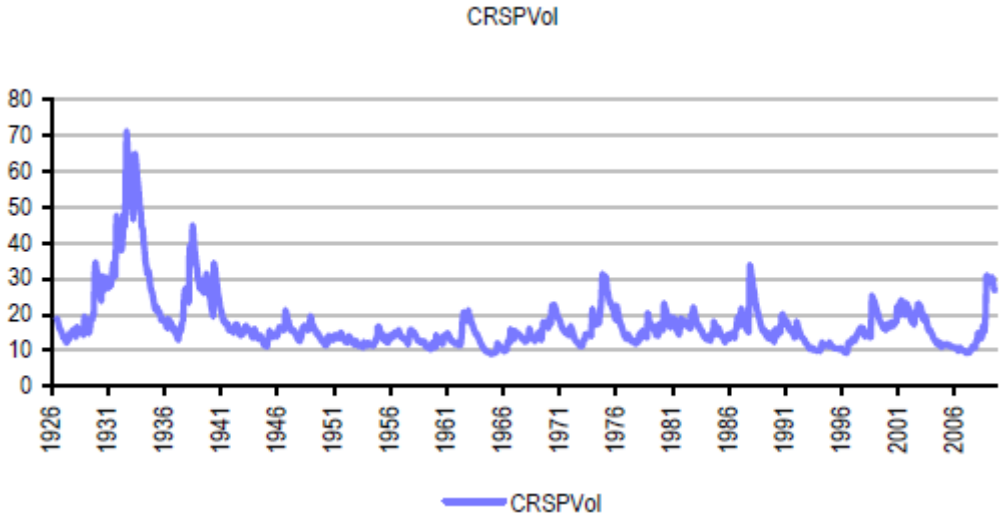
Market risk

If we do not accept that quant strategies have been arbitraged away, then what caused the recent poor performance? The alternative view contends that the current crisis has brought about a temporary dislocation of the markets. How does the credit crisis compare to the previous episodes of market dislocation in the data? It is difficult to answer this question as no single measure of market dislocation can be found for this purpose. However, if we look at volatility as a measure of risk on the US equity market, then by fitting a simple GARCH(1,1) model we can compare the current environment to the historical evidence as far back as 1926 (Figure 35).

The visual impression is that, while the current imbalances appear more significant than the effects of the tech bubble at the end of the 1990s, the surge in volatility triggered by the credit crunch is by no means unprecedented. Even excluding the pre-war period, during which the number of stocks in the US market was much lower than the current one, still the oil crisis and the

1987 market crash have resulted in similar volatility levels. What is interesting to note is that the peak seen in 1932 (at the end of a long upward trend started after the crash in October 1929) corresponds to the worst setback for the momentum strategy in the Fama-French database. Given that it is often argued that the 1929 crisis is the only episode that can be likened to the current crisis in recent history, it is reassuring to observe that that episode coincided with a loss for the price momentum strategy and the magnitude of the loss was much more significant than what we have observed so far between 2008 and 2009.

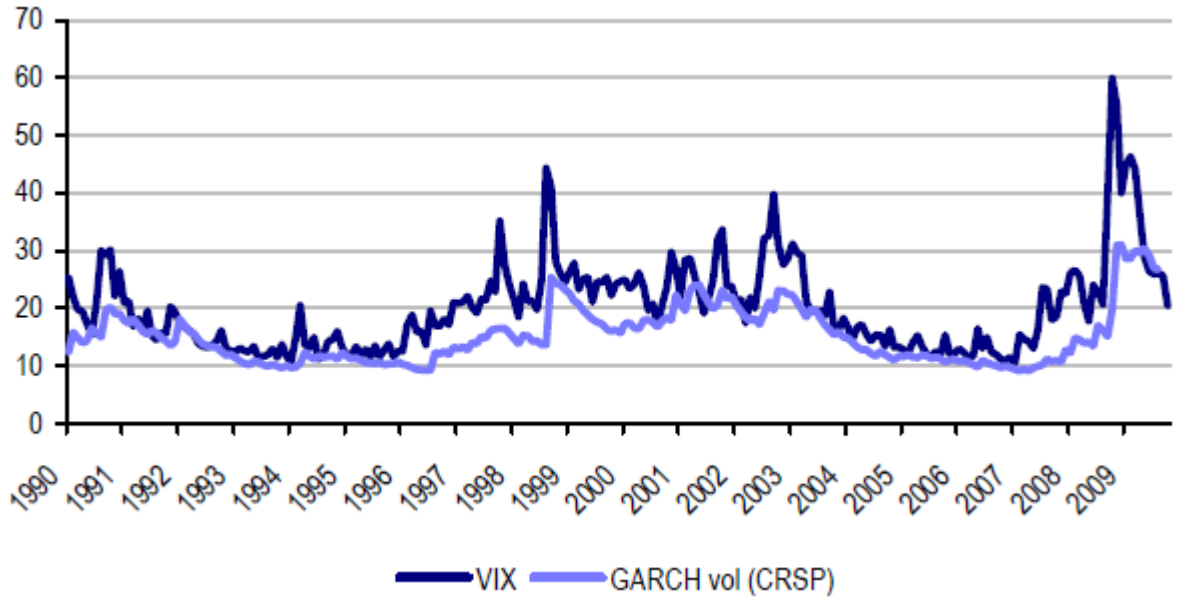
Figure 35: Estimated volatility, US market 7/1926 – 8/2009



Source: Author estimates. Volatility is estimated via a GARCH(1,1) model. The data consists of returns to the value weighted CRSP index.

How does our volatility measure compare with the most common ones, like the VIX index? The estimated conditional volatility from the GARCH model should be lower than the implied because the index is different (S&P 500 is less diversified) and because of the existence of a risk premium. Moreover, because of the autoregressive structure of the GARCH model, any reaction to volatility shocks appears to be lagged by one period. Figure 36 shows that the two measures display very similar patterns, although the end-of-month VIX index is noisier and, as expected, mostly higher than the estimated GARCH counterpart.

Figure 36: Estimated GARCH vs Implied Volatility



Source: Bloomberg, Author estimates. The GARCH volatility series is the same used in Figure 35, albeit the sample period is much shorter.

Hedge fund returns

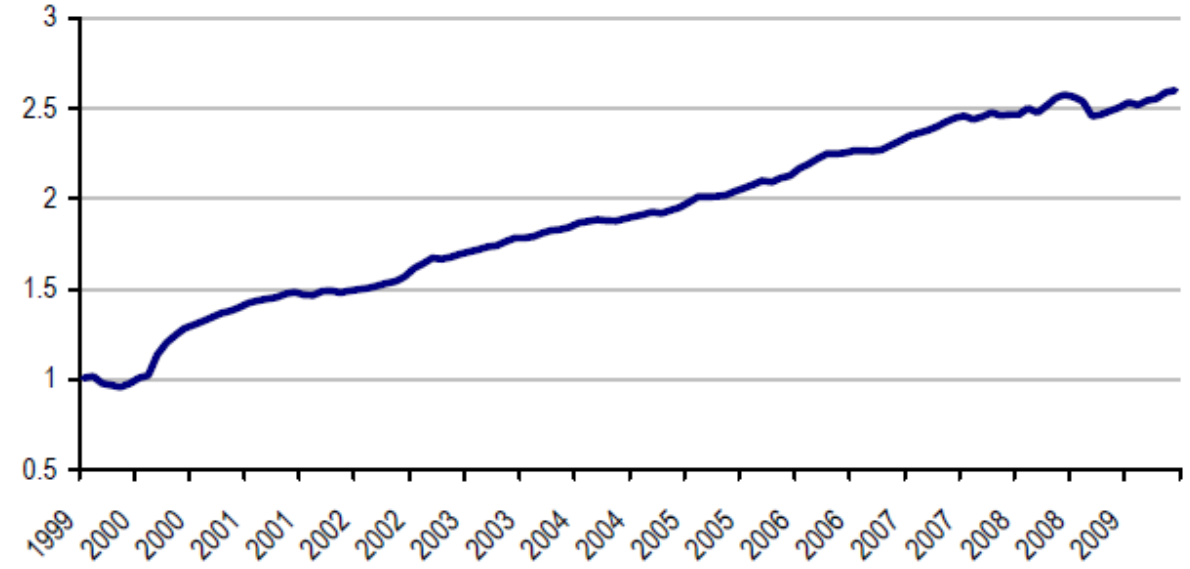
The last statistic that we consider in this section is the average monthly return of hedge funds that are categorised as long short equity, market neutral. The data consists of all the funds available in the HFR universe. While hedge fund strategies are clearly not limited to value and price momentum, one would expect to see a sharp deterioration of the performance if a break in the predictive power of the main quant factors.

Except for three consecutive drops in 2008Q3, the series does not suggest a sharp decrease in the profitability of this asset class over the last ten years.

Sector neutral returns

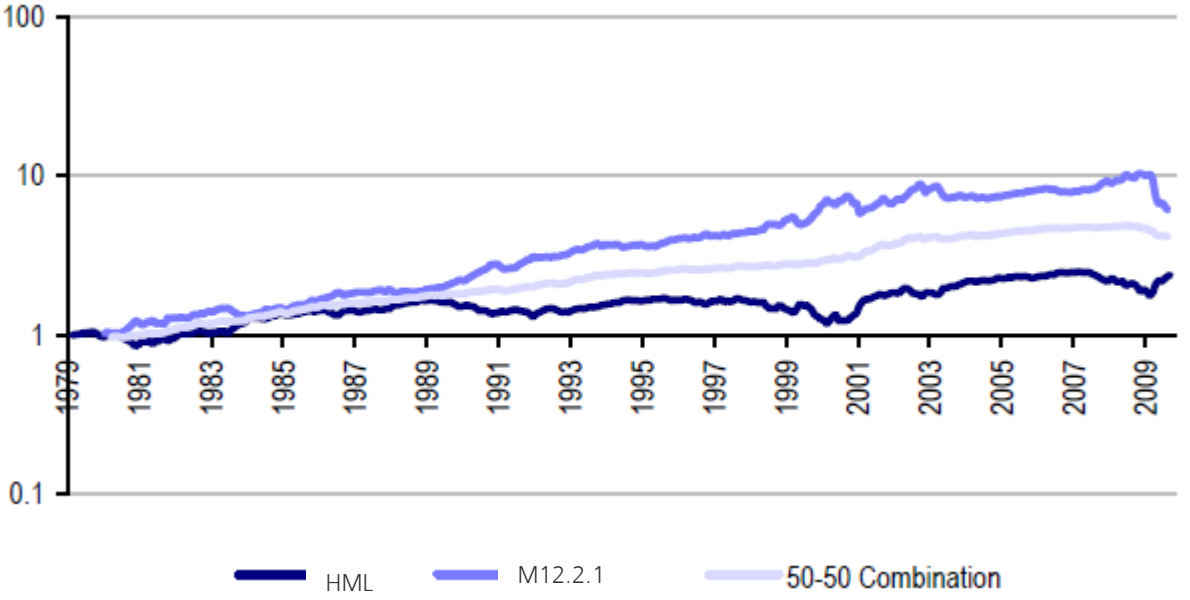
Another important check is carried out by obtaining sector neutral versions of the main value and momentum strategies. To this end we sorted stocks according to the value of each style indicator within each of the ten GICS industries and then aggregated the ten resulting long short portfolios. The overall long short portfolio is market cap weighted. From the plot in Figure 37 no significant differences emerge compared to the simple strategies illustrated in Figure 33, i.e. with no constraints on the sector exposures.

Figure 37: Performance of long short equity market neutral hedge funds



Source: HFR

Figure 38 Performance of value and price momentum, sector neutral, 1/1980 – 8/2009



Source: Author calculations

9.1.3 A formal test: Has alpha disappeared?

The visual impression from the previous section is that no evidence can be found in the data of a breakdown in the positive trend displayed over the long period by the stylised quant strategies taken into account here. This section describes the results of a formal test, devised by Andrews (1993, 2003)[102], which deals with the hypothesis of a structural break in alpha. The advantage of Andrews's test, as argued in the Appendix, is that it is based on the assumption that the date in which the break occurred is unknown – much more methodologically robust than, say, a simple Chow test in which we assume that the breakpoint is known in advance.

The same approach has been used in Viceira (1997)[103] to test for the stability of some popular predictive models of stock returns. Here we deal with the stability of contemporaneous relations among time series of portfolio returns. We adopt the simplest possible model of asset returns, the workhorse of much of modern finance theory (see Campbell, Lo and McKinlay (1997))[104]

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right) = \alpha + \varepsilon_t \quad (208)$$

where α is a constant and ε is a (potentially heteroscedastic and mildly autocorrelated) disturbance term having zero mean. The null hypothesis of our test is that α is constant throughout the sample period. A rejection in Andrews's test suggests either a structural break (the value of alpha has changed abruptly at some point in time) or multiple breaks (unstable alpha).

In order to carry out the test we must select a range of dates in which we believe the structural change might have occurred. Common practice is to use the central portion of the sample so as to leave enough observations on either side to identify the change in parameter values. We leave 25% of the observations, i.e. roughly 90 months, at either end of the sample period, thus allowing for breaks to occur roughly between May 1987 and April 2002. While the recent credit turmoil is left out, because of data limitations, the tech bubble period is included. We can still answer the question: Have value and momentum stopped generating alpha since the end of the tech bubble? It has often been suggested that the growth in assets under management by quant funds in the late 1990s might have significantly reduced the profitability of value and momentum.

Table 6: Test of parameter stability. Has the expected return changed over time?

	HML	12.2.1	Combination
Test stat	**4.60**	**4.87**	**5.11**

Source: The critical values are 6.35, 7.87 and 11.28, respectively, for 10%, 5% and 1% significance. In all cases we fail to reject the null hypothesis, i.e. we do not find evidence of a break in the alpha of of simple stylized quantitative strategies.

.

As can be seen from Table 6, we fail to reject the null hypothesis for value, price momentum and the simple equally weighted combination of the two. From the statistical point of view, we find no evidence that the expected return of simple quantitative strategies has been significantly lower over the last seven years than in the 1980s and 1990s. It is worth stressing that this is a very robust result, since the test takes unknown breakpoints and heteroscedasticity into account. Nevertheless, even at a coarse significance level (10%) we still fail to reject the null.

The same result is obtained by looking at the much longer time series of returns from the Fama-French database. Here the test is more powerful because we only need to drop about 20% of the observations in order to leave roughly 100 data points on either side to identify the change in parameter value. As the test statistics is 5.33, again we fail to reject even at the 10% significance level.The critical value when dropping 20% of the observations is 7.58.

Is the test powerful enough to detect a sharp drop in alpha? This question is particularly important because the uncertainty in the estimation of the expected return is typically very high. The results of a small Monte Carlo study show that if a break had occurred then the probability of correctly rejecting the null would have been high. The details can be found in the Appendix.

9.2. Systematic strategies and market risk: What has changed?

9.2.1 Correlation with market returns

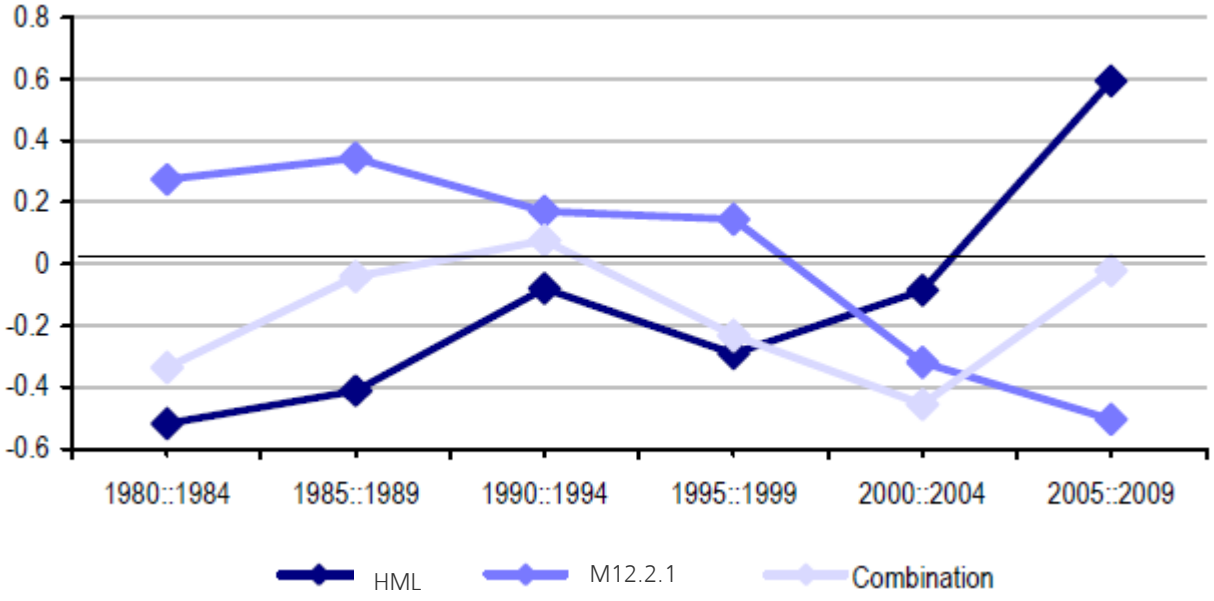
As we argued in the introduction, long short quantitative strategies like value and momentum are understood to have little exposure to market risk. We examined this assumption more closely. The simplest measure of market risk exposure is arguably correlation. We considered the correlation between hedge fund and market returns, using a test that took heteroscedasticity into account: 29% of funds that claimed to follow a market neutral strategy had correlations to the market that was significantly different to zero. Please see the appendix for further details about our tests and results here.

It seems that hedge funds have not been as market neutral as has been assumed.

However, looking at our quantitative strategies, over the full sample period from 1980-2009 the correlation between our value and momentum styles on one side and the market on the other is respectively -0.11 and -0.07, and neither of these are significantly different to zero. These strategies are what Patton (2009)[105] terms correlation neutral.

But the correlations appear to change noticeably over time (Figure 39). A long term pattern seems to emerge from calculating correlations over five year windows: value becomes more and more positively correlated with the market, while momentum, which used to be mildly positively correlated with the market, becomes more and more negatively correlated. The equally weighted composite strategy averages out the two tendencies, so that its own correlation with market returns is most of the time smaller in absolute value than each component's.

Figure 39 Correlation with market returns



Source: Author estimates. Data is monthly for the US.

The picture that emerges from the analysis illustrated in Figure 39 is clear; however do we have enough data to conclude that the correlations of value and momentum with the market have switched sign? The answer to this question can be given by running a formal test of the stability of the correlation coefficient

To run such a test we can set up a linear model which allows for breaks in the relation between asset returns and market returns:

$$\begin{aligned} r_t &= \alpha_1 + \beta_1 r_{Mt} + \varepsilon_t \quad t < \tau \\ r_t &= \alpha_2 + \beta_2 r_{Mt} + \varepsilon_t \quad t \geq \tau \end{aligned} \tag{209}$$

where the break is assumed to have occurred at some date τ and rMt is the return to the market index. An advantage of the methodology described in the Appendix is that is yields, as a byproduct, an estimate of the breakpoint τ.

It is interesting to note that expression (119) can be seen as a CAPM with time varying parameters, similar to the models proposed for example by Jagannathan and Wang (1996)[106] and Ang and Chen (2007)[107]. In their analyses the time variation stems from the evolution of conditioning variables which affect the risk premium and the CAPM beta. The approach has been used, among others, by Lewellen and Nagel (2007[108]) and Petkova and Zhang (2005)[109] in analysing the value premium. Patton (2009), in his study on the market neutrality of hedge funds, concedes that the exposure to market risk of individual hedge funds varies over time but does not incorporate this aspect into his models due to data limitations.

Table 7: Testing for a break in the correlation with market returns

	HML	12.2.1	Combination
Test stat	**52.13**	**42**	**4.44**
Estimated break date	**sep-01**	**Jan-01**	

Source: The critical values are 6.35, 7.87 and 11.28, respectively, for 10%, 5% and 1% significance. Rejecting the null hypothesis means that we find evidence of a break in the correlation between the the strategy returns and market returns

Table 7 displays the results of Andrews's (1993) test. The main conclusion is that both value and momentum appear to have experienced a structural break, while the correlation between the composite strategy and the market appears to have been stable throughout the sample period. The conclusion is true regardless of whether we consider a coarse significance level like 10%, which should make it easier to reject, or a strict criterion like the 1% significance level.

Table 8: Estimated linear models, before and after the break

	HML	Momentum
Pre-break		
Alpha (bp per month)	**45.82**	**61.50**
t-stat	**2.22**	**2.59**
Beta	**-0.31**	**0.23**
t-stat	**-5.99**	**3.42**
Post-break		
Alpha	**31.30**	**16.59**
t-stat *	**(-0.38)**	**(-0.92)**
Beta	**0.47**	**-0.70**
t-stat *	**[7.22]**	**(-6.23)**

Source: The linear model is specified in (1). The dates of the breaks are taken from Table 3. Post break t-stats are for the difference n parameter itself.

In Table 8 we take a look at the linear model before and after the break. The after break exposure to market risk (as measured by the CAPM beta) has significantly increased and switched direction in both cases, changing from-0.31 to 0.47 for value and from 0.23 to -0.70 for price momentum. Changes in the estimated alphas are not statistically significant, as can be seen from the robust tstats.

The point estimates are lower in the post-break than in the pre-break period, but the uncertainty surrounding the estimate of alpha is typically large.

Table 9: Testing for a break in the CAPM alpha

	HML	12.2.1	Combination
Test stat	**3.63**	**4.82**	**6.32**

Source: The critical values are 6.35, 7.87 and 11.28, respectively, for 10%, 5% and 1% significance. Rejecting the null hypothesis means that we find evidence of a break in the alpha of simple stylised quantitatve strategies.

Table 10: Testing for a break in the correlation with market returns and/or CAPM alpha

	HML	12.2.1	Combination
Test stat	**26.10**	**21.00**	**4.15**

Source: The critical values are 6.35, 7.87 and 11.28, respectively, for 10%, 5% and 1% significance. To reject the null hypothesis means that we find evidence of a break either in the CAPM-Alpha or in the CAPM-Beta of simple stylised quantitative strategies (or in both) from Equation (1).

As a robustness check, we tested the hypotheses of a break in the CAPM α (Table 9) and a break in either α or β (Table 10). The results are consistent with our analysis so far: No evidence of a break in the alpha can be found in any of the individual series, while testing for the stability of both parameters leads to a rejection in the case of value and momentum.

Table 11: Test of parameter stability, sector neutral value and price momentum

	HML	12.2.1
Test stat (alpha)	**2.72**	**5.39**
Test stat (beta)	**45.64**	**29.92**

Source: The critical values are 6.35, 7.87 and 11.28, respectively, for 10%, 5% and 1% significance.

We ran the same analysis with sector neutral style returns: The idea is to ascertain whether our results are ultimately driven by sector effects. The conclusions are unchanged compared to the version with no constraints on sector exposures. In fact, as can be seen from Table 11, we reject the hypothesis of parameter instability for the alpha but we cannot reject it for the beta, which measures the exposure to systematic risk.

9.2.2 Market neutrality beyond correlation

We have already found that correlation neutrality is violated. However, what investors dislike is market exposure when markets are falling. In other words, what we should assess, ideally, is the conditional correlation of a portfolio's returns with the market. This aspect is particularly relevant in the light of the findings of the literature on asymmetric dependence in asset returns. A recent contribution, along with a summary of the existing evidence, can be found in Chua, Kritzman and Page (2009)[110].

Downside mean neutrality

This section investigates whether the expected return of our strategies is affected by market returns in times of falling markets. To take a robust approach we must allow for a nonlinear relation between portfolio returns and market return. A simple nonparametric solution consists of using a third order polynomial as the expected return function:

$$\mu_i(r_{mt}) = \beta_o + \beta_1 r_{mt} + \beta_2 r_{mt}^2 + \beta_3 r_{mt}^3 \quad (210)$$

Does the expected return μi depend on the performance of the market as a whole? We then follow Patton (2009) in checking that the derivative

$$E\left[\frac{\partial \mu_i(r_{mt})}{\partial r_{mt}}\right] = \hat{B}_1 + 2\hat{B}_2 E\left[r_{mt} | r_{mt} \le 0\right] + 3\hat{B}_3 E\left[r_{mt}^2 | r_{mt} \le 0\right] \quad (211)$$

is either negative (i.e. the strategy offers a hedge against falling markets) or equal to zero. In practice, we estimate the value of the derivative and then carry out a simple one sided significance test.

Over our sample of hedge fund returns we rejected the null hypothesis of downside mean neutrality for nearly a quarter of funds which claimed market neutrality was part of their strategy and for 54% of our other funds. The results for our value and momentum strategies are presented in Table 10:

Table 12: Testing for downside mean neutrality

	HML	Momentum	Combination
Full sample			
Estimated derivative	**-0.108**	**0.027**	**-0.041**
Standard error	**0.081**	**0.095**	**0.040**
Subsample 1980-1 to 2001-6			
Estimated derivative	**-0.352**	**0.231***	**-0.061**
Standard error	**0.068**	**0.104**	**0.050**

Source: An asterisk indicates that at 95% level we reject neutrality. The test is one-sided.

Conditional on the market experiencing a negative return, prior to 2001 the expected return of value (HML) was negatively related to the market's, i.e. value acted as a hedge (bottom panel of Table 12). The simple combination did not display any significant relation with market returns. Momentum, however, failed to provide market neutrality, as it gave positive exposure to the market conditional on the market experiencing a drop. Once the full sample period is taken into account all significant exposures disappear from the estimation results (top panel). This suggests that the negative downside exposure of value and the positive exposure of momentum switch signs simultaneously, as suggested by Figure 39.

Variance neutrality

Does the volatility of each strategy change when market volatility changes? In other words, have traditional quant strategies offered protection against shocks in volatility? We can formalise the concept of variance neutrality by requiring that

$$Var\left(r_{it} - \mu_i\left(r_{mt}\right) \mid r_{mt}, F_{t-1}\right) = Var\left(r_{it} - \mu_i\left(r_{mt}\right) \mid F_{t-1}\right) \quad (212)$$

where Ft is the information about the historical performance of the strategy up to time t. In words, a strategy is considered variance neutral if the volatility of its returns (conditional on past realisations, ,K i,t−1 i,t−2 r r) is unaffected by the return to the market portfolio.

Here we model the variance of each strategy's returns, 2 i σ, as a stochastic process, potentially displaying ARCH effects. A third order polynomial is used to fit the expected return function, as in the previous section, while a second order polynomial represents the potential links between the variance of portfolio returns and market returns. Finally, the last term in the variance equation is the ARCH effect, which is meant to capture volatility clustering.

$$\begin{aligned} r_{it} &= \mu_i\left(r_{mt}\right) + e_{it} \\ e_{it} &= \sigma_i\left(r_{mt}\right)\varepsilon_{it}, \qquad \varepsilon_{it} \approx N(0,1) \\ \mu_i\left(r_{mt}\right) &= \beta_o + \beta_1 r_{mt} + \beta_2 r_{mt}^2 + \beta_3 r_{mt}^3 \\ \sigma_i^2\left(r_{mt}\right) &= \alpha_o + \alpha_1 r_{mt} + \alpha_2 r_{mt}^2 + \alpha_3 e_{i,t-1}^2 \end{aligned} \quad (213)$$

The results are illustrated in Table 11. The hypothesis of market neutrality corresponds to the case in which α1= α2=0, i.e. the variable mt r does not affect volatility at all. We test for this by simply using a likelihood ratio test. Given that at the strictest confidence level, 99%, the critical value is 9.21, we can reject neutrality in all cases. As both estimated parameters are positive in each case, we can conclude that an increase in market volatility tends to result in an increase in the volatility of quant strategies. It is interesting to note that the phenomenon is less pronounced in the case of the combined value-momentum strategy.

Upon inspection of the individual time series of returns, it becomes apparent that this result is due at least partly to the fact that in the early stages of the technology bubble value and momentum displayed negative correlation. Peaks and troughs of the two series, as a consequence, tend to offset each other over that period when an equally weighted composite is created, thereby reducing volatility.

Table 13: Testing for variance neutrality

	α0 x 10^4	α1 x 100	α2	α3	LR stat
HML	**3.83**	**0.004**	**0.33**	**0.28**	**35.9**
Mom 12.2.1	**3.77**	**0.661**	**0.52**	**0.28**	**55.4**
50-50 Combination	**1.4**	**0.001**	**0.03**	**0.27**	**10.8**

Source:The first four columns contain the estimates of model (2). The last column contains the statistic of a likelihood ratio test of the null hypothesis that each strategy is variance neutral with respect to the market, i.e. α1= α2=0. The critical values are 4.61, 5.99 and 9.21 respectively at 10%, 5% and 1% level.

VaR neutrality

We now turn to the tail of the portfolio return distribution. The failure of our variance neutrality tests suggests that the occurrence of extreme events should be affected by market volatility. Here we consider a formal framework to test if indeed this is the case. The definition of VaR neutrality is analogous to the definition of variance neutrality presented in the previous section. We require that:

$$VaR_{\alpha\%}\left(r_{it}-\mu_i\left(r_{mt}\right)\middle|\ r_{mt},F_{t-1}\right)=VaR_{\alpha\%}\left(r_{it}-\mu_i\left(r_{mt}\right)\middle|\ F_{t-1}\right) \quad (214)$$

The natural choice, given the relatively large size of our monthly dataset, is to run directly a quantile regression focussing on the lower tail of portfolio returns.Recall that the VaRα% is defined as (the negative of) the α% quantile of the return distribution. A very accessible introduction to quantile regression can be found in the initial chapter of Koenker (2005)[111].
This can be written as a test for the model:

$$\begin{aligned} r_{it} &= \mu_i\left(\varepsilon_{it}\right)+e_{it} \\ VaR_{\alpha\%}\left(e_{it}\right) &= \delta_o+\delta_1 r_{mt}+\delta_2 r_{mt}^2 \end{aligned} \quad (215)$$

where we are interested in the hypothesis δ1= δ2=0, i.e. that the VaR of our portfolio is unaffected by market returns and squared market returns.

Table 14: Quantile regression results

	δ1			δ2		
		Confidence Interval			Confidence Interval	
	Estimate	Upper Bound	Lower Bound	Estimate	Upper Bound	Lower Bound
HML	**0.05**	**0.29**	**-0.14**	**4.72**	**13.34**	**1.50**
Mom 12.2.1	**0.02**	**0.37**	**-0.14**	**8.67**	**21.19**	**3.88**
50-50 Combination	**0.11**	**0.15**	**0.05**	**2.18**	**5.45**	**0.64**

Source: The 95% confidence intervals are asymmetric around the point estimates.
The quantile of interest is the 10% VaR of each strategy

The results are displayed in Table 14 Because of the nature of quantile regression, confidence intervals are not symmetric around the point estimate; the table simply shows lower and upper bound of each interval. The results are not surprising: all estimates of δ2 are positive and statistically significant, thus suggesting that increases in market volatility result in higher levels of tail risk (i.e. a higher 10% VaR).

9.2.3 The quest for market neutrality

We have seen in the previous sections that simple quant strategies achieve correlation neutrality and downside mean neutrality, i.e. their expected returns do not appear to be significantly

affected by market dynamics, not even if one focuses on periods of falling markets. However, the volatility of simple quant strategies does seem to be affected by market dynamics.

A simple tool that can be employed by long short portfolio managers in order to mitigate such dependence is leverage. Intuitively, we can think of a simple mechanism whereby the fund is deleveraged when volatility is predicted to be high over the next period. This section considers a very simple implementation of such a strategy and assesses its performance.

In order to keep our approach as simple as possible, we have used a GARCH(1,1) model to simulate a predictor of volatility over the sample period:

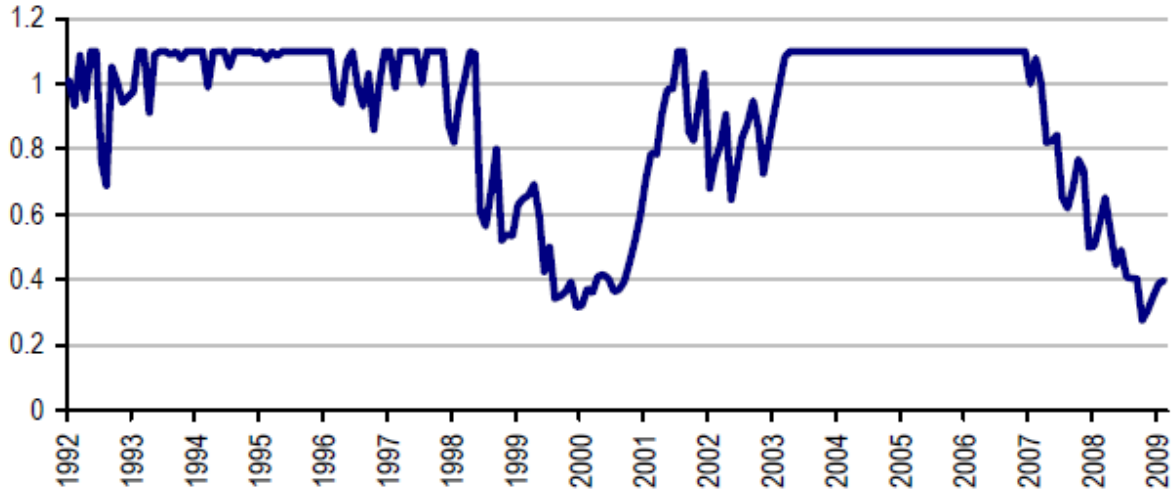
$$\begin{aligned} r_{it} &= \mu_{it} + \sigma_{it}\varepsilon_{it}, \qquad \varepsilon_{it} \approx N(0,1) \\ \sigma_{it} &= \varpi_i + \alpha_1 \varepsilon^2_{it-1} + \beta_i \sigma^2_{it-1} \end{aligned} \tag{216}$$

where α, β and ω are constant.

In practice, we have re-estimated the model each month (leaving 150 observations for the initial estimation) and derived the one month ahead predicted volatility of each strategy.

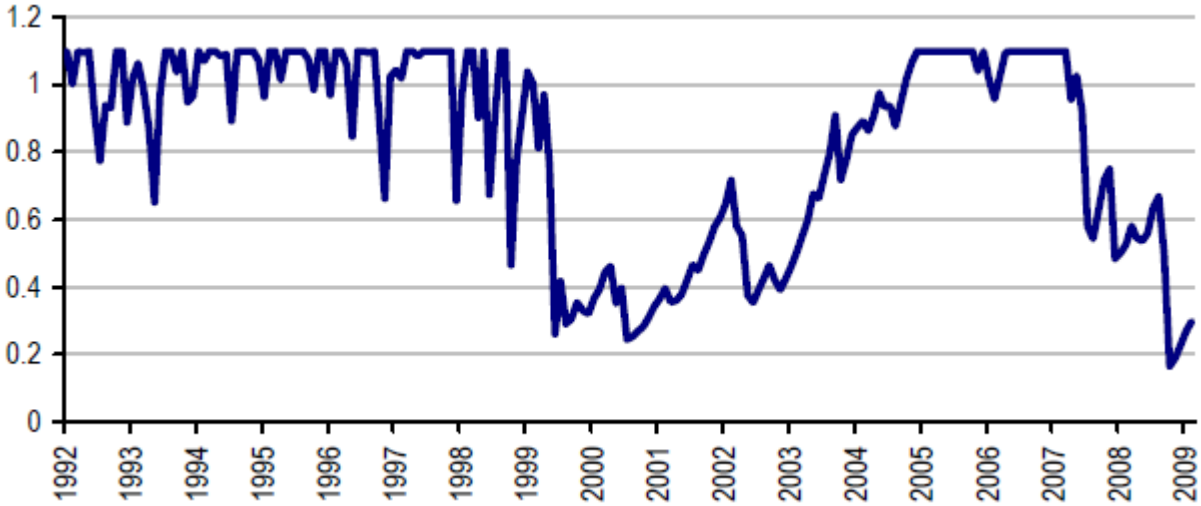
Then, given these volatility predictions, we have calculated the level of leverage necessary to keep the predicted GARCH volatility of each fund constant. More precisely, we assume that if the cash component of the portfolio is worth x, then the equity exposure consists of a long portfolio and a short portfolio each of value ax. The coefficient a is determined so as to keep the (predicted ex ante) volatility constant over time. The strategy is then adjusted by imposing an arbitrary ceiling of 110% on the value of the long position in terms of the cash holdings (i.e. we cannot borrow any more than 10% of the value of the assets held).

Figure 40 : HML .Dynamically adjusted leverage, value portfolio.



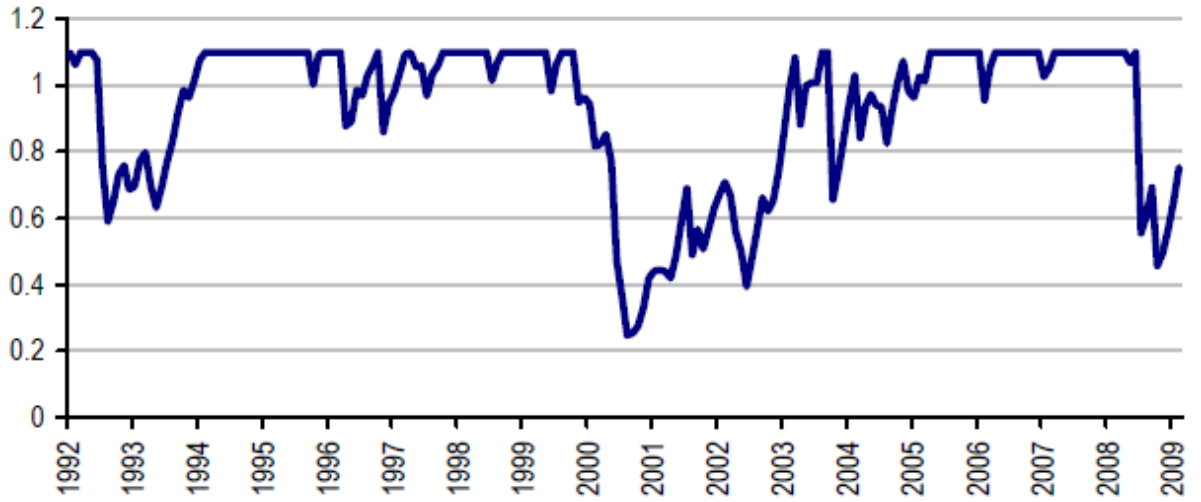
Source: Author calculations.

Figure 41 Mom12.2.1 .Dynamically adjusted leverage, value portfolio.



Source: Author calculations.

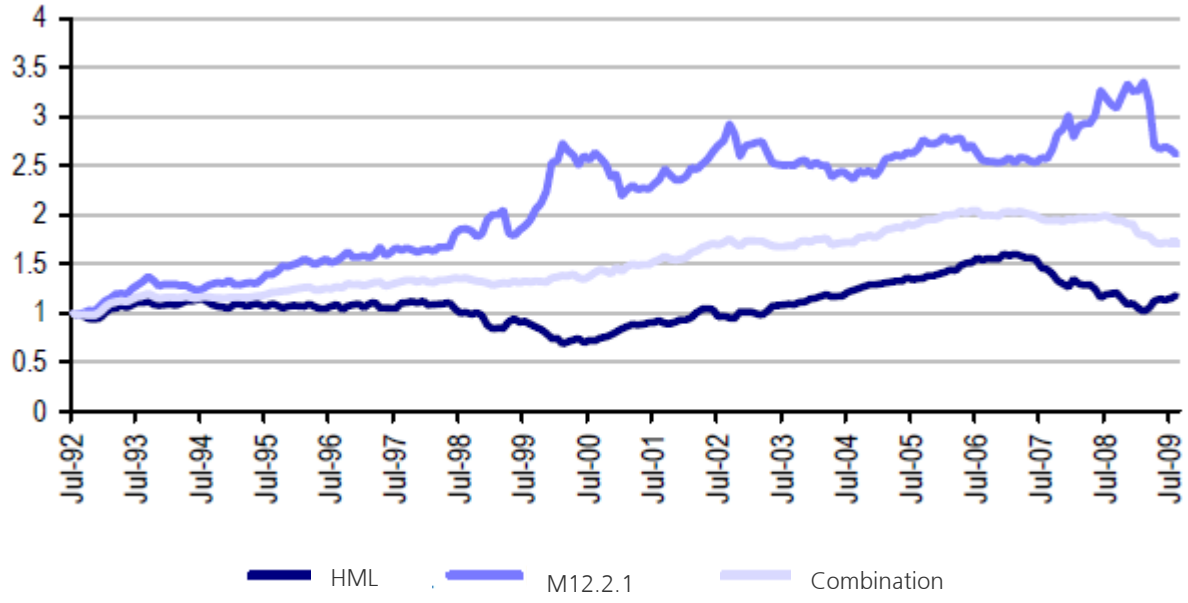
Figure 42 Combination. Dynamically adjusted leverage, value portfolio.



Source: Author calculations.

Figures 40 through 42 show the time varying leverage calculated from out of sample GARCH forecasts of the return volatility of each portfolio. In all cases, the episodes that prompted a dramatic deleveraging of the portfolio are the bursting of the tech bubble and the current crisis. It is clear from the plots that the combined value-momentum strategy experiences a surge in volatility later than the individual strategies do and, as a result, it requires the portfolio manager to deleverage later. Similarly, the volatility of the value portfolio appears to be more erratic than the volatility of price momentum. This is probably due to the occurrence of extreme returns in the latter, as extreme returns tend to generate spikes in volatility when fed to an autoregressive model like GARCH.

Figure 43 Performance of quant strategies with time varying leverage



Source: Author calculations.

Even after dropping 150 monthly observations at the beginning of the sample period, we still have 205 months, between August 1992 and August 2009, to assess the variance neutrality of the modified strategies. Along the lines of the analysis carried out in the previous section, we estimated GARCH models on the returns and used the results to test the hypothesis that the volatility of quant strategies is unaffected by market events.

Table 15: Testing for variance neutrality, strategies with time varying leverage

	α0 x 10^4	α1 x 100	α2	α3	LR stat
HML	**4.00**	**0.350**	**0.19**	**NA**	**21.3**
Mom 12.2.1	**4.60**	**0.447**	**0.28**	**0.00**	**31.0**
50-50 Combination	**1.62**	**0.001**	**0.02**	**0.12**	**5.6**

Source:The first four columns contain the estimates of model (1). The last column contains the statistic of a likelihood ratio test of the null hypothesis that each strategy is variance neutral with respect to the market, i.e. α1= α2=0. The critical values are 4.61, 5.99 and 9.21 respectively at 10%, 5% and 1% level.

Table 15 can be directly compared to Table 11. The third parameter, which represents the magnitude of ARCH effects, is now much smaller and, in the case of the value portfolio, insignificant. Another difference that can be found between the two tables concerns the third parameter, i.e. the coefficient of the square market return term. Its size is drastically reduced once we allow for time varying leverage. The changes in α2 and α3 are offset by an increase in the estimated constant across the board. These findings suggest that the mechanism to adjust leverage based on predicted volatility seems to have worked in the right direction.

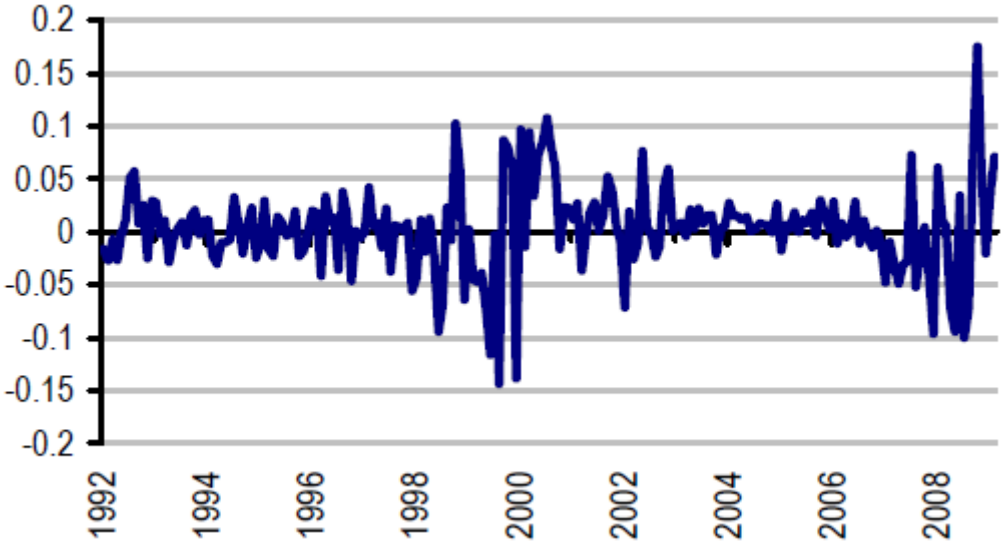
The same impression results from a comparison of the time series of returns with and without the leverage adjustment (Figures 44 through 49). In each pair of plots, the one on the left shows two clear clusters of high volatility, corresponding to the tech bubble and the current credit crisis.

Once we adjust the amount of leverage dynamically, the spikes in volatility almost disappear. But do we have enough evidence to state that this effect is statistically significant?

We repeat the test of market neutrality with the new versions of our basic quant strategies

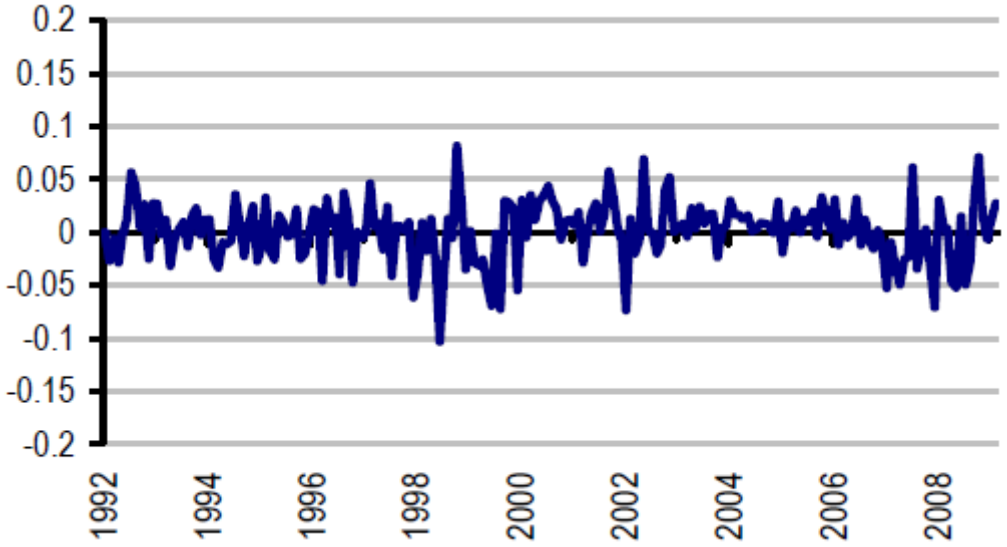
Graphically, the leverage adjustment seems to work in the right direction (volatility becomes more stable over time)...

Figure 44: HML. Returns of the value portfolio, unadjusted leverage.



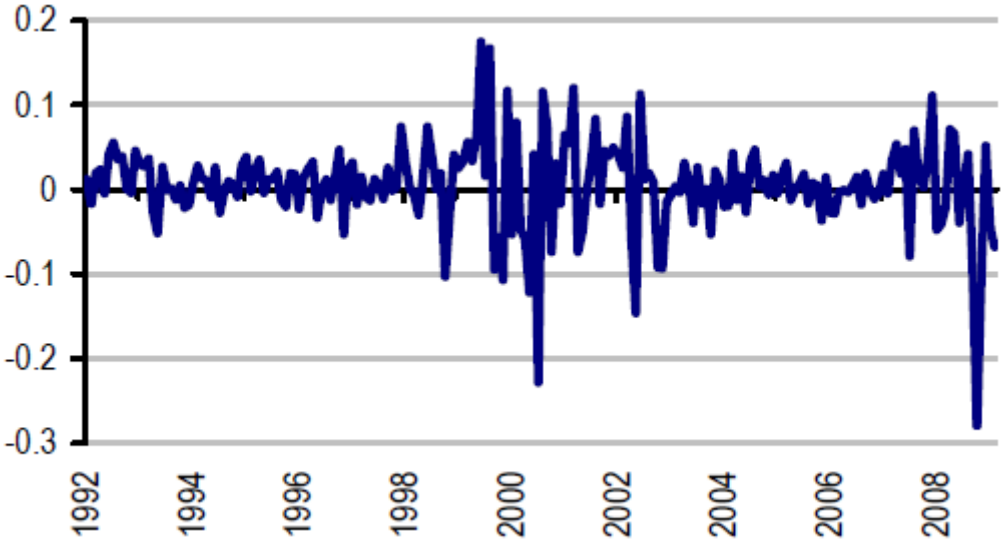
Source. Author calculations.

Figure 45 HML. Returns of the value portfolio, adjusted leverage



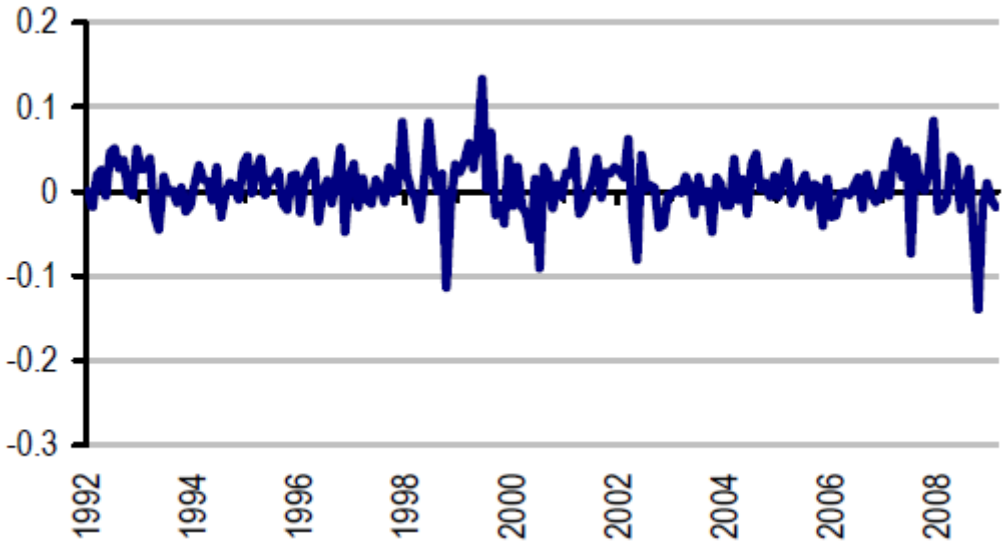
Source. Author calculations.

Figure 46: Mom 12.2.1. Returns of the value portfolio, unadjusted leverage



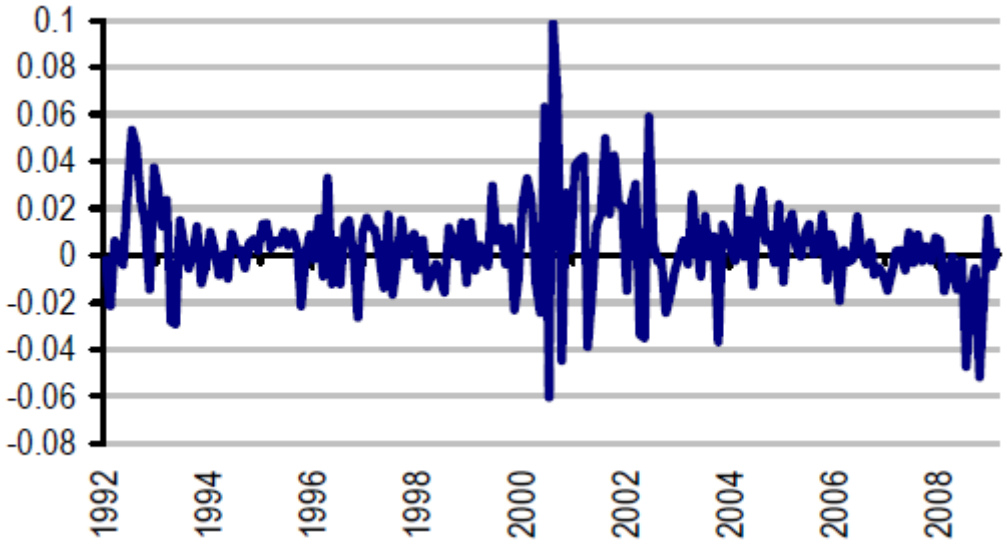
Source. Author calculations.

Figure 47 Mom 12.2.1. Returns of the value portfolio, adjusted leverage



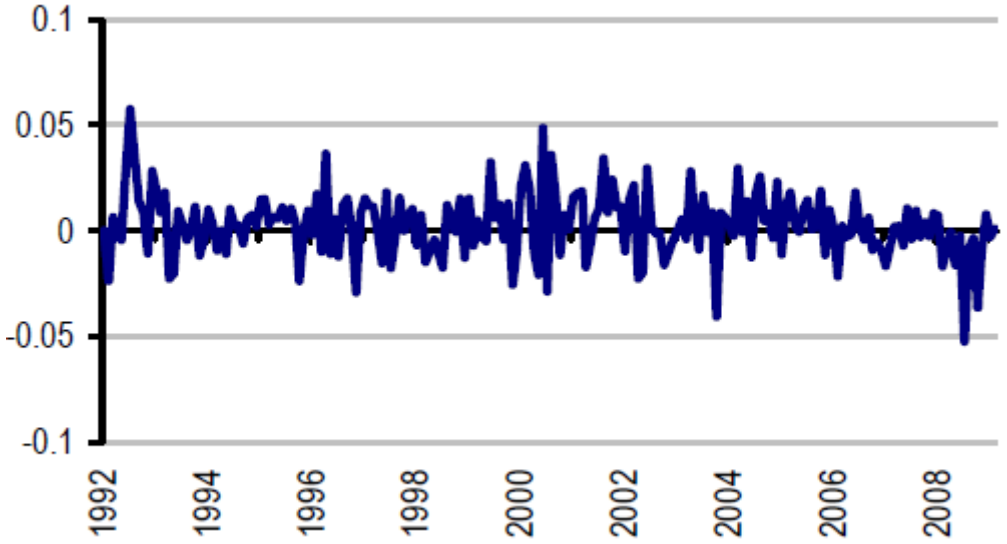
Source. Author calculations.

Figure 48: Combination. Returns of the value portfolio, unadjusted leverage



Source. Author calculations.

Figure 49: Combination. Returns of the value portfolio, adjusted leverage



Source. Author calculations.

The last column in Table 15 answers the question, as it reports the test statistic for a test of the null hypothesis that market returns do not affect the volatility of each portfolio. The critical value is 5.99 and therefore we reject the null for the individual components of the strategy, i.e. value and momentum. Adjusting the leverage according to this simple mechanism is not enough to ensure variance neutrality. However, the combined strategy yields a LR test statistic of 5.6, which indicates that it is indeed variance neutral.

Does the leverage adjustment affect performance? In other words, do we hinder the profitability of quant strategies by deleveraging when risk increases? Table 16 reports the gross Sharpe ratios of both versions of each strategy. The figures in the first column are different from the ones reported in Table 1 because of the different sample period. By dynamically adjusting the amount

of leverage we would have penalised the value portfolio, whose Sharpe ratio would have decreased from 0.26 to 0.15 (second column). However, the momentum strategy would have benefitted from the leverage adjustment. The two effects almost exactly offset each other, so that the performance of the equally weighted combination of value and momentum is virtually unchanged.

Table 16: Performance when leverage is adjusted dynamically

	Sharpe ratio	Sharpe ratio, dynamic leverage
HML	**0.26**	**0.15**
Mom 12.2.1	**0.29**	**0.56**
50-50 Combination	**0.65**	**0.63**

An even stronger result is obtained by repeating the set of quantile regressions on the series of returns after dynamically adjusting the amount of leverage. The estimated parameters are displayed in Table 17. For comparison, the corresponding estimates obtained from the unadjusted returns can be found in Table 14. As regards value and momentum, the 95% confidence interval around the estimate of δ2 has clearly shifted to the left towards the origin, indicating that the estimated effect of market returns on the tail of the distribution of portfolio returns is weaker. By combining value and momentum we obtain tail neutrality:
The confidence intervals around both δ1 and δ2 now include zero, which indicates that neither effect is statistically significant.

Table 17: Quantile regression results, adjusted leverage

	δ1			δ2		
		Confidence Interval			Confidence Interval	
	Estimate	Upper Bound	Lower Bound	Estimate	Upper Bound	Lower Bound
HML	**0.07**	**0.20**	**-0.15**	**4.36**	**9.76**	**0.38**
Mom 12.2.1	**0.14**	**0.30**	**-0.10**	**4.50**	**10.45**	**0.86**
50-50 Combination	**0.06**	**0.07**	**-0.15**	**1.42**	**6.84**	**-0.23**

9.3. Momentum in a multiasset class context

Momentum applied across a set of 60 futures and forward contracts on different commodities, equity indices, currencies and government bonds. This Strategy generated positive hypothetical returns in each of the 60 contracts over a period of more than two decades. Moreover, since these 60 trend-following strategies have exhibited low correlation to each other, the strategy produced strong risk adjusted returns by diversifying across all of them. One of the most powerful attributes of this Simple Momentum Futures Strategy is depicted in Figure 50.

Figure 50: Sharpe ratios different assets

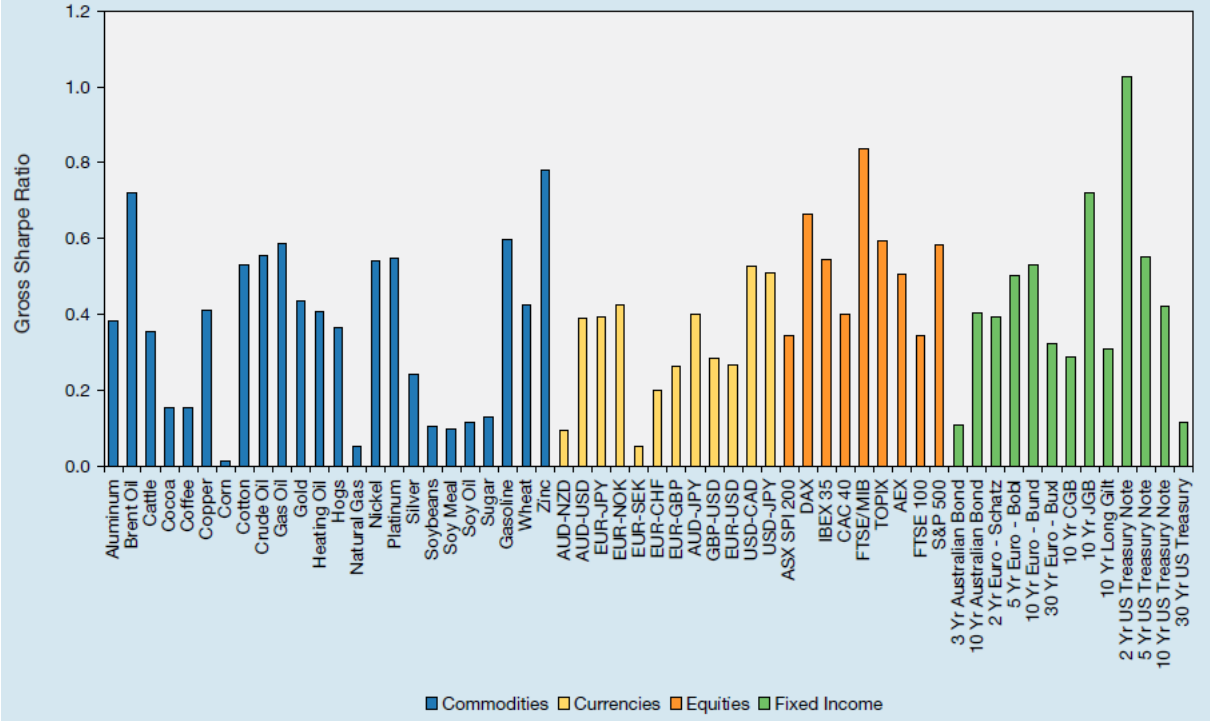


Figure 51: Portfolio results

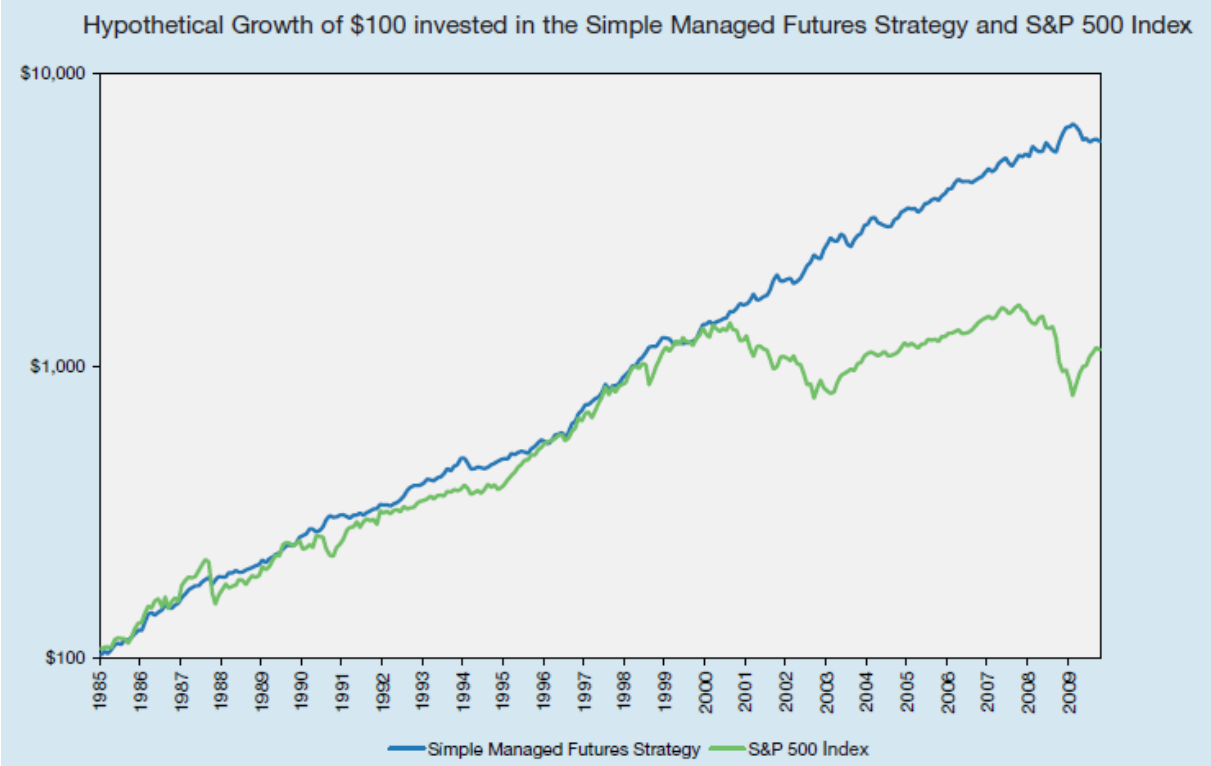


Table 18: Performance statistics gross of transaction costs

	1985 – 2009
Annualized Return	17.8%
Annualized Standard Deviation	9.3%
Sharpe ratio	1.42
Worst Month	-6.4%
Worst Drawdown	-13.3%

When the hypothetical returns to the Strategy are plotted against the returns to the stock market Figure 51, the Strategy exhibits a "smile." In other words, the Strategy produced its best performance in extreme up and extreme down stock markets.

Certainly any investment that produced positive returns in bear markets would have been beneficial to most investors' portfolios, but why has a simple trend-following strategy exhibited this kind of return characteristic? One reason is that most extreme bear or bull markets have not happened overnight, but instead have occurred as the result of continued deterioration or improvement in economic conditions. In bear markets, managed futures strategies position themselves short as markets begin to decline and can profit if markets continue to fall.

Similarly, in bull markets, managed futures strategies position themselves long as markets begin to rise and can profit if the rise continues.

The most recent downturn represents a classic example. Going into the fourth quarter of 2008, equity and energy prices had been declining, government bond prices had been rising, and currencies with high interest rates had been depreciating. This led to managed futures funds

being positioned short equities, short energies, long government bonds and gold, and short “carry” currencies. These hypothetical positions profited as the same trends continued throughout the quarter, while markets and other strategies suffered.

This hypothetical strategy trades 60 highly liquid futures and currency forwards during the period from January 1985 to December 2009.

Identifying Trends and Sizing of Positions

To determine the direction of the trend in each asset, the strategy considers the excess return over cash of each asset for the prior 12 months. The portfolio takes a long position if the return was positive and a short position if the return was negative. The strategy always holds positions in each of 24 commodity futures, 9 equity index futures, 15 bond futures and 12 currency forwards.

The size of each position is determined by volatility, with a target of 0.60% (10% for the overall portfolio) annualized volatility for each asset. This yields a portfolio that is equal risk weighted across the instruments to provide diversification and to limit the portfolio risk from any one asset. The portfolio is rebalanced at the end of each month.

9.4. Moving average model SP500

Figure 52: First test

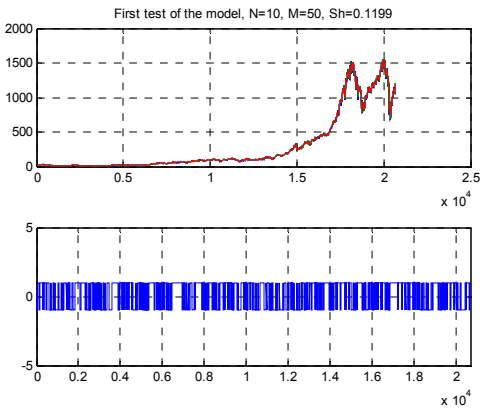


Figure 53: First test P&L

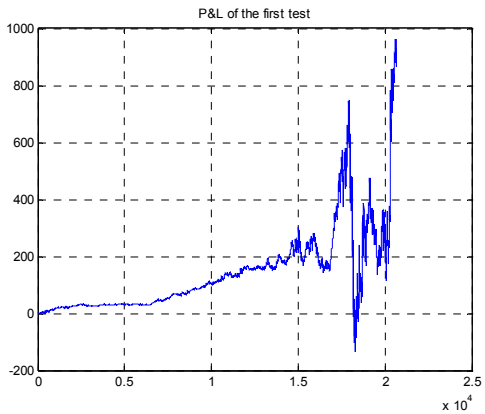


Figure 54: Sharpe ratios heatmap

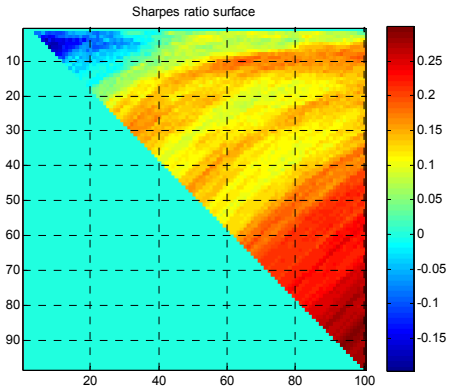


Figure 55: Sharpe ratios surface

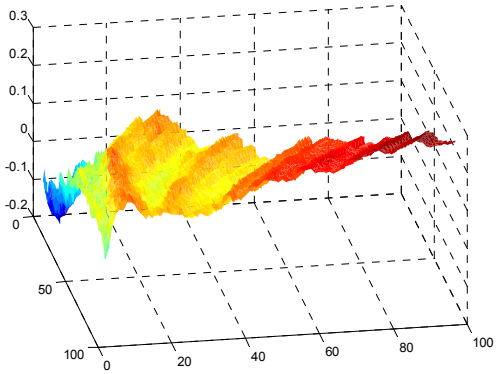
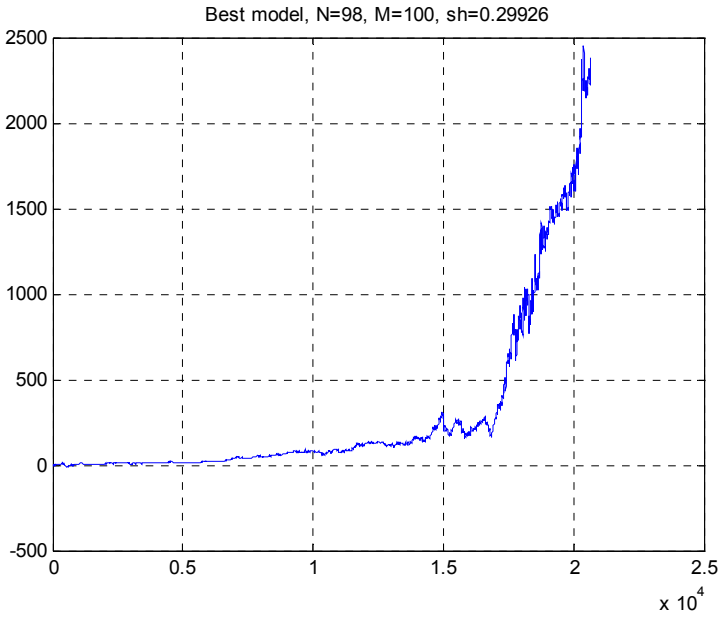
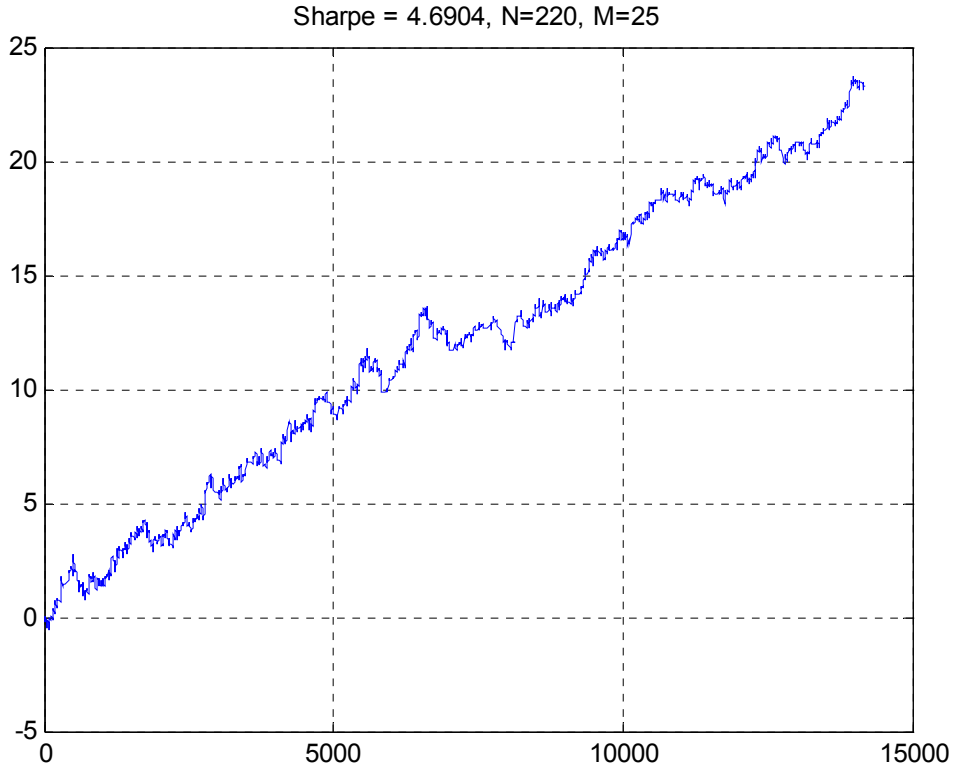


Figure 56: Best model



9.5. Moving Average + RSI bund 1 minute

Figure 57: Best Model Bund 1 minute



9.6. Statistical analysis

9.6.1 SP500-Descriptive statistics + ACF + PACF + GARCH modelling

Descriptive Statistics	
AVERAGE:	0,00%
STD DEV:	1,39%
SKEW:	0,09
EXCESS-KURTOSIS:	7,96
MEDIAN:	0,05%
MIN:	-9,03%
MAX:	11,58%
Q 1:	-0,64%
Q 3:	0,63%

Significance Test			5,00%
Target	**p-value**	**Different?**	
0,000	36,61%	FALSO	
0,000	0,96%	VERDADERO	
0,000	0,00%	VERDADERO	

Test	p-value	Result?
White-noise	0,00%	FALSO
Normal Distributed?	0,00%	FALSO
ARCH Effect?	0,00%	VERDADERO

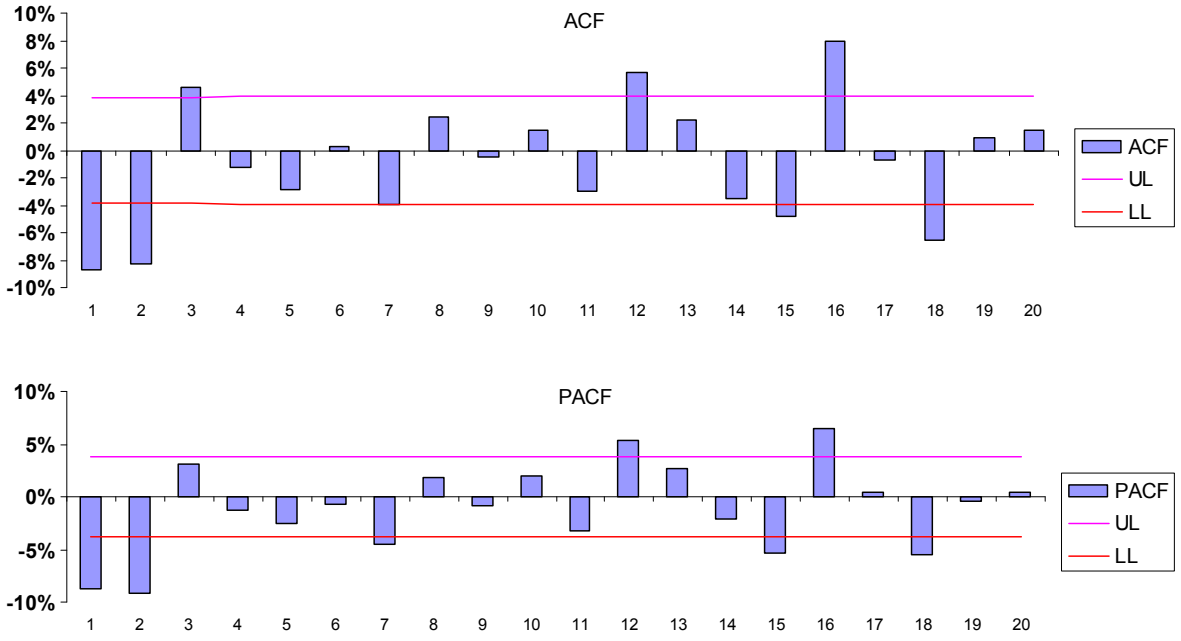
I use the adjusted closing prices of S&P 500 index between Jan 1st, 2000 and April 23th, 2010 (2591 observations)

The summary statistics describe a symmetric fat-tailed (leptokurtic) probability distribution for daily log returns. In other words, extreme movements are more probable than those predicted by normal-distribution model. As one may expect, the probability distribution is not normally distributed, and the logreturns exhibits significant serial correlation and ARCH effect.

Correlogram Analysis

Lag	ACF	UL	LL	PACF	UL	LL
1	-8,69%	3,85%	-3,85%	-8,69%	3,85%	-3,85%
2	-8,28%	3,85%	-3,85%	-9,11%	3,85%	-3,85%
3	4,61%	3,88%	-3,88%	3,08%	3,85%	-3,85%
4	-1,24%	3,91%	-3,91%	-1,29%	3,85%	-3,85%
5	-2,90%	3,91%	-3,91%	-2,49%	3,85%	-3,85%
6	0,23%	3,91%	-3,91%	-0,63%	3,85%	-3,85%
7	-3,97%	3,92%	-3,92%	-4,45%	3,85%	-3,85%
8	2,44%	3,92%	-3,92%	1,87%	3,85%	-3,85%
9	-0,49%	3,92%	-3,92%	-0,86%	3,85%	-3,85%
10	1,44%	3,93%	-3,93%	1,95%	3,85%	-3,85%
11	-3,03%	3,93%	-3,93%	-3,19%	3,85%	-3,85%
12	5,71%	3,93%	-3,93%	5,42%	3,85%	-3,85%
13	2,24%	3,93%	-3,93%	2,73%	3,85%	-3,85%
14	-3,50%	3,94%	-3,94%	-2,09%	3,85%	-3,85%
15	-4,84%	3,94%	-3,94%	-5,29%	3,85%	-3,85%
16	7,95%	3,95%	-3,95%	6,46%	3,85%	-3,85%
17	-0,72%	3,96%	-3,96%	0,49%	3,85%	-3,85%
18	-6,59%	3,98%	-3,98%	-5,48%	3,85%	-3,85%
19	0,90%	3,98%	-3,98%	-0,44%	3,85%	-3,85%
20	1,42%	4,00%	-4,00%	0,37%	3,85%	-3,85%

The white-noise test above indicates the presence of significant serial correlations in the time series. To drill more into this phenomenon, we computed and plotted the autocorrelation (ACF) and partial-autocorrelation (PACF) functions.



The ACF and PACF diagrams show serial correlations for lag-order of two. An ARMA model of second order can capture the serial correlation but falls short from producing heavy tails (i.e. excess-kurtosis) seen in the data.

We are ready to examine the different models, compare, and select the one that better-fits the data. By considering GARCH(1,1) model, and then move on to GARCH-M(1,1) and EGARCH(1,1) models, we evaluate the model assuming normal and non-normal distributed innovations. Finally, we summarize those models properties and recommend the one that fits the data best. LLF value (LLF is the log likelihood function):

Table 19: Goodness of fit - LLF different models

	Innovations		
	Normal	GED	t-Dist
GARCH(1,1)	7294	7321	7320
GARCH-M(1,1)	7294	7321	7320
EGARCH(1,1)	7348	**7365**	
ARMA(1,1)	6695		

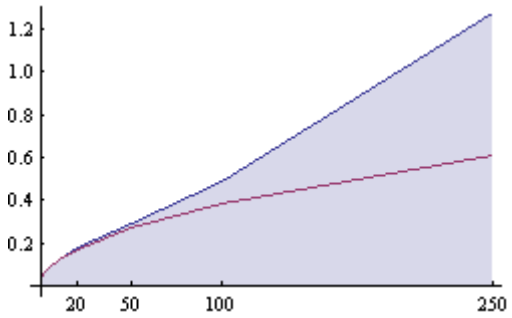
The EGARCH(1,1) with GED innovations seems like a reasonable model for the S&P 500 daily-log returns; it has the highest log-likelihood value and the model assumptions are largely satistifed. Nevertheless, the daily log-returns exhibits serial correlation that EGARCH does not capture.

Examining the in-sample model fitted conditional volatilities displays patterns similar to EWMA, but lower in value and faster to detect and to reflect changes in the underlying volatility. The forecast characteristic of EGARCH model emphasizes the mean-reversion phenomenon of the volatility.

9.6.2 Single stock-Microsoft: Variance ratio test

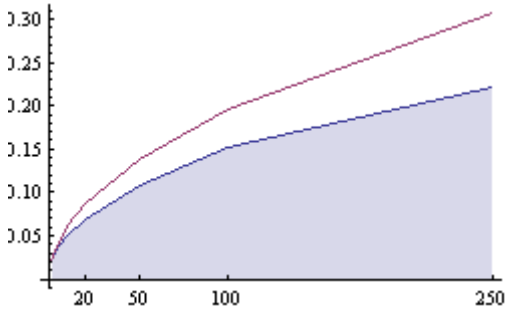
Figure 58: Microsoft 2002-5/2010

Figure 59: Variance ratio test



1	1.
2	0.999524
5	1.00586
10	1.02746
20	1.02531
50	1.06331
100	1.25116
250	2.07565

Figure 60: Variace ratio test: Longer time scale



1	1.
2	0.941002
5	0.873182
10	0.821617
20	0.79655
50	0.784726
100	0.775326
250	0.717192

On short time scales, single stocks may approximate a lognormal random walk - here the expected dispersion for periods up to 50 days are close to those one would infer from extending the daily moves with uncorrelated random daily changes (purple reference line).

But on longer scales, they can show significant trending behavior (serial correlation, total movement greater than expected for uncorrelated randomness, Figure 59) or mean reverting behavior (negative serial correlation, total movement less than expected for uncorrelated randomness, Figure 60).

9.6.3 Time series analysis SP500

Figure 61: SP500 daily

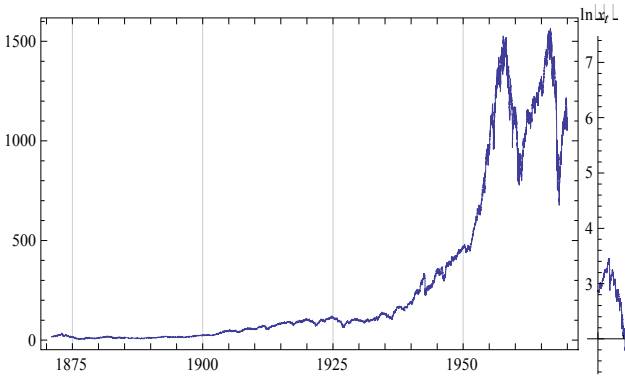
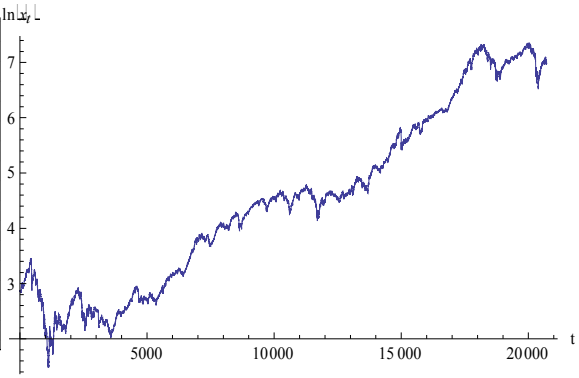


Figure 62: Log transformation



We see that the series rises sharply with time, clearly signaling the nonstationary nature of the series, and indicating theneed to transform the data to make it stationary. Here is the time plot of the series after the logarithmic transformation.

The variability in the series has been reduced. However, a rising trend still persists. We difference the series to obtain a constant mean series.

We saw in the previous section that both the correlation and partial correlation functions decay quickly to small values. This indicates that the series can be fitted to models of relatively low order. We use the Hannan-Rissanen procedure to select models and get preliminary estimates.

Table 20: HannanRissanenEstimate[data, 10, 6, 6, 5]

Models	Parameters + Vol				AIC
ARModel	0.0197641	-0.0342961	0.000142996		–8.8525
MAMode	0.0194216	-0.0336177	0.000143039		–8.85258
ARModel	0.0200571	-0.034452	0.00878715	0.000142988	–8.85246
ARMAModel	0.155957	-0.0368424	-0.136533	0.00014302	–8.85229
ARModel	0.0191037	0.000143158			–8.85147

These selected models with their estimated parameters are used as input to the function that performs the conditional maximum likelihood estimate. Model selected are AR(2),MAM(2),AR(3),ARMA(2,1),AR(1). Best models are then MA(2),AR(3).

10. **Machine Learning review**

Machine learning (ML) approach to financial modelling is an attempt to find financial models automatically – without many theoretical assumptions – through a process of progressive adaptation. Machine learning is deeply rooted in statistics and Artificial Intelligence (AI).

In the literature authors tend to use AI, ML and nonlinear models as synonims. As we are going to explore forecasting models we are going to review and analyze the most useful models to forecast financial markets through AI and ML techniques.

Learning in Machine learning is seen as a process of progressive adaptation that can be described in purely mechanical terms. In a general sense, learning is considered to be the ability to produce the right patterns in response to a given set of inputs. In principle, this notion of learning includes such high level mental activities as automatic language translation.

In machine learning, two fundamental forms of learning are distinguished: supervised and supervised learning.

Machine learning (and AI techniques in general) is one of the many techniques used in specific applications in financial modelling. In the 1990s, AI and its application to financial forecasting generated a lot of exaggerated claims and hype and received a lot of professional and media attention. AI was considered a revolutionary technology that could completely change the way people were doing business in finance. Today, as noted by one of the pioneers in the application of AI to finance, David Leinweber, those days are over. The hyperbole has made way for a more pragmatic attitude: AI is useful, but its merits and limits are now more clearly understood.[112].

10.1. Supervised, unsupervised and statistical learning

In machine learning, supervised learning is learning from examples. From a numer of approximately correct examples of pairs of output patterns and input stimuli, the machine learns how to respond to future stimuli.

In finance , the data-generation process (DGP) of a time series of prices is a function that links future returns with present and past returns. If only two lags are used the DGP that we want to learn is the function, if it exists, that links prices at time t+1 with prices at time t and t-1. The sample in this case will be all the triples formed by prices in 3 consecutive instants for all the sample set. The sample is formed by approximate realizations of the DGP.

Supervised learning thus means approximation of a function for which a number of samples are known.

In unsupervised learning a system discovers the the structure of data through a process of endogenous evolution. Clustering is the typical example of unsupervised learning. One starts with a set of points and discovers the grouping of points into clusters. In principle, the method of unsupervised learning applies to all available data. One can apply unsupervised learning to a sample, however, and then generalize to the entire population.

Supervised learning as defined before is the process of approximating a function from examples. The key ingredients in this concept are a mathematical model able to approximate any function and a set of rules for the approximation process from the examples.

The concept of supervised learning does not, in itself, include any notion of prediction or generalization. The main objective of a considerable body of research has been to show that specific learning rules can effectively learn any pattern. Learning patterns of small samples with high precision, however produces poor generalization and forecasting results for the population from which the sample came.

The approach that places the learning process in the context of generalization and forecasting is statistical learning. Given an efficient process of learning, statistical learning deals with how one can make sure the learning process has good generalization and forecasting abilities.

Statistical learning attempts to answer this question in a probabilistic framework. Classical learning theory places limits on the model complexity in order to improve the forecasting capabilities of the learned models. It does so by adding a penalty function that constraints model complexity. Vapnik (1995,1998)[113] introduced an important conceptual novelty by showing that not only does the complexity of the model matter but also does the type of functions used. He was able to construct mathematical framework to predict how well a given set of approximating functions would generalize.

10.2. Artificial Neural Networks

An artificial neural network (ANN), usually called "neural network" (NN), is a mathematical model or computational model that tries to simulate the structure and/or functional aspects of biological neural networks.[114] It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modelling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.[115]

A neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain.

Figure 63: A simple neural network

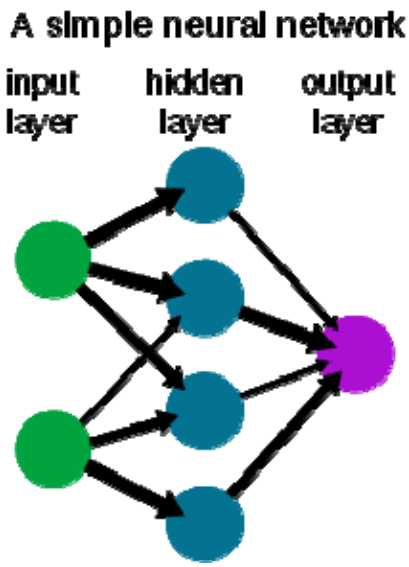
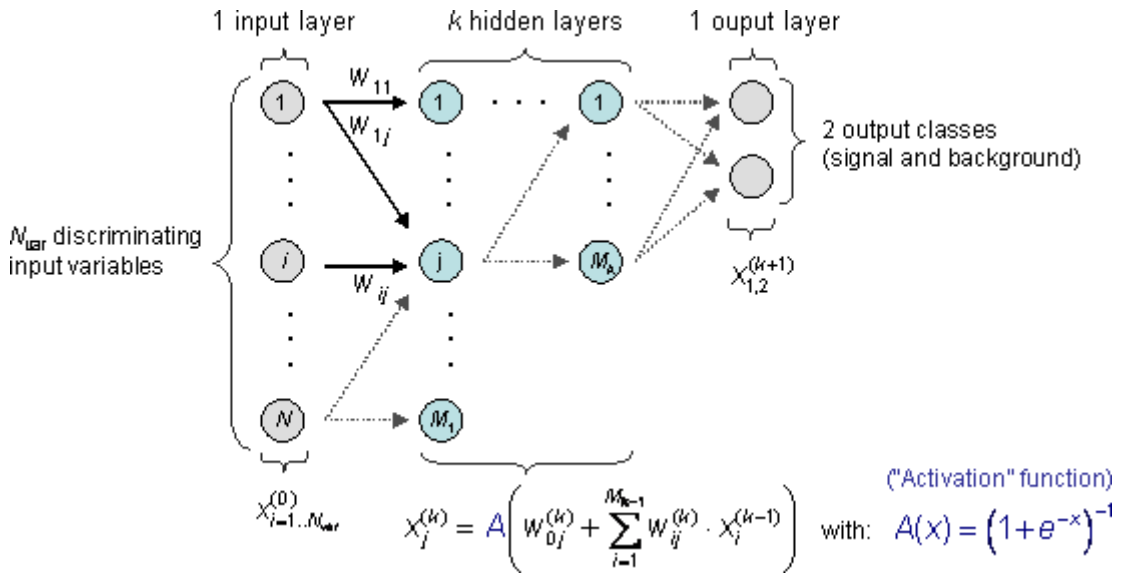


Figure 64: Another neural network



10.2.1 Background

There is no precise agreed-upon definition among researchers as to what a neural network is, but most would agree that it involves a network of simple processing elements (neurons), which can exhibit complex global behavior, determined by the connections between the processing elements and element parameters. The original inspiration for the technique came from examination of the central nervous system and the neurons (and their axons, dendrites and synapses) which constitute one of its most significant information processing elements (see neuroscience). In a neural network model, simple nodes, called variously "neurons", "neurodes", "PEs" ("processing elements") or "units", are connected together to form a network of nodes — hence the term "neural network". While a neural network does not have to be adaptive per se, its practical use comes with algorithms designed to alter the strength (weights) of the connections in the network to produce a desired signal flow.

These networks are also similar to the biological neural networks in the sense that functions are performed collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which various units are assigned (see also connectionism). Currently, the term Artificial Neural Network (ANN) tends to refer mostly to neural network models employed in statistics, cognitive psychology and artificial intelligence. Neural network models designed with emulation of the central nervous system (CNS) in mind are a subject of theoretical neuroscience (computational neuroscience).

In modern software implementations of artificial neural networks the approach inspired by biology has for the most part been abandoned for a more practical approach based on statistics and signal processing. In some of these systems, neural networks or parts of neural networks (such as artificial neurons) are used as components in larger systems that combine both adaptive and non-adaptive elements. While the more general approach of such adaptive systems is more suitable for real-world problem solving, it has far less to do with the traditional artificial intelligence connectionist models. What they do have in common, however, is the principle of non-linear, distributed, parallel and local processing and adaptation.

10.2.2 Models

Neural network models in artificial intelligence are usually referred to as artificial neural networks (ANNs); these are essentially simple mathematical models defining a function .

$$\boldsymbol{f} : \boldsymbol{X} \rightarrow \boldsymbol{Y} \quad (217)$$

Each type of ANN model corresponds to a class of such functions.

The word network in the term 'artificial neural network' arises because the function f(x) is defined as a composition of other functions gi(x), which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the

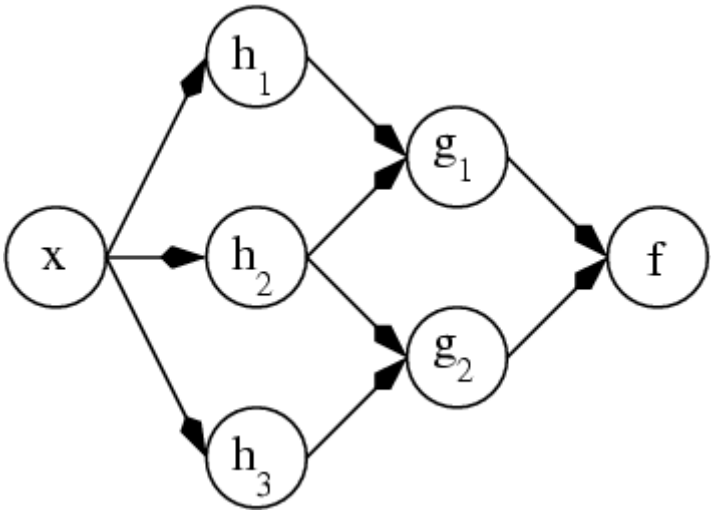
dependencies between variables. A widely used type of composition is the nonlinear weighted sum, where

$$f(x) = K\left(\sum_i w_i g_i(x)\right) \quad (218)$$

where K (commonly referred to as the activation function) is some predefined function, such as the hyperbolic tangent. It will be convenient for the following to refer to a collection of functions gi as simply a vector

$$g = (g_1, g_2, ..., g_n) \quad (219)$$

Figure 65: ANN dependency graph



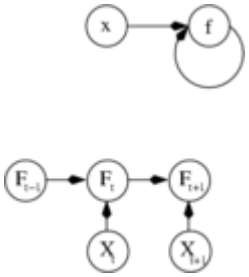
This figure depicts such a decomposition of f, with dependencies between variables indicated by arrows. These can be interpreted in two ways.

The first view is the functional view: the input x is transformed into a 3-dimensional vector h, which is then transformed into a 2-dimensional vector g, which is finally transformed into f. This view is most commonly encountered in the context of optimization.

The second view is the probabilistic view: the random variable F = f(G) depends upon the random variable G = g(H), which depends upon H = h(X), which depends upon the random variable X. This view is most commonly encountered in the context of graphical models.

The two views are largely equivalent. In either case, for this particular network architecture, the components of individual layers are independent of each other (e.g., the components of g are independent of each other given their input h). This naturally enables a degree of parallelism in the implementation.

Figure 66: Recurrent ANN dependency graph



Networks such as the previous one are commonly called feedforward, because their graph is a directed acyclic graph. Networks with cycles are commonly called recurrent. Such networks are commonly depicted in the manner shown at the top of the figure, where f is shown as being dependent upon itself. However, there is an implied temporal dependence which is not shown.

Learning

What has attracted the most interest in neural networks is the possibility of learning. Given a specific task to solve, and a class of functions F, learning means using a set of observations to find $\boldsymbol{f}^* \in \boldsymbol{F}$ which solves the task in some optimal sense.

This entails defining a cost function $\boldsymbol{C} : \boldsymbol{F} \to \Re$ such that, for the optimal solution f * , (i.e., $\boldsymbol{C}(\boldsymbol{f}^*) \leq \boldsymbol{C}(\boldsymbol{f}) \forall \boldsymbol{f} \in \boldsymbol{F}$ (ie no solution has a cost less than the cost of the optimal solution).

The cost function C is an important concept in learning, as it is a measure of how far away a particular solution is from an optimal solution to the problem to be solved. Learning algorithms search through the solution space to find a function that has the smallest possible cost.

For applications where the solution is dependent on some data, the cost must necessarily be a function of the observations, otherwise we would not be modelling anything related to the data. It is frequently defined as a statistic to which only approximations can be made. As a simple example consider the problem of finding the model f which minimizes

$$\boldsymbol{C} = \boldsymbol{E}\left[(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y})^2\right] \quad (220)$$

, for data pairs (x,y) drawn from some distribution . In practical situations we would only have N samples from D and thus, for the above example, we would only minimize

$$\hat{\boldsymbol{C}} = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \left(\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{y}_i\right)^2 \quad (221)$$

Thus, the cost is minimized over a sample of the data rather than the entire data set.

When $\boldsymbol{N} \to \infty$ some form of online machine learning must be used, where the cost is partially minimized as each new example is seen. While online machine learning is often used when is fixed, it is most useful in the case where the distribution changes slowly over time. In neural network methods, some form of online machine learning is frequently used for finite datasets.

10.2.3 Choosing a cost function

While it is possible to define some arbitrary, ad hoc cost function, frequently a particular cost will be used, either because it has desirable properties (such as convexity) or because it arises

naturally from a particular formulation of the problem (e.g., in a probabilistic formulation the posterior probability of the model can be used as an inverse cost). Ultimately, the cost function will depend on the task we wish to perform. The three main categories of learning tasks are overviewed below.

10.2.4 Learning paradigms

There are three major learning paradigms, each corresponding to a particular abstract learning task. These are supervised learning, unsupervised learning and reinforcement learning. Usually any given type of network architecture can be employed in any of those tasks.

Supervised learning

In supervised learning, we are given a set of example pairs $(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x} \in \boldsymbol{X}, \boldsymbol{y} \in \boldsymbol{Y}$ and the aim is to find a function $\boldsymbol{f} : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ in the allowed class of functions that matches the examples. In other words, we wish to infer the mapping implied by the data; the cost function is related to the mismatch between our mapping and the data and it implicitly contains prior knowledge about the problem domain.

A commonly used cost is the mean-squared error which tries to minimize the average squared error between the network's output, f(x), and the target value y over all the example pairs. When one tries to minimize this cost using gradient descent for the class of neural networks called Multi-Layer Perceptrons, one obtains the common and well-known backpropagation algorithm for training neural networks.

Tasks that fall within the paradigm of supervised learning are pattern recognition (also known as classification) and regression (also known as function approximation). The supervised learning paradigm is also applicable to sequential data (e.g., for speech and gesture recognition). This can be thought of as learning with a "teacher," in the form of a function that provides continuous feedback on the quality of solutions obtained thus far.

Unsupervised learning

In unsupervised learning we are given some data x and the cost function to be minimized, that can be any function of the data x and the network's output, f.

The cost function is dependent on the task (what we are trying to model) and our a priori assumptions (the implicit properties of our model, its parameters and the observed variables).

As a trivial example, consider the model f(x) = a, where a is a constant and the cost

$$\boldsymbol{C} = \boldsymbol{E}\left[\left(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}\right)^2\right] \quad (222)$$

Minimizing this cost will give us a value of a, that is equal to the mean of the data. The cost function can be much more complicated. Its form depends on the application: for example, in compression it could be related to the mutual information between x and y, whereas in statistical modelling, it could be related to the posterior probability of the model given the data. (Note that in both of those examples those quantities would be maximized rather than minimized). Tasks that

fall within the paradigm of unsupervised learning are in general estimation problems; the applications include clustering, the estimation of statistical distributions, compression and filtering.

Reinforcement learning

In reinforcement learning, data x are usually not given, but generated by an agent's interactions with the environment. At each point in time t, the agent performs an action yt and the environment generates an observation xt and an instantaneous cost ct, according to some (usually unknown) dynamics. The aim is to discover a policy for selecting actions that minimizes some measure of a long-term cost; i.e., the expected cumulative cost. The environment's dynamics and the long-term cost for each policy are usually unknown, but can be estimated.

More formally, the environment is modeled as a Markov decision process (MDP) with states and actions with the following probability distributions: the instantaneous cost distribution P(ct | st), the observation distribution P(xt | st) and the transition P(st + 1 | st,at), while a policy is defined as conditional distribution over actions given the observations. Taken together, the two define a Markov chain (MC). The aim is to discover the policy that minimizes the cost; i.e., the MC for which the cost is minimal. ANNs are frequently used in reinforcement learning as part of the overall algorithm. Tasks that fall within the paradigm of reinforcement learning are control problems, games and other sequential decision making tasks.

10.2.5 Learning algorithms

Training a neural network model essentially means selecting one model from the set of allowed models (or, in a Bayesian framework, determining a distribution over the set of allowed models) that minimizes the cost criterion. There are numerous algorithms available for training neural network models; most of them can be viewed as a straightforward application of optimization theory and statistical estimation.

Most of the algorithms used in training artificial neural networks employ some form of gradient descent. This is done by simply taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction.

Evolutionary methods, simulated annealing, expectation-maximization and non-parametric methods are some commonly used methods for training neural networks. See also machine learning.

Temporal perceptual learning relies on finding temporal relationships in sensory signal streams. In an environment, statistically salient temporal correlations can be found by monitoring the arrival times of sensory signals. This is done by the perceptual network.

Employing artificial neural networks

Perhaps the greatest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism which 'learns' from observed data. However, using them is not so straightforward and a relatively good understanding of the underlying theory is essential.

Choice of model: This will depend on the data representation and the application. Overly complex models tend to lead to problems with learning.
Learning algorithm: There are numerous tradeoffs between learning algorithms. Almost any algorithm will work well with the correct hyper-parameters for training on a particular fixed dataset. However selecting and tuning an algorithm for training on unseen data requires a significant amount of experimentation.
Robustness: If the model, cost function and learning algorithm are selected appropriately the resulting ANN can be extremely robust.
With the correct implementation ANNs can be used naturally in online learning and large dataset applications. Their simple implementation and the existence of mostly local dependencies exhibited in the structure allows for fast, parallel implementations in hardware.

10.2.6 Applications

The utility of artificial neural network models lies in the fact that they can be used to infer a function from observations. This is particularly useful in applications where the complexity of the data or task makes the design of such a function by hand impractical.
Real life applications
The tasks to which artificial neural networks are applied tend to fall within the following broad categories:

- Function approximation, or regression analysis, including time series prediction, fitness approximation and modelling.
- Classification, including pattern and sequence recognition, novelty detection and sequential decision making.
- Data processing, including filtering, clustering, blind source separation and compression.
- Robotics, including directing manipulators, Computer numerical control.

Application areas include system identification and control (vehicle control, process control), quantum chemistry, game-playing and decision making (backgammon, chess, racing), pattern recognition (radar systems, face identification, object recognition and more), sequence recognition (gesture, speech, handwritten text recognition), medical diagnosis, financial applications (automated trading systems), data mining (or knowledge discovery in databases, "KDD"), visualization and e-mail spam filtering.

10.2.7 Types of neural networks

Feedforward neural network
The feedforward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the

input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network.

Radial basis function (RBF) network

Radial Basis Functions are powerful techniques for interpolation in multidimensional space. A RBF is a function which has built into a distance criterion with respect to a center. Radial basis functions have been applied in the area of neural networks where they may be used as a replacement for the sigmoidal hidden layer transfer characteristic in Multi-Layer Perceptrons. RBF networks have two layers of processing: In the first, input is mapped onto each RBF in the 'hidden' layer. The RBF chosen is usually a Gaussian. In regression problems the output layer is then a linear combination of hidden layer values representing mean predicted output. The interpretation of this output layer value is the same as a regression model in statistics. In classification problems the output layer is typically a sigmoid function of a linear combination of hidden layer values, representing a posterior probability. Performance in both cases is often improved by shrinkage techniques, known as ridge regression in classical statistics and known to correspond to a prior belief in small parameter values (and therefore smooth output functions) in a Bayesian framework.

RBF networks have the advantage of not suffering from local minima in the same way as Multi-Layer Perceptrons. This is because the only parameters that are adjusted in the learning process are the linear mapping from hidden layer to output layer. Linearity ensures that the error surface is quadratic and therefore has a single easily found minimum. In regression problems this can be found in one matrix operation. In classification problems the fixed non-linearity introduced by the sigmoid output function is most efficiently dealt with using iteratively re-weighted least squares.

RBF networks have the disadvantage of requiring good coverage of the input space by radial basis functions. RBF centres are determined with reference to the distribution of the input data, but without reference to the prediction task. As a result, representational resources may be wasted on areas of the input space that are irrelevant to the learning task. A common solution is to associate each data point with its own centre, although this can make the linear system to be solved in the final layer rather large, and requires shrinkage techniques to avoid overfitting.

Associating each input datum with an RBF leads naturally to kernel methods such as Support Vector Machines and Gaussian Processes (the RBF is the kernel function). All three approaches use a non-linear kernel function to project the input data into a space where the learning problem can be solved using a linear model. Like Gaussian Processes, and unlike SVMs, RBF networks are typically trained in a Maximum Likelihood framework by maximizing the probability (minimizing the error) of the data under the model. SVMs take a different approach to avoiding overfitting by maximizing instead a margin. RBF networks are outperformed in most classification applications by SVMs. In regression applications they can be competitive when the dimensionality of the input space is relatively small.

Kohonen self-organizing network

The self-organizing map (SOM) invented by Teuvo Kohonen[116] performs a form of unsupervised learning. A set of artificial neurons learn to map points in an input space to coordinates in an output space. The input space can have different dimensions and topology from the output space, and the SOM will attempt to preserve these.

Figure 67: Self organizing map. Source: Wolfram research

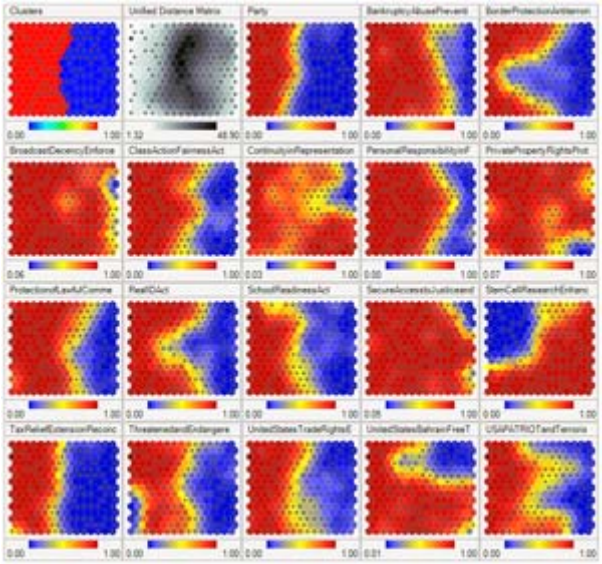


Figure 67 above shows a self-organizing map showing US Congress voting patterns visualized in Synapse. The first two boxes show clustering and distances while the remaining ones show the component planes. Red means a yes vote while blue means a no vote in the component planes (except the party component where red is Republican and blue is Democrat).

Recurrent network

Contrary to feedforward networks, recurrent neural networks (RNs) are models with bi-directional data flow. While a feedforward network propagates data linearly from input to output, RNs also propagate data from later processing stages to earlier stages.

Simple recurrent network

A simple recurrent network (SRN) is a variation on the Multi-Layer Perceptron, sometimes called an "Elman network" due to its invention by Jeff Elman[117]. A three-layer network is used, with the addition of a set of "context units" in the input layer. There are connections from the middle (hidden) layer to these context units fixed with a weight of one. At each time step, the input is propagated in a standard feed-forward fashion, and then a learning rule (usually back-propagation) is applied. The fixed back connections result in the context units always maintaining a copy of the previous values of the hidden units (since they propagate over the connections before the learning rule is applied). Thus the network can maintain a sort of state, allowing it to perform such tasks as sequence-prediction that are beyond the power of a standard Multi-Layer Perceptron. In a fully recurrent network, every neuron receives inputs from every other neuron in the network. These networks are not arranged in layers. Usually only a subset of the neurons

receive external inputs in addition to the inputs from all the other neurons, and another disjunct subset of neurons report their output externally as well as sending it to all the neurons. These distinctive inputs and outputs perform the function of the input and output layers of a feed-forward or simple recurrent network, and also join all the other neurons in the recurrent processing.

Hopfield network

The Hopfield network is a recurrent neural network in which all connections are symmetric. Invented by John Hopfield in 1982[118], this network guarantees that its dynamics will converge. If the connections are trained using Hebbian learning then the Hopfield network can perform as robust content-addressable (or associative) memory, resistant to connection alteration.

Echo state network

The echo state network (ESN) is a recurrent neural network with a sparsely connected random hidden layer. The weights of output neurons are the only part of the network that can change and be learned. ESN are good to (re)produce temporal patterns.

Long short term memory network

The Long short term memory is an artificial neural net structure that unlike traditional RNNs doesn't have the problem of vanishing gradients. It can therefore use long delays and can handle signals that have a mix of low and high frequency components.

Stochastic neural networks

A stochastic neural network differs from a typical neural network because it introduces random variations into the network. In a probabilistic view of neural networks, such random variations can be viewed as a form of statistical sampling, such as Monte Carlo sampling.

Boltzmann machine

The Boltzmann machine can be thought of as a noisy Hopfield network. Invented by Geoff Hinton and Terry Sejnowski in 1985[119], the Boltzmann machine is important because it is one of the first neural networks to demonstrate learning of latent variables (hidden units). Boltzmann machine learning was at first slow to simulate, but the contrastive divergence algorithm of Geoff Hinton (circa 2000) allows models such as Boltzmann machines and products of experts to be trained much faster.

Modular neural networks

Biological studies have shown that the human brain functions not as a single massive network, but as a collection of small networks. This realization gave birth to the concept of modular neural networks, in which several small networks cooperate or compete to solve problems.

Committee of machines

A committee of machines (CoM) is a collection of different neural networks that together "vote" on a given example. This generally gives a much better result compared to other neural network models. Because neural networks suffer from local minima, starting with the same architecture and training but using different initial random weights often gives vastly different networks. A CoM tends to stabilize the result.

The CoM is similar to the general machine learning bagging method, except that the necessary variety of machines in the committee is obtained by training from different random starting weights rather than training on different randomly selected subsets of the training data.

Associative neural network (ASNN)

The ASNN is an extension of the committee of machines that goes beyond a simple/weighted average of different models. ASNN represents a combination of an ensemble of feed-forward neural networks and the k-nearest neighbor technique (kNN). It uses the correlation between ensemble responses as a measure of distance amid the analyzed cases for the kNN. This corrects the bias of the neural network ensemble. An associative neural network has a memory that can coincide with the training set. If new data become available, the network instantly improves its predictive ability and provides data approximation (self-learn the data) without a need to retrain the ensemble. Another important feature of ASNN is the possibility to interpret neural network results by analysis of correlations between data cases in the space of models.

Physical neural network

A physical neural network includes electrically adjustable resistance material to simulate artificial synapses. Examples include the ADALINE neural network developed by Bernard Widrow in the 1960's and the memristor based neural network developed by Greg Snider of HP Labs in 2008.

Other types of networks

These special networks do not fit in any of the previous categories.

Holographic associative memory

Holographic associative memory represents a family of analog, correlation-based, associative, stimulus-response memories, where information is mapped onto the phase orientation of complex numbers operating.

Instantaneously trained networks

Instantaneously trained neural networks (ITNNs) were inspired by the phenomenon of short-term learning that seems to occur instantaneously. In these networks the weights of the hidden and the output layers are mapped directly from the training vector data. Ordinarily, they work on binary data, but versions for continuous data that require small additional processing are also available.

Spiking neural networks

Spiking neural networks (SNNs) Wulfram Gerstner (2001)[120] are models which explicitly take into account the timing of inputs. The network input and output are usually represented as series of spikes (delta function or more complex shapes). SNNs have an advantage of being able to process information in the time domain (signals that vary over time). They are often implemented as recurrent networks. SNNs are also a form of pulse computer. Spiking neural networks with axonal conduction delays exhibit polychronization, and hence could have a very large memory capacity.

Networks of spiking neurons — and the temporal correlations of neural assemblies in such networks — have been used to model figure/ground separation and region linking in the visual system (see, for example, Reitboeck et al.in Haken and Stadler: Synergetics of the Brain. Berlin,

1989). In June 2005 IBM announced construction of a Blue Gene supercomputer dedicated to the simulation of a large recurrent spiking neural network.

Dynamic neural networks

Dynamic neural networks not only deal with nonlinear multivariate behaviour, but also include (learning of) time-dependent behaviour such as various transient phenomena and delay effects.

Cascading neural networks

Cascade-Correlation is an architecture and a supervised learning algorithm developed by Scott Fahlman and Christian Lebiere[121]. Instead of just adjusting the weights in a network of fixed topology, Cascade-Correlation begins with a minimal network, then automatically trains and adds new hidden units one by one, creating a multi-layer structure. Once a new hidden unit has been added to the network, its input-side weights are frozen. This unit then becomes a permanent feature-detector in the network, available for producing outputs or for creating other, more complex feature detectors. The Cascade-Correlation architecture has several advantages over existing algorithms: it learns very quickly, the network determines its own size and topology, it retains the structures it has built even if the training set changes, and it requires no back-propagation of error signals through the connections of the network. See: Cascade correlation algorithm.

Neuro-fuzzy networks

A neuro-fuzzy network is a fuzzy inference system in the body of an artificial neural network. Depending on the FIS type, there are several layers that simulate the processes involved in a fuzzy inference like fuzzification, inference, aggregation and defuzzification. Embedding an FIS in a general structure of an ANN has the benefit of using available ANN training methods to find the parameters of a fuzzy system.

Compositional pattern-producing networks

Compositional pattern-producing networks (CPPNs) are a variation of ANNs which differ in their set of activation functions and how they are applied. While typical ANNs often contain only sigmoid functions (and sometimes Gaussian functions), CPPNs can include both types of functions and many others. Furthermore, unlike typical ANNs, CPPNs are applied across the entire space of possible inputs so that they can represent a complete image. Since they are compositions of functions, CPPNs in effect encode images at infinite resolution and can be sampled for a particular display at whatever resolution is optimal.

One-shot associative memory

This type of network can add new patterns without the need for re-training. It is done by creating a specific memory structure, which assigns each new pattern to an orthogonal plane using adjacently connected hierarchical arrays [5]. The network offers real-time pattern recognition and high scalability, it however requires parallel processing and is thus best suited for platforms such as Wireless sensor networks (WSN), Grid computing, and GPGPUs.

10.2.8 Theoretical properties

Computational power

The multi-layer perceptron (MLP) is a universal function approximator, as proven by the Cybenko theorem. However, the proof is not constructive regarding the number of neurons required or the settings of the weights.

Work by Hava Siegelmann and Eduardo D. Sontag has provided a proof that a specific recurrent architecture with rational valued weights (as opposed to the commonly used floating point approximations) has the full power of a Universal Turing Machine[6] using a finite number of neurons and standard linear connections. They have further shown that the use of irrational values for weights results in a machine with super-Turing power.

Capacity

Artificial neural network models have a property called 'capacity', which roughly corresponds to their ability to model any given function. It is related to the amount of information that can be stored in the network and to the notion of complexity.

Convergence

Nothing can be said in general about convergence since it depends on a number of factors. Firstly, there may exist many local minima. This depends on the cost function and the model. Secondly, the optimization method used might not be guaranteed to converge when far away from a local minimum. Thirdly, for a very large amount of data or parameters, some methods become impractical. In general, it has been found that theoretical guarantees regarding convergence are an unreliable guide to practical application.

Generalisation and statistics

In applications where the goal is to create a system that generalises well in unseen examples, the problem of overtraining has emerged. This arises in overcomplex or overspecified systems when the capacity of the network significantly exceeds the needed free parameters. There are two schools of thought for avoiding this problem: The first is to use cross-validation and similar techniques to check for the presence of overtraining and optimally select hyperparameters such as to minimize the generalisation error. The second is to use some form of regularisation. This is a concept that emerges naturally in a probabilistic (Bayesian) framework, where the regularisation can be performed by selecting a larger prior probability over simpler models; but also in statistical learning theory, where the goal is to minimize over two quantities: the 'empirical risk' and the 'structural risk', which roughly corresponds to the error over the training set and the predicted error in unseen data due to overfitting.

Confidence analysis of a neural network

Supervised neural networks that use an MSE cost function can use formal statistical methods to determine the confidence of the trained model. The MSE on a validation set can be used as an estimate for variance. This value can then be used to calculate the confidence interval of the output of the network, assuming a normal distribution. A confidence analysis made this way is

statistically valid as long as the output probability distribution stays the same and the network is not modified. By assigning a softmax activation function on the output layer of the neural network (or a softmax component in a component-based neural network) for categorical target variables, the outputs can be interpreted as posterior probabilities. This is very useful in classification as it gives a certainty measure on classifications. The softmax activation function is:

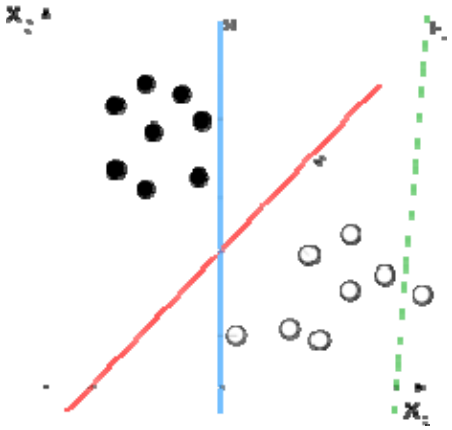
$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{c} e^{x_i}} \quad (223)$$

10.2.9 Support vector machines

Support vector machines (SVMs) [122] are a set of related supervised learning methods used for classification and regression. In simple words, given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training datapoints of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Figure 68: Support Vector Machines



Motivation

H3 (green) doesn't separate the 2 classes. H1 (blue) does, with a small margin and H2 (red) with the maximum margin.Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a p − 1-dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier.

Formalization

We are given some training data , a set of n points of the form

$$D = \{(x_i, c_i) | x_i \in R^p, c_i \in \{-1,1\}\}_{i=1}^n \quad (224)$$

where the ci is either 1 or −1, indicating the class to which the point x_i belongs. Each x_i is a p-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having ci = 1 from those having ci = − 1. Any hyperplane can be written as the set of points x satisfying

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \quad (225)$$

where . denotes the dot product. The vector $\mathbf{w}$ a normal vector: it is perpendicular to the hyperplane. The parameter $b/\|w\|$ determines the offset of the hyperplane from the origin along the normal vector w.

We want to choose the w and b to maximize the margin, or distance between the parallel hyperplanes that are as far apart as possible while still separating the data. These hyperplanes can be described by the equations

$$\mathbf{w} \cdot \mathbf{x} - \boldsymbol{b} = 1 \quad (226)$$

and

$$\mathbf{w} \cdot \mathbf{x} - b = -1 \quad (227)$$

Note that if the training data are linearly separable, we can select the two hyperplanes of the margin in a way that there are no points between them and then try to maximize their distance. By using geometry, we find the distance between these two hyperplanes is $2/\|w\|$, so we want to minimize $\|w\|$. As we also have to prevent data points falling into the margin, we add the following constraint: for each i either

$$\mathbf{w} \cdot \mathbf{x}_i - b \geq 1 \quad (228)$$

for xi of the first class or

$$\mathbf{w} \cdot \mathbf{x}_i - b \leq -1 \quad (229)$$

fo xi of the second. This can be rewritten as:

$$c_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \quad (230)$$

We can put this together to get the optimization problem:

Minimize (in w,b)

$$\|\mathbf{w}\| \quad (231)$$

subject to (for any i=1,…n)

$$c_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \quad (232)$$

Primal form

The optimization problem presented in the preceding section is difficult to solve because it depends on ||w||, the norm of w, which involves a square root. Fortunately it is possible to alter the equation by substituting ||w|| with $\frac{1}{2}\|\mathbf{w}\|^2$ (the factor of 1/2 being used for mathematical convenience) without changing the solution (the minimum of the original and the modified equation have the same w and b). This is a quadratic programming (QP) optimization problem. More clearly:

Minimize (in w,b)

$$\frac{1}{2}\|\mathbf{w}\|^2 \quad (233)$$

subject to (for any i=1…..n)

$$c_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \quad (234)$$

One could be tempted to express the previous problem by means of non-negative Lagrange multipliers αi as

$$c_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1 \geq 0 \quad (235)$$

but this would be wrong. The reason is the following: suppose we can find a family of hyperplanes which divide the points; then all $c_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1 \geq 0$. Hence we could find the minimum by sending all αi to +∞, and this minimum would be reached for all the members of the family, not only for the best one which can be chosen solving the original problem.

Nevertheless the previous constrained problem can be expressed as

$$\min_{w,b} \max_{\alpha} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 - \sum \alpha_i [c_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1] \right\} \quad (236)$$

that is we look for a saddle point. In doing so all the points which can be separated as $c_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1 > 0$ do not matter since we must set the corresponding αi to zero.

This problem can now be solved by standard quadratic programming techniques and programs. The solution can be expressed by terms of linear combination of the training vectors as only a few

αi will be greater than zero. The corresponding xi are exactly the support vectors, which lie on the margin and satisfy $c_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1$. From this one can derive that the support vectors also satisfy

$$\mathbf{w}\cdot\mathbf{x}_i - b = 1/c_i = c_i \Leftrightarrow \mathbf{w}\cdot\mathbf{x}_i - c_i \quad (237)$$

which allows one to define the offset b. In practice, it is more robust to average over all Nsv support vectors:

$$b = \frac{1}{N_{SV}}\sum_{i=1}^{N_{SV}}(\mathbf{w}\cdot\mathbf{x}_i - c_i) \quad (238)$$

Dual form

Writing the classification rule in its unconstrained dual form reveals that the maximum margin hyperplane and therefore the classification task is only a function of the support vectors, the training data that lie on the margin.

Using the fact, that $\|\mathbf{w}\|^2 = w\cdot w$ and substituting $\mathbf{w} = \sum_{i=0}^{n}\alpha_i c_i \mathbf{x_i}$, one can show that the dual of the SVM boils down to the following optimization problem:

Maximize (in αi)

$$\widetilde{L}(\alpha) = \sum_{i=0}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i c_i \alpha_j c_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=0}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i c_i \alpha_j c_j k(\mathbf{x}_i \mathbf{x}_j) \quad (239)$$

subject to (for any)

$$\alpha_i \geq 0 \quad (240)$$

and to the constraint from the minimization in b

$$\sum_{i=0}^{n}\alpha_i c_i = 0 \quad (241)$$

Here the kernel is defined by .

$$k(\mathbf{x}_i \mathbf{x}_j) = \mathbf{x_i}\cdot\mathbf{x_j} \quad (242)$$

The α terms constitute a dual representation for the weight vector in terms of the training set:

$$\mathbf{w} = \sum_{i=0}^{n}\alpha_i c_i \mathbf{x_i} \quad (243)$$

Biased and unbiased hyperplanes

For simplicity reasons, sometimes it is required that the hyperplane passes through the origin of the coordinate system. Such hyperplanes are called unbiased, whereas general hyperplanes not necessarily passing through the origin are called biased. An unbiased hyperplane can be enforced by setting b = 0 in the primal optimization problem. The corresponding dual is identical to the dual given above without the equality constraint

$$\sum_{i=0}^{n}\alpha_i c_i = 0 \quad (244)$$

Transductive support vector machines

Transductive support vector machines extend SVMs in that they could also treat partially labeled data in semi-supervised learning. Here, in addition to the training set , the learner is also given a set

$$D^* = \left\{ x_i^* \, x_i^* \in R^P \right\}_{i=1}^k \text{(245)}$$

of test examples to be classified. Formally, a transductive support vector machine is defined by the following primal optimization problem:

Minimize (in w,b,c*)

$$\frac{1}{2}\|\mathbf{w}\|^2 \text{(246)}$$

subject to (for any and any i=1,…,n and any j=1,…,k)

$$c_i\left(\mathbf{w}\cdot\mathbf{x}_i - b\right) \geq 1$$

$$c_j^*\left(\mathbf{w}\cdot\mathbf{x}_j^* - b\right) \geq 1 \text{ (247)}$$

and

$$c_j^* \in \{-1,1\} \text{ (248)}$$

Transductive support vector machines were introduced by Vladimir Vapnik in 1998.

Properties

SVMs belong to a family of generalized linear classifiers. They can also be considered a special case of Tikhonov regularization. A special property is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum margin classifiers.

A comparison of the SVM to other classifiers has been made by Meyer, Leisch and Hornik.[123]

Extensions to the linear SVM

Soft margin

In 1995, Corinna Cortes and Vladimir Vapnik[124] suggested a modified maximum margin idea that allows for mislabeled examples. If there exists no hyperplane that can split the "yes" and "no" examples, the Soft Margin method will choose a hyperplane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. The method introduces slack variables, ξi, which measure the degree of misclassification of the datum xi

$$\boldsymbol{c}_i\left(\mathbf{w}\cdot\mathbf{x}_i - \boldsymbol{b}\right) \geq 1 - \varepsilon_i \quad 1 \leq \mathrm{i} \leq \mathrm{n} \text{ (249)}$$

The objective function is then increased by a function which penalizes non-zero ξi, and the optimization becomes a trade off between a large margin, and a small error penalty. If the penalty function is linear, the optimization problem becomes:

$$\min_{w,\varepsilon} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\varepsilon_i \text{ (250)}$$

subject to (for any i=1,..,n)

$$c_i(\mathbf{w}\cdot\mathbf{x}_i - b) \geq 1-\varepsilon_i \quad \varepsilon_i \geq 0 \text{ (251)}$$

This constraint in (250) along with the objective of minimizing $\|\mathbf{w}\|$ can be solved using Lagrange multipliers as done above. One has then to solve the following problem

$$\min_{w,\varepsilon,b}\max_{\alpha,\beta}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\alpha_i[c_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 + \varepsilon_i] - \sum_{i=1}^{n}\varepsilon_i\beta_i\right\} \text{ (252)}$$

with $\alpha_i, \beta_i \geq 0$ (253).

The key advantage of a linear penalty function is that the slack variables vanish from the dual problem, with the constant C appearing only as an additional constraint on the Lagrange multipliers. For the above formulation and its huge impact in practice, Cortes and Vapnik received the 2008 ACM Paris Kanellakis Award[125]. Non-linear penalty functions have been used, particularly to reduce the effect of outliers on the classifier, but unless care is taken, the problem becomes non-convex, and thus it is considerably more difficult to find a global solution.

Non-linear classification

The original optimal hyperplane algorithm proposed by Vladimir Vapnik in 1963 was a linear classifier. However, in 1992, Bernhard Boser, Isabelle Guyon and Vapnik suggested a way to create non-linear classifiers by applying the kernel trick (originally proposed by Aizerman et al.[126]) to maximum-margin hyperplanes. The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be non-linear and the transformed space high dimensional; thus though the classifier is a hyperplane in the high-dimensional feature space, it may be non-linear in the original input space.

If the kernel used is a Gaussian radial basis function, the corresponding feature space is a Hilbert space of infinite dimension. Maximum margin classifiers are well regularized, so the infinite dimension does not spoil the results. Some common kernels include,

Polynomial (homogeneous): $k(\mathbf{x}_i\,\mathbf{x}_j) = (\mathbf{x_i}\cdot\mathbf{x}_j)^d$ (254)

Polynomial (inhomogeneous): $k(\mathbf{x}_i\,\mathbf{x}_j) = (\mathbf{x_i}\cdot\mathbf{x}_j + 1)^d$ (255)

Radial Basis Function: , $k(\mathbf{x}_i\,\mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)^d$ (256)for γ > 0

Gaussian Radial basis function: $k(\mathbf{x}_i\,\mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)^d$ (257)

Hyperbolic tangent: $k(\mathbf{x}_i\,\mathbf{x}_j) = \tanh(\kappa\mathbf{x_i}\cdot\mathbf{x}_j + c)$(258), for some (not every) κ > 0 and c < 0

The kernel is related to the transform $\varphi(x_i)$ by the equation $k(\mathbf{x}_i\,\mathbf{x}_j) = \varphi(x_i)\cdot\varphi(x_j)$. The value w is also in the transformed space, with $w = \sum_i \alpha_i c_i \varphi(x_i)$. Dot products with w for classification can

again be computed by the kernel trick, i.e. $w \cdot \varphi(x_i) = \sum_i \alpha_i c_i \kappa(x_i, x)$. However, there does not in general exist a value w' such that $w \cdot \varphi(x_i) = \kappa(w', x)$.

Issues

Potential drawbacks of the SVM are the following two aspects:

- Uncalibrated Class membership probabilities
- The SVM is only directly applicable for two-class tasks. Therefore, algorithms that reduce the multi-class task to several binary problems have to be applied, see the Multi-class SVM section.

Multiclass SVM

Multiclass SVM aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements. The dominating approach for doing so is to reduce the single multiclass problem into multiple binary classification problems. Each of the problems yields a binary classifier, which is assumed to produce an output function that gives relatively large values for examples from the positive class and relatively small values for examples belonging to the negative class. Two common methods to build such binary classifiers are where each classifier distinguishes between (i) one of the labels to the rest (one-versus-all) or (ii) between every pair of classes (one-versus-one). Classification of new instances for one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class (it is important that the output functions be calibrated to produce comparable scores). For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with most votes determines the instance classification.

Structured SVM

SVMs have been generalized to Structured SVM, where the label space is structured and of possibly infinite size.

Regression

A version of SVM for regression was proposed in 1996 by Vladimir Vapnik, Harris Drucker, Chris Burges, Linda Kaufman and Alex Smola[127] This method is called support vector regression (SVR). The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction (within a threshold ε).

Implementation

The parameters of the maximum-margin hyperplane are derived by solving the optimization. There exist several specialized algorithms for quickly solving the QP problem that arises from SVMs, mostly reliant on heuristics for breaking the problem down into smaller, more-manageable

chunks. A common method for solving the QP problem is the Platt's Sequential Minimal Optimization (SMO) algorithm, which breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating the need for a numerical optimization algorithm.
Another approach is to use an interior point method that uses Newton-like iterations to find a solution of the Karush-Kuhn-Tucker conditions of the primal and dual problems.[128] Instead of solving a sequence of broken down problems, this approach directly solves the problem as a whole. To avoid solving a linear system involving the large kernel matrix, a low rank approximation to the matrix is often used to use the kernel trick.

10.3. Classification and regression trees

Trees are ubiquitous in Artificial Intelligence because they implement a natural way of searching variables. In AI, trees are called “identification trees”. Ross Quinlan (1979)[129] built a well known family of identification trees work by applying a sequence of binary classification tests to a set of examples. In this way, the method constructs a tree that exactly classifies each example.
Trees can be constructed as hierarchical sets of rules, for example, ID3, the best known system of the ID family of trees developed by Quinlan.
The generalization of a tree depends on when the tree is stopped, thus creating an approximate classification. As in every application of automatic learning, understanding where to stop the accuracy of in-sample analysis is one of the critical tasks.
Classification and regression trees are the statistical counterpart of Quinlan’s identification trees. CART work by splitting variables into two or more segments, such as return $\leq$ 3%, return > 3%. The objective is to identify what combination of values of the independent variables corresponds to a given value of the dependent variable. Each item might be identified by both continuous and categorical variables (i.e. discrete variables that identify categories). For example, a group of companies could be identified by both continuous variables (e.g. financial ratios) and by categorical variables (e.g. Industrial sector); by successive splitting, one can identify what financial ratios and sector identifier correspond to a given credit rating.
Standard regressions work only on with continuous variables, but CART accepts as inputs a combination of different types of variables including discrete variables.

10.4. Genetic Algorithms

A genetic algorithm (GA) is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms (also known as evolutionary computation) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination).

Genetic algorithms in particular became popular through the work of John Holland in the early 1970s, and particularly his book Adaptation in Natural and Artificial Systems (1975[130]). His work originated with studies of cellular automata, conducted by Holland and his students at the University of Michigan. Holland introduced a formalized framework for predicting the quality of the next generation, known as Holland's Schema Theorem. Research in GAs remained largely theoretical until the mid-1980s, when The First International Conference on Genetic Algorithms was held in Pittsburgh, Pennsylvania.

Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

Genetic algorithms find application in bioinformatics, phylogenetics, computer science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields.

A typical genetic algorithm requires two things to be defined:

- a genetic representation of the solution domain,
- a fitness function to evaluate the solution domain.

A standard representation of the solution is as an array of bits. Arrays of other types and structures can be used in essentially the same way. The main property that makes these genetic representations convenient is that their parts are easily aligned due to their fixed size, that facilitates simple crossover operation. Variable length representations may also be used, but crossover implementation is more complex in this case. Tree-like representations are explored in Genetic programming and graph-form representations are explored in Evolutionary programming.

The fitness function is defined over the genetic representation and measures the quality of the represented solution. The fitness function is always problem dependent. For instance, in the knapsack problem we want to maximize the total value of objects that we can put in a knapsack of some fixed capacity. A representation of a solution might be an array of bits, where each bit represents a different object, and the value of the bit (0 or 1) represents whether or not the object is in the knapsack. Not every such representation is valid, as the size of objects may exceed the capacity of the knapsack. The fitness of the solution is the sum of values of all objects in the knapsack if the representation is valid, or 0 otherwise. In some problems, it is hard or even

impossible to define the fitness expression; in these cases, interactive genetic algorithms are used.

Once we have the genetic representation and the fitness function defined, GA proceeds to initialize a population of solutions randomly, then improving it through repetitive application of mutation, crossover, inversion and selection operators.

1 - Initial random population

Candidate	String	Fitness
C	00000110	2
B	11101110	6
C	00100000	1
D	00110100	3

2 - Crossover Applied

Initial Parent	Candidate B	Candidate C
	11101110	00100000
Resulting Child	Candidate E	Candidate F
	01101110	10100000

3 - No Crossover Applied

Initial Parent	Candidate B	Candidate D
	11101110	00110100
Resulting Child	Candidate G	Candidate H
	11101110	00110100

4 - Final new generation of solutions

Candidate	String	Fitness
E (Mutation)	01001110	4
F	10100000	2
G	11101110	6
H	00110100	3

10.4.1 Initialization

Initially many individual solutions are randomly generated to form an initial population. The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Traditionally, the population is generated randomly, covering the entire range of possible solutions (the search space). Occasionally, the solutions may be "seeded" in areas where optimal solutions are likely to be found.

10.4.2 Selection

During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. Certain selection methods rate the fitness of each solution and preferentially select the best solutions. Other methods rate only a random sample of the population, as this process may be very time-consuming. Most functions are stochastic and designed so that a small proportion of less fit solutions are selected. This helps keep the diversity of the population large, preventing premature convergence on poor solutions. Popular and well-studied selection methods include roulette wheel selection and tournament selection.

10.4.3 Reproduction

The next step is to generate a second generation population of solutions from those selected through genetic operators: crossover (also called recombination), and/or mutation.
For each new solution to be produced, a pair of "parent" solutions is selected for breeding from the pool selected previously. By producing a "child" solution using the above methods of crossover and mutation, a new solution is created which typically shares many of the characteristics of its "parents". New parents are selected for each child, and the process continues until a new population of solutions of appropriate size is generated.
These processes ultimately result in the next generation population of chromosomes that is different from the initial generation. Generally the average fitness will have increased by this procedure for the population, since only the best organisms from the first generation are selected for breeding, along with a small proportion of less fit solutions, for reasons already mentioned above.

10.4.4 Termination

This generational process is repeated until a termination condition has been reached. A solution is found that satisfies minimum criteria

- Fixed number of generations reached
- Allocated budget (computation time/money) reached
- The highest ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better results
- Manual inspection
- Combinations of the above.
-

10.4.5 Solutions

There are several general observations about the generation of solutions via a genetic algorithm:
In many problems, GAs may have a tendency to converge towards local optima or even arbitrary points rather than the global optimum of the problem. This means that it does not "know how" to sacrifice short-term fitness to gain longer-term fitness. The likelihood of this occurring depends on the shape of the fitness landscape: certain problems may provide an easy ascent towards a global optimum, others may make it easier for the function to find the local optima. This problem may be alleviated by using a different fitness function, increasing the rate of mutation, or by using selection techniques that maintain a diverse population of solutions, although the No Free Lunch theorem proves that there is no general solution to this problem. A common technique to maintain diversity is to impose a "niche penalty", wherein, any group of individuals of sufficient similarity (niche radius) have a penalty added, which will reduce the representation of that group in subsequent generations, permitting other (less similar) individuals to be maintained in the population. This trick, however, may not be effective, depending on the landscape of the problem.

Diversity is important in genetic algorithms (and genetic programming) because crossing over a homogeneous population does not yield new solutions. In evolution strategies and evolutionary programming, diversity is not essential because of a greater reliance on mutation.

Operating on dynamic data sets is difficult, as genomes begin to converge early on towards solutions which may no longer be valid for later data. Several methods have been proposed to remedy this by increasing genetic diversity somehow and preventing early convergence, either by increasing the probability of mutation when the solution quality drops (called triggered hypermutation), or by occasionally introducing entirely new, randomly generated elements into the gene pool (called random immigrants). Recent research has also shown the benefits of using biological exaptation (or preadaptation) in solving this problem. Again, evolution strategies and evolutionary programming can be implemented with a so-called "comma strategy" in which parents are not maintained and new parents are selected only from offspring. This can be more effective on dynamic problems.

GAs cannot effectively solve problems in which the only fitness measure is right/wrong, as there is no way to converge on the solution. (No hill to climb.) In these cases, a random search may find a solution as quickly as a GA.

Selection is clearly an important genetic operator, but opinion is divided over the importance of crossover versus mutation. Some argue that crossover is the most important, while mutation is only necessary to ensure that potential solutions are not lost. Others argue that crossover in a largely uniform population only serves to propagate innovations originally found by mutation, and in a non-uniform population crossover is nearly always equivalent to a very large mutation (which is likely to be catastrophic). There are many references in Fogel (2006) that support the importance of mutation-based search, but across all problems the No Free Lunch theorem holds, so these opinions are without merit unless the discussion is restricted to a particular problem.

Often, GAs can rapidly locate good solutions, even for difficult search spaces. The same is of course also true for evolution strategies and evolutionary programming.

For specific optimization problems and problem instantiations, simpler optimization algorithms may find better solutions than genetic algorithms (given the same amount of computation time). Alternative and complementary algorithms include evolution strategies, evolutionary programming, simulated annealing, Gaussian adaptation, hill climbing, and swarm intelligence (e.g.: ant colony optimization, particle swarm optimization).

As with all current machine learning problems it is worth tuning the parameters such as mutation probability, recombination probability and population size to find reasonable settings for the problem class being worked on. A very small mutation rate may lead to genetic drift (which is non-ergodic in nature). A recombination rate that is too high may lead to premature convergence of the genetic algorithm. A mutation rate that is too high may lead to loss of good solutions unless there is elitist selection. There are theoretical but not yet practical upper and lower bounds for these parameters that can help guide selection. The implementation and evaluation of the fitness function is an important factor in the speed and efficiency of the algorithm.

10.4.6 Variants

The simplest algorithm represents each chromosome as a bit string. Typically, numeric parameters can be represented by integers, though it is possible to use floating point representations. The floating point representation is natural to evolution strategies and evolutionary programming. The notion of real-valued genetic algorithms has been offered but is really a misnomer because it does not really represent the building block theory that was proposed by Holland in the 1970s. This theory is not without support though, based on theoretical and experimental results. The basic algorithm performs crossover and mutation at the bit level. Other variants treat the chromosome as a list of numbers which are indexes into an instruction table, nodes in a linked list, hashes, objects, or any other imaginable data structure. Crossover and mutation are performed so as to respect data element boundaries. For most data types, specific variation operators can be designed. Different chromosomal data types seem to work better or worse for different specific problem domains.

When bit strings representations of integers are used, Gray coding is often employed. In this way, small changes in the integer can be readily affected through mutations or crossovers. This has been found to help prevent premature convergence at so called Hamming walls, in which too many simultaneous mutations (or crossover events) must occur in order to change the chromosome to a better solution.

Other approaches involve using arrays of real-valued numbers instead of bit strings to represent chromosomes. Theoretically, the smaller the alphabet, the better the performance, but paradoxically, good results have been obtained from using real-valued chromosomes.

A very successful (slight) variant of the general process of constructing a new population is to allow some of the better organisms from the current generation to carry over to the next, unaltered. This strategy is known as elitist selection.

Parallel implementations of genetic algorithms come in two flavours. Coarse grained parallel genetic algorithms assume a population on each of the computer nodes and migration of individuals among the nodes. Fine grained parallel genetic algorithms assume an individual on each processor node which acts with neighboring individuals for selection and reproduction. Other variants, like genetic algorithms for online optimization problems, introduce time-dependence or noise in the fitness function.

It can be quite effective to combine GA with other optimization methods. GA tends to be quite good at finding generally good global solutions, but quite inefficient at finding the last few mutations to find the absolute optimum. Other techniques (such as simple hill climbing) are quite efficient at finding absolute optimum in a limited region. Alternating GA and hill climbing can improve the efficiency of GA while overcoming the lack of robustness of hill climbing.

An algorithm that maximizes mean fitness (without any need for the definition of mean fitness as a criterion function) is Gaussian adaptation, See Kjellström 1970[131] provided that the ontogeny of an individual may be seen as a modified recapitulation of evolutionary random steps in the past and

that the sum of many random steps tend to become Gaussian distributed (according to the central limit theorem).
This means that the rules of genetic variation may have a different meaning in the natural case. For instance - provided that steps are stored in consecutive order - crossing over may sum a number of steps from maternal DNA adding a number of steps from paternal DNA and so on. This is like adding vectors that more probably may follow a ridge in the phenotypic landscape. Thus, the efficiency of the process may be increased by many orders of magnitude. Moreover, the inversion operator has the opportunity to place steps in consecutive order or any other suitable order in favour of survival or efficiency.
Gaussian adaptation is able to approximate the natural process by an adaptation of the moment matrix of the Gaussian. So, because very many quantitative characters are Gaussian distributed in a large population, Gaussian adaptation may serve as a genetic algorithm replacing the rules of genetic variation by a Gaussian random number generator working on the phenotypic level. See Kjellström 1996. Population-based incremental learning is a variation where the population as a whole is evolved rather than its individual members.

10.4.7 Problem domains

Problems which appear to be particularly appropriate for solution by genetic algorithms include timetabling and scheduling problems, and many scheduling software packages are based on GAs. GA's have also been applied to engineering. Genetic algorithms are often applied as an approach to solve global optimization problems. As a general rule of thumb genetic algorithms might be useful in problem domains that have a complex fitness landscape as recombination is designed to move the population away from local optima that a traditional hill climbing algorithm might get stuck in.

11. Statistical analysis of genetic algorithms in discovering technical analysis trading strategies

In this PhD thesis, the performance of ordinal GA-based trading strategies is evaluated under six classes of time series model, namely, the linear ARMA model, the bilinear model, the ARCH model, the GARCH model, the threshold model and the chaotic model. The performance criteria employed are the winning probability, accumulated returns, Sharpe ratio and luck coefficient. Asymptotic test statistics for these criteria are derived. The hypothesis as to the superiority of GA over a benchmark, say, buy-and-hold, can then be tested using Monte Carlo simulation. From this rigorously established evaluation process, we find that simple genetic algorithms can work very well in linear stochastic environments, and that they also work very well in nonlinear deterministic (chaotic) environments. However, they may perform much worse in pure nonlinear stochastic

cases. These results shed light on the superior performance of GA when it is applied to the two tick-by-tick time series of foreign exchange rates: EUR/USD and USD/JPY.

11.1. Introduction

Genetic algorithms (GAs) have been developed by Holland (1975)[132] to mimic some of the processes observed in natural evolution. They are based on the genetic processes of natural selection which have become widely known as the "survival of the fittest" since Darwin's celebrated work. In recent years, Gas have been successfully applied to find good solutions to real-world problems whose search space is complex, such as the traveling salesman problem, the knapsack problem, large scheduling problems, graph partitioning problems, and engineering problems, too.

In finance, Bauer (1994)[133] provides the first application of GAs to discover trading strategies. Since then, GAs have gradually become a standard tool for enhancing investment decisions.While many studies have supported the effectiveness of GAs in investment decisions; however, the foundation of these applications has not been well established. The thing that concerns us, therefore, is the robustness of these empirical results. For example, if GAs are effective for the investment in one market at one time, would the same result apply to the same market or different markets at different times?

It is for the purpose of pursuing this generality, that we see the necessity of building a solid foundation upon which a rigorous evaluation can be made. In this PhD thesis, a statistical approach to testing the performance of GA-based trading strategies is proposed. Instead of testing the performance of GAs in specific markets as a number of conventional studies already have, we are interested in a market-independence issue: what makes GAs successful and what makes them not? Since the data to whichGAsare applied consist of financial time series, the question can be rephrased as follows: what are the statistical properties which distinguish a successful application of GA from an unsuccessful one? One way to think of the question is to consider two markets following different stochastic processes. One market follows stochastic process A, and the other stochastic process B. If Gas can work well with stochastic process A, but not B, then the successful experience of GAs in the first market is certainly not anticipated in the second market.

Having said that, this PhD thesis follows the following research methodology. First, some financially-related stochastic processes are singled out as the standard scenarios (testbeds) to test the performance of GA. Second, appropriate performance criteria are used to evaluate the performance of the GA over these testbeds. Third, the associated asymptotic statistical tests are applied to examine whether the GAs perform significantly differently as opposed to a familiar benchmark. By this procedure, we may be able to distinguish the processes in which the GA has competence from others in which it does not. Once the critical properties are grasped, we can then apply the GA to the financial time series whose stochastic properties are well-known, and

test whether the GA behaves consistently with what we have learned from the previous statistical analysis. By means of the procedure established in this PhD thesis, we hope to push forward the current applications of GAs or, more generally, computational intelligence (CI), toward a more mature status. After all, whether GA will work has been asked too intensely in the literature. The very mixed results seem to suggest that we look at the same question at a finer level and start to inquire why it works or why it doesn't. We believe that there are other ways to do something similar to what we propose in this research.

We do not exclude these possibilities. In fact, little by little, these efforts will eventually enable GA or CI tools to rid themselves of their notoriety for being black boxes. The rest of the chapter is organized as follows. Section 11.2 introduces a specific version of GA, referred as to the ordinary GA (OGA), used in this chapter. Section 11.3 will detail the classes of stochastic processes considered in this chapter and the reasons for this choice. Section 11.4 reviews the four performance criteria and establishes their associated asymptotic test. Section 11.5 sets up the Monte Carlo simulation procedure. Section 11.6 summarizes and discusses the actual performance of the GA over the artificial data, whereas the counterpart over the real data is given in Section 11.7. Section 11.8 concludes this chapter.

11.2. Trading with gas

A trading strategy g can be formally defined as a mapping:

$$\boldsymbol{g}: \Omega \to \{0,1\} \quad (259)$$

In this chapter, Ω is assumed to be a collection of finite-length binary strings. This simplification can be justified by the data-pre processing procedure which transforms the rawd ata into binary strings. The range of the mapping g is simplified as a 0–1 action space. In terms of simple market-timing strategy, "1" means to "act" and "0" means to "wait." Here, for simplicity, we are only interested in day trading. So, "act" means to buy it at the opening time and sell it at the closing time. Like all financial applications of GA, the start-off question is the representation issue. In our case, it is about how to effectively characterize the mapping g by a finite-length binary string, also known as a chromosome in GA. Research on this issue is very much motivated by the format of existing trading strategies, and there are generally two approaches to this issue. The first approach, called the decision tree approach, was pioneered by Bauer (1994). In this approach each trading strategy is represented by a decision tree. Bauer used bit strings to encode this decision trees, and generated and evolved them with genetic algorithms. The second approach, called the combinatoric approach, was first seen in Palmer et al. (1994)[134]. The combinatoric approach treats each trading strategy as one realization.

From $\binom{n}{k}$ combinations, where l ≤ k ≤ n, and n is the total number of given trading rules. Using GAs, one can encode the inclusion or exclusion of a specific trading rule as a bit and the whole trading strategy as a bit string (chromosome).

Both approaches have very limited expression power. While various enhancements are possible, they all lead to non-standard GAs in the sense that their representations are not based on finite-length binary strings. Since the main focus of this chapter is to illustrate a statistical foundation of the GA, we try to avoid all unnecessary complications, including the use of those non-standard representations. In other words, at this initial stage, we only make the illustration with the ordinary genetic algorithm (OGA), and, for that reason, Bauer's simple decision-tree representation is employed. However, it is clear that the statistical foundation presented in this chapter is also applicable to GAs with different representations.

Bauer's decision-tree representation corresponds to the following general form of trading strategies

(IF (CONDS) THEN (BUY AND SELL [DAY TRADING]) ELSE (WAIT)).

The CONDS appearing in the trading strategy is a predicate. CONDS itself is a logical composition of several primitive predicates. In this chapter, all CONDSs are composed of three primitive predicates. Each primitive predicate can be represented as:

$$\boldsymbol{Cond}(\boldsymbol{Z}) = \begin{cases} 1(\boldsymbol{True}), & \text{if } \boldsymbol{Z} \oplus \boldsymbol{a}, \\ 0(\boldsymbol{False}), & \text{if } \boldsymbol{Z} \Theta \boldsymbol{a}, \end{cases} \quad (260)$$

where Z, in this application, can be considered as a time series of returns indexed by t, e.g. rt−1, rt−2, etc., and a can be regarded as a threshold or critical value ($\boldsymbol{a} \in \aleph$, a set of integers). $\oplus \in \{\geq, <\}$ and $\theta = \{\geq, <\}$ and $\theta = \{\geq, <\} - \oplus$.

An example ofCONDS with three primitive predicates is

$$\boldsymbol{CONDS}(\boldsymbol{r}_{t-1}, \boldsymbol{r}_{t-2}, \boldsymbol{r}_{t-3}) = \boldsymbol{Cond}(\boldsymbol{r}_{t-1}, ,) \vee (\boldsymbol{Cond}(\boldsymbol{r}_{t-2}) \wedge \boldsymbol{Cond}(\boldsymbol{r}_{t-3})) \quad (261)$$

where "V" refers to the logic operator "OR," and "Λ" refers to "AND."

Following Bauer, we use a 21-bit string to encode a trading strategy of this kind. Details can be found in the Appendix (Section A.1). Let G be the collection of all trading strategies encoded as above. Then the cardinality of G is 2^21 (#(G) = 2^21), which is more than 2 million. The search over the space G can be interpreted as a numerical algorithm as well as a machine learning algorithm for solving a mathematical optimization problem without losing generality, consider the trading strategy with only one primitive predicate,

$$\boldsymbol{Cond}(\boldsymbol{Z}) = \begin{cases} 1(\boldsymbol{True}), & \text{if } \boldsymbol{r}_{\text{t-1}} \geq \boldsymbol{a}, \\ 0(\boldsymbol{False}), & \text{if } \boldsymbol{r}_{\text{t-1}} < \boldsymbol{a}, \end{cases} \quad (262)$$

Suppose the stochastic process of rt is strictly stationary and denote the joint density of rt−1 and rt by f(rt−1, rt). In this simplest case, a trading strategy is parameterized by a single parameter a.

Denote it by ga. Then the optimal strategy ga□ can be regarded as a solution to the optimization problem

$$\max_{a} \boldsymbol{E}\left(\ln\left(\pi_{n}\right)\right) \text{(263)}$$

where

$$\pi_{n} = \prod_{t=1}^{n}(1+\boldsymbol{r}_{t}) \text{ (264)}$$

is the accumulated returns of ga over n consecutive periods. It can be shown that the solution to the problem (265) is

$$\boldsymbol{a}^{*} = \boldsymbol{F}^{-1}(0), \text{ if } \boldsymbol{F}^{-1}(0) \text{ exists (265)}$$

where

$$\boldsymbol{F}(0) = \int_{-\infty}^{\infty} \ln(1+\boldsymbol{r}_{t})\boldsymbol{f}(\boldsymbol{a}, \boldsymbol{r}_{t})\boldsymbol{dr}_{t} \text{ (266)}$$

To solve Eq. (266), one has to know the density function of f(rt−1, rt), which can only be inferred from the historical data. In this case, GAs are used as a machine learning tool to obtain an estimate of this joint density. Also, to arrive at a value for a□, we have to know the inverse function of F(a), which in general can only be solved numerically. In this case, GAs are used as a numerical technique to solve this problem. Therefore, in the trading-strategy problem, GAs are used simultaneously as a numerical technique and a machine learning tool to determine the critical parameter $\boldsymbol{a}^{*}$. In the general case when CONDS has more than one predicate, the mathematical formulation of the problem can become very complicated, but the dual role of GAs remains unchanged. This interpretation justifies the mathematical significance of using GAs to discover the trading strategies.

The GA employed in this chapter is a very basic version, which we shall call the ordinary genetic algorithm (OGA). In this study, we only focus on the OGA. Nonetheless, in a further study, it will be interesting to see whether a better result can be expected from advanced versions of GAs. The technical details of the OGA are given in the Appendix (Section 17.14.2).

11.3. Testing different models

There are six stochastic processes used to evaluate the performance of GAs. They are:

(1) the linear stationary time series (also known as the Auto-Regressive and Moving-Average (ARMA) processes),

(2) the bilinear processes,

(3) the Auto-Regressive Conditional Heteroskedasticity (ARCH) processes,

(4) the Generalized ARCH (GARCH) processes,

(5) the threshold bilinear processes, and

(6) the chaotic processes.

All of the six classes have been frequently applied to modelling financial time series. Linear ARMA processes are found to be quite useful in high-frequency financial data (Campbell et al., 1997; Roll, 1984)[135]. Bilinear processes are often used to model the nonlinear dependence in both low- and high-frequency data (Drunat et al., 1998[136]; Granger & Andersen, 1978[137]). The ARCH processes are the most popular econometric tools for capturing the nonlinear dependence in the form of the second moment (Bollerslev et al., 1992[138]). The threshold processes are good for asymmetric series and bursts (Tong, 1990)[139]. Finally, chaotic time series have been a topic of interest in finance over the last decade (Brock et al., 1991)[140]. Some details of these classes of processes are briefly reviewed from Sections 11.3.1 to 11.3.6 some of them have been also reviewed in previous chapters. These six processes are general enough to cover three important classes of dynamic processes, namely, linear stochastic processes, nonlinear stochastic processes, and nonlinear deterministic processes. This enables us to analyze the GA's performance in terms of some generic properties. For example, would it be easier for the GA to perform better with the linear (stochastic) process than with the nonlinear (stochastic) process, and with the deterministic (nonlinear) processes than with the stochastic (nonlinear) processes? The answers to these questions can certainly help us to delineate the effectiveness of GAs.

11.3.1 Linear Time Series

The linear time series model, also known as the Auto-Regressive and Moving- Average (ARMA(p,q)) model, was initiated by Box and Jenkings (1976)[141]. It has the following general form:

$$r_t = \mu + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{j=1}^{q} \theta_{ji} \varepsilon_{t-j} + \varepsilon_t \quad (267)$$

Table 21: Data Generating Processes – ARMA

Code	Model	Parameters			
		Φ1	Φ2	θ1	θ2
L-1	ARMA(1,0)	0.3	0	0	0
L-2	ARMA(1,0)	0.6	0	0	0
L-3	ARMA(2,0)	0.3	-0.6	0	0
L-4	ARMA(2,0)	0.6	-0.3	0	0
L-5	ARMA(0,1)	0	0	0.3	0
L-6	ARMA(0,1)	0	0	0.6	0
L-7	ARMA(0,2)	0	0	0.3	-0.6
L-8	ARMA(0,2)	0	0	0.6	-0.3
L-9	ARMA(1,1)	0.3	0	-0.6	0
L-10	ARMA(1,1)	0.6	0	-0.3	0
L-11	ARMA(2,2)	0.4	-0.4	0.4	0.4
L-12	ARMA(2,2)	0.6	-0.3	-0.3	-0.6
L-13	White Noise	Gaussian (0,0,1)			

where $\varepsilon_t \overset{iid}{\sim} N(0,\sigma^2)$.In all Monte Carlo simulations conducted in this chapter, μ is set to 0 and σ^2 is set to 0.01. Thirteen ARMA(p,q) models were tested. The parameters of these thirteen ARMA(p,q) models are detailed in Table 21. Among these thirteen models, there are four pure AR models (L1–L4), four pure MA models (L5–L8), and four mixtures (L9–L12). The last one is simply Gaussian white noise.

11.3.2 Bilinear Process

The second class of stochastic processes considered in this chapter is the bilinear process (BL), which was first studied by Granger and Anderson (1978)[142], and subsequently by Subba-Rao (1981)[143] and Subba-Rao and Gabr (1980)[144]. The BL process is constructed simply by adding the cross-product terms of r_{t-i} and ε_{t-j} to a linear ARMA process so it can be regarded as a second-order nonlinear time series model. In other words, if the parameters of all cross-product terms are zero,then the BL process can be reduced to the ARMA process. The general form of a bilinear process, BL(p, q, u, v) is:

$$r_t = \mu + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{m=1}^{u}\sum_{n=1}^{v} \Psi_{mn} r_{t-m}\varepsilon_{t-n} + \varepsilon_t \quad (268)$$

where $\varepsilon_t \sim N(0,\sigma^2)$. Eight specific bilinear processes are employed for the Monte- Carlo simulation. In all of these processes, μ is set to 0 and σ^2 is set to 0.01. Other parameters

Table 22: Data Generating Processes – Bilinear.

Code	Model	Parameters					
		Φ1	θ1	ψ11	ψ12	ψ21	ψ22
BL-1	BL(0,0,1,1)	0	0	0.6	0	0	0
BL-2	BL(0,0,1,1)	0	0	0.3	0	0	0
BL-3	BL(0,1,1,2)	0	0.3	0	0.6	0	0
BL-4	BL(0,1,1,2)	0	0.6	0	0.3	0	0
BL-5	BL(1,0,2,1)	0.3	0	0	0	0.6	0
BL-6	BL(1,0,2,1)	0.6	0	0	0	0.3	0
BL-7	BL(1,1,2,2)	0.3	0.3	0	0	0	0.3
BL-8	BL(1,1,2,2)	0.3	0.3	0	0	0	0.6

are given in Table 22. Notice that the first two (BL-1, BL-2) do not have the linear component, and only the nonlinear cross-product terms are presented.

11.3.3 ARCH Processes

The third class of models considered is the Auto-Regressive Conditional Heteroskedasticity (ARCH) process introduced by Engle (1982)[145], which has played a dominant role in the field of

financial econometrics. The ARCH process is mainly used to replicate the three stylized facts of financial time series, namely, the fat tailed marginal distribution of returns, the time-variant volatility of the returns, and clustering outliers. Consequently, unlike the ARMA process, ARCH mainly works only on the second moment, rather than the first moment. Nonetheless, by combining the two, one can attach an ARMA(p, q) process with an ARCH (q) process, called the ARMA(p, q)-ARCH(q) process. Its general form is

$$r_t = \mu + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{j=1}^{q} \theta_{ji} \varepsilon_{t-j} + \varepsilon_t \quad (269)$$

$$\varepsilon_t \left| \Omega_{t-1} \sim N\left(0, \sigma_t^2\right) \right. \quad (270)$$

$$\sigma_t^2 = \omega + \sum_{m=1}^{q'} \alpha_m \varepsilon_{t-m}^2 \quad (271)$$

where ω > 0, σm ≥ 0, m = 1, . . . , q’ and Ωt denotes the information set available at time t.

Table 23: Data Generating Processes – ARCH

Code	Model	Parameters				
		ω	α1	α2	Φ1	θ1
AH-1	AR(0)-ARCH(1)	0.005	0.3	0	0	0
AH-2	AR(0)-ARCH(1)	0.005	0.6	0	0	0
AH-3	AR(0)-ARCH(2)	0.001	0.3	0.5	0	0
AH-4	AR(0)-ARCH(2)	0.001	0.5	0.3	0	0
AH-5	AR(1)-ARCH(1)	0.005	0.6	0	0.6	0
AH-6	AR(1)-ARCH(2)	0.001	0.5	0.3	0.6	0
AH-7	MA(1)-ARCH(1)	0.005	0.3	0	0	-0.06

Seven ARCH processes are included in this study. They share a common value of μ, which is 0. Values of other parameters are detailed in Table 23. Notice that the first four processes do not have linear signals (Φ1 = 0, θ1 = 0), whereas the fifth and the sixth processes are associated with an AR(1) linear signal (Φ1= 0.6), and the last process has a MA(1) linear signal (θ1 = −0.6).

11.3.4 GARCH Processes

A generalized version of the ARCH process, known as the generalized ARCH (GARCH) process, was introduced by Bollerslev (1986)[146]. GARCH generalizes Engle’s ARCH process by adding additional conditional autoregressive terms. An ARMA(p, q) process with a GARCH error term of order (p , q), ARMA(p, q)- GARCH(p , q), can be written as

$$r_t = \mu + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{j=1}^{q} \theta_{ji} \varepsilon_{t-j} + \varepsilon_t \quad (272)$$

$$\varepsilon_t \left| \Omega_{t-1} \sim N\left(0, \sigma_t^2\right) \right. \quad (273)$$

$$\sigma_t^2 = \omega + \sum_{m=1}^{q'} \alpha_m \varepsilon_{t-m}^2 + \sum_{n=1}^{p} \beta_n \sigma_{t-n}^2 \quad (274)$$

with ω > 0, m = 0 and βn ≥ 0,m = 1, . . . , q', n = 1, . . . , p' .Again,Ωt denotes the information set available at time t.

Nine GARCH processes are attempted. In all cases, μ = 0 and ω = 0.001. Specifications of other parameters are given in Table 24. The 7th, 8th and 9th models (GH-7, GH-8, GH-9) are AR(1) processes combined with a GARCH error term, whereas the last model (GH-10) is a MA(1) process plus a GARCH error term. For the remaining six, there are no linear signals but just pure GARCH processes.

Table 24: Data Generating Processes – GARCH.

Code	Model	Parameters					
		β1	β2	α1	α2	Φ1	θ1
GH-1	AR(0)-GARCH(1,1)	0.3	0	0.5	0	0	0
GH-2	AR(0)-GARCH(1,1)	0.5	0	0.3	0	0	0
GH-3	AR(0)-GARCH(1,2)	0.2	0	0.2	0.4	0	0
GH-4	AR(0)-GARCH(1,2)	0.2	0	0.4	0.2	0	0
GH-5	AR(0)-GARCH(2,1)	0.2	0.4	0.2	0	0	0
GH-6	AR(0)-GARCH(2,1)	0.4	0.2	0.2	0	0	0
GH-7	AR(1)-GARCH(1,1)	0.5	0	0.3	0	0.6	0
GH-8	AR(1)-GARCH(1,2)	0.2	0	0.4	0.02	0.6	0
GH-9	AR(1)-GARCH(1,1)	0.3	0	0.2	0	0.6	0
GH-10	MA(1)-GARCH(1,1)	0.4	0	0.5	0	0	-0.06

11.3.5 Threshold Processes

Tong (1983)[147] proposed a threshold autoregressive (TAR) model which is of the form,

$$r_t = \mu^{(l)} + \sum_{i=1}^{p} \phi_i^{(l)} r_{t-i} + \varepsilon_t \quad (275)$$

if $r_{t-d} \in \Omega_1 \left(l = 1,2,.....k\right)$ where $\Omega_i \in \Omega_j = \Theta\left(i, j = 1,2,.....k\right)$if $i \neq j$ and $\bigcup_{l=1}^{k} \Omega_l = \Re$.

The parameter k represents the number of thresholds and d is called the threshold lag (or delay parameter). Producing various limit cycles is one of the important features of the threshold models, and the TAR process can be applied to the time series which has an asymmetric cyclical form. The threshold idea can be used as a module to add and to extend other processes. Here, we apply the threshold idea to the bilinear process (268), and extend it to a threshold bilinear

(TBL) process. Let us denote a bilinear process (BL(p, q, u, v)) with k-thresholds by TBL(k, p, q, u, v), which can be written as

$$r_t = \mu^{(l)} + \sum_{i=1}^{p} \phi_i^{(l)} r_{t-i} + \sum_{j=1}^{q} \theta_j^{(l)} \varepsilon_{t-j} + \sum_{m=1}^{u}\sum_{n=1}^{v} \Psi_{mn}^{(l)} r_{t-m}\varepsilon_{t-n} + \varepsilon_t \quad (276)$$

Table 25: Data Generating Processes – Threshold Processes

Code	Model	Parameters						
		Φ1 (1)	Φ2 (1)	θ1 (1)	ψ11 (1)	ψ12 (1)	ψ21 (1)	ψ22 (1)
0 0		Φ1 (2)	Φ2 (2)	θ1 (2)	ψ11 (2)	ψ12 (2)	ψ21 (2)	ψ22 (2)
TH-1	TBL(2;1,0,0,0)	0.3	0	0	0	0	0	0
		0.6	0	0	0	0	0	0
TH-2	TBL(2;1,1,0,0)	0.3	0	0.6	0	0	0	0
		0.6	0	0.3	0	0	0	0
TH-3	TBL(2;0,0,1,1)	0	0	0	0.3	0	0	0
		0	0	0	0.6	0	0.6	0
TH-4	TBL(2;1,1,2,2)	0.3	0	0	0	0	0	0
		0	0	0.3	0	0.6	0	0
TH-5	TBL(2;2,0,2,2)	0	0	0	0.3	0	0	-0.6
		0.3	-0.6	0	0	0	0	0

Note: The lag period d is set to 1 and μ(1) = μ (2) = 0 in all of the models. In addition, Ω1 ≡ {rt−d |rt−d ≥ 0} and ┌ 2 ≡ {rt−d |rt−d < 0 }

.It is trivial to show that TBL can be reduced to a threshold ARMA if Ω (l)mm = 0 for all m, n and l. Table 25 lists the five TBL processes considered in this chapter.The motives for choosing these five series will become clear when we come to Section 11.6.4.

11.3.6 Chaotic Processes

All of the above-mentioned processes are stochastic. However, the time series that appear to be random does not necessary imply that they are generated from a stochastic process. Chaotic time series as an alternative description of this seemingly random phenomenon was a popular econometrics topic in the 1990s. While it is hard to believe that a financial time series is just a deterministic chaotic time series, the chaotic process can still be an important module for the working of a nonlinear time series. Five chaotic processes are employed in this study.

C-1: Logistic Map

$$r_t = 4r_{t-1}(1-r_{t-1}),\ r_t \in [0,1]\ \forall t \quad (277)$$

C-2: Henon Map

$$r_t = 1 + 0.3r_{t-2} - 1.4r_{t-1}^2,\ r_{-1}, r_0 \in [-1,1] \quad (278)$$

C-3: Tent Map

$$\begin{cases} r_t = 2r_{t-1}, \ \textit{if}\ 0 \le r_{t-1} < 0.5 \\ r_t = 2(1-r_{t-1}), \ \textit{if}\ 0.5 \le r_{t-1} \le 1 \end{cases} \quad (279)$$

C-4: Poly. 3

$$r_t = 4r_{t-1}^3 - 3r_{t-1},\ r_t \in [-1,1]\ \forall t \quad (280)$$

C-5: Poly. 4

$$r_t = 8r_{t-1}^4 - 8r_{t-1}^2,\ r_t \in [-1,1]\ \forall t \quad (281)$$

The series generated by all these stochastic processes (from Sections 11.3.1 to 11.3.6) may have a range which does not fit the range of the normal return series. For example, the process (278) is always positive. As a result, a contracting or a dilating map is needed. We, therefore, contract or dilate all series linearly and monotonically into an acceptable range, which is (−0.3, 0.3) in this chapter.

11.4. Performance criteria and statistical tests

Basic performance metrics to evaluate the performance of trading strategies have long existed in the literature. Following Refenes (1995)[148], we consider the following four main criteria: returns, the winning probability, the Sharpe ratio and the luck coefficient. In this chapter, the performance of the trading strategies generated by the ordinal genetic algorithm (OGA) is compared with that using a benchmark based on these four criteria. To make the evaluation process rigorous, performance differences between the OGA-based trading strategies and the benchmark are tested statistically. Tests for returns and winning probability are straightforward. Tests for the Sharpe ratio are available in the literature (see, for example, Jobson and Korkie (1981)[149] and Arnold (1990))[150]. However, tests for the luck coefficient are more demanding, and it has not been derived in the literature. In this chapter, we develop asymptotic tests for the luck coefficient.

Returns

Let X and Y be the accumulated returns of an one-dollar investment by applying OGA-based trading strategies and the benchmark strategy, say, the buy-and-hold (B&H) strategy, respectively. Assume that E(X) = μ and E(Y) = v. Let us estimate the μ and v by the respective sample average $\tilde{\pi}^2$ and $\tilde{\pi}^1$ via the Monte Carlo simulation. Then one can test the null

$$H_0 : \mu - \nu \le 0 \quad (282)$$

with the following test statistic

$$Z\pi = \frac{\sqrt{n}\left(\tilde{\pi}^2 - \tilde{\pi}^1\right)}{\left(\hat{\sigma}^2 + \hat{\tau}^2 - 2\hat{\rho}\hat{\sigma}\hat{\tau}\right)^{1/2}} \quad (283)$$

where $\hat{\sigma}^2$ and $\hat{\tau}^2$ are the sample variances of X and Y, $\hat{\rho}$ is the sample correlation coefficient of X and Y, and n is the sample size (the number of ensembles generated during the Monte Carlo simulation). By using the central limit theorem, it is straightforward to show that Zπ is an asymptotically standard normal test. While testing the difference between $\tilde{\pi}^2$ and $\tilde{\pi}^1$ can tell us the performance of the GA as opposed to a benchmark, it provides us with nothing more than a point evaluation. In some cases, we may also wish to know whether the superiority, if shown, can extend to a large class of trading strategies. A common way to address this question is to

introduce an omniscient trader. Let us denote the respective accumulated returns earned by this omniscient trader as $\tilde{\pi}^*$.4 Now, subtracting $\tilde{\pi}^1$ from $\tilde{\pi}^*$ gives us the total unrealized gain, if we only know the benchmark. Then, the ratio, also called the exploitation ratio,

$$\tilde{\pi} \equiv \frac{\tilde{\pi}^2 - \tilde{\pi}^1}{\tilde{\pi}^* - \tilde{\pi}^1} \quad (284)$$

is a measure of the size of those unrealized gains which can be exploited by using a GA. Based on its formulation, $\tilde{\pi}$ may be positive, negative or zero, but has one as its maximum. If $\tilde{\pi}$ is not only positive, but is also close to one, then its superiority is not just restricted to the benchmark, but may also have global significance. In addition to the accumulated gross returns, one can also base the comparison on the excess return by simply subtracting one from the accumulated gross returns. A relative superiority measure of the GA as opposed to the benchmark can be defined accordingly as

$$\tilde{\pi} \equiv \frac{\left(\tilde{\pi}^2 - 1\right) - \left(\tilde{\pi}^1 - 1\right)}{\left|\tilde{\pi}^1 - 1\right|} = \frac{\tilde{\pi}^2 - \tilde{\pi}^1}{\left|\tilde{\pi}^1 - 1\right|} \quad (285)$$

11.4.1 Winning Probability

The mean return can sometimes be sensitive to outliers. Therefore, it is also desirable to base our performance criterion on some robust statistics, and the winning probability is one of this kind. The winning probability basically tells us, by randomly picking up an ensemble from one stochastic process, the probability that the GA will win. Formally, let (X, Y) be a random vector with the joint density function h(x, y). Then pw, defined as follows, is called the winning probability.

$$p_w = P_r\left(X > Y\right) = \int\int_{x>y} h(x, y)dxdy \quad (286)$$

Based on the winning probability, we can say that X is superior to Y if pw > 0.5, and inferior to Y if pw < 0.5, and equivalent to Y if pw = 0.5. The null hypothesis to test is

$$H_0 : p_w \leq 0.5 \quad (287)$$

The rejection of (288) shows the superiority of the GA over the benchmark. An asymptotic standard normal test of (288) can be derived as

$$Z_w = \frac{\sqrt{n}\left(\hat{p}_w - 0.5\right)}{\sqrt{\hat{p}_w\left(1 - \hat{p}_w\right)}} \quad (288)$$

where $\hat{p}_w$ is the sample counterpart of pw.

11.4.2 Sharpe Ratio

One criterion which has been frequently ignored by machine learning people in finance is the risk associated with a trading rule. Normally, a higher profit known as the risk premium is expected

when the associated risk is higher.Without taking the risk into account, we might exaggerate the profit performance of a highly risky trading rule. Therefore, to evaluate the performance of our GA-based trading rule on a risk-adjusted basis, we have employed the well-known Sharpe ratio as the third performance criterion (Sharpe, 1966)[151]. The Sharpe ratio s is defined as the excess return divided by a risk measure. The higher the Sharpe ratio, the higher the risk-adjusted return. Formally, let X ~ f(x) with E(X) = μ and Var(X) = σ^2. Then the value

$$\boldsymbol{s} = \frac{\mu - \boldsymbol{c}}{\sigma} \quad (289)$$

is called the Sharpe ratio of X where c is one plus a risk-free rate. Furthermore, to compare the performance of two trading strategies in the Sharpe ratio, let X ~ f(x) and Y ~ g(y) with E(X) = μ, E(Y) = v, Var(X) = σ^2 and Var(Y)= τ^2.

Then the difference

$$\boldsymbol{d} = \frac{\mu - \boldsymbol{c}}{\sigma} - \frac{\nu - \boldsymbol{c}}{\tau} \quad (290)$$

Jobson and Korkie (1981) derive an asymptotic standard normal test for the Sharpe-ratio differential. However, we do not follow their Taylor expansion formulation. Instead, by applying Slutzky's theorem, the Cramer theorem, and the multivariate central limit theorem, a standard normal test for the null

$$\boldsymbol{H}_0 : \boldsymbol{d} \le 0 \quad (291)$$

can be derived as follows:

$$\boldsymbol{Z}_d = \frac{\sqrt{\boldsymbol{n}}\left(\hat{\boldsymbol{d}} - \boldsymbol{d}\right)}{\hat{\omega}_1} \quad (292)$$

Where

$$\hat{\boldsymbol{d}} = \frac{\overline{\pi}^2 - \boldsymbol{c}}{\hat{\sigma}} - \frac{\overline{\pi}^1 - \boldsymbol{c}}{\hat{\tau}} \quad (293)$$

and

$$\begin{aligned}\hat{\omega}_1^2 = {} & 2(1-\hat{\rho}) + \frac{\overline{\pi}^2 - \boldsymbol{c}}{\hat{\sigma}}\left(\hat{\theta} - \hat{\delta}\right) - \frac{\overline{\pi}^1 - \boldsymbol{c}}{\hat{\tau}}\left(\hat{\psi} - \hat{\xi}\right) \\ & - \frac{\left(\overline{\pi}^2 - \boldsymbol{c}\right)\left(\overline{\pi}^1 - \boldsymbol{c}\right)}{\hat{\sigma}\hat{\tau}}\frac{\left(\hat{\phi} - 1\right)}{2} + \frac{\left(\overline{\pi}^2 - \boldsymbol{c}\right)^2}{\hat{\sigma}^2}\frac{\left(\hat{\gamma} - 1\right)}{4} + \frac{\left(\overline{\pi}^1 - \boldsymbol{c}\right)^2}{\hat{\tau}}\frac{\left(\hat{\eta} - 1\right)}{4}\end{aligned} \quad (294)$$

$\hat{\delta}, \hat{\gamma}, \hat{\xi}$ and $\hat{\eta}$ are the corresponding sample third and fourth moments of X and Y, whereas $\hat{\rho}, \hat{\theta}, \hat{\psi}, \hat{\phi}$ are the corresponding sample mixed moments between X andY (also expressed as Eq. (296)).

$$\begin{bmatrix} \dfrac{\boldsymbol{E}(\boldsymbol{X}-\mu)^3}{\sigma^3} \\ \dfrac{\boldsymbol{E}(\boldsymbol{X}-\mu)^4}{\sigma^4} \\ \dfrac{\boldsymbol{E}(\boldsymbol{Y}-\nu)^3}{\tau^3} \\ \dfrac{\boldsymbol{E}(\boldsymbol{Y}-\nu)^4}{\tau^4} \end{bmatrix} = \begin{bmatrix} \delta \\ \gamma \\ \xi \\ \eta \end{bmatrix}, \begin{bmatrix} \dfrac{\boldsymbol{E}(\boldsymbol{X}-\mu)\boldsymbol{E}(\boldsymbol{X}-\nu)}{\sigma\tau} \\ \dfrac{\boldsymbol{E}(\boldsymbol{X}-\mu)^2(\boldsymbol{Y}-\nu)}{\sigma^2\tau} \\ \dfrac{\boldsymbol{E}(\boldsymbol{X}-\mu)(\boldsymbol{Y}-\nu)^2}{\sigma\tau^2} \\ \dfrac{\boldsymbol{E}(\boldsymbol{X}-\mu)^2(\boldsymbol{Y}-\nu)^2}{\sigma^2\tau^2} \end{bmatrix} = \begin{bmatrix} \rho \\ \theta \\ \psi \\ \phi \end{bmatrix} \quad (295)$$

11.4.3 Luck Coefficient

The largest positive trade can be very important if it makes a significant contribution towards skewing the average profit dramatically. When this happens, people can be severely misled by the sample mean. As a solution to this problem, the trimmed mean is often used in statistics. A similar idea in finance is known as the luck coefficient. The luck coefficient l_ε is defined as the sum of the largest 100ε% returns, ε ∈ (0, 1), divided by the sum of total returns. In a sense, the larger the luck coefficient, the weaker the reliability of the performance. The luck coefficient, as a performance statistic, is formally described below.

Let {X1, X2, . . ., Xm} be a random sample from f(x) with E(X) = μ. The order statistic of this random sample can be enumerated as X(1), X(2), . . ., X(m), where X(1) ≤ X(2) ≤ · · · ≤ X(m). Then, from the order statistics, it is well known that

$$\boldsymbol{X}_{(\boldsymbol{m})} \sim \mathrm{g}\left(\mathrm{x}_{(\mathrm{m})}\right) = \boldsymbol{m}\left[\boldsymbol{F}\left(\mathrm{x}_{(\mathrm{m})}\right)\right]^{\boldsymbol{m}-1} \mathrm{f}\left(\mathrm{x}_{(\mathrm{m})}\right) \quad (296)$$

where F is the distribution function of X. Furthermore, let $\boldsymbol{X}_i \overset{\text{iid}}{\sim} \mathrm{f}(\mathrm{x})$ i=1,2…,m and X(m) ~ g(x(m)) as described above with E(X(m)) = μ_ε .Then the ratio

$$\boldsymbol{l}_\varepsilon = \frac{\varepsilon\mu_\varepsilon}{\mu} \quad (297)$$

is called the luck coefficient of X where ε = 1/m. In this thesis, ε is set to 0.05. Here we want to see how much of the contribution to mean returns comes from the largest 5% of trades. For making a comparison between strategies, the luck coefficient ratio is defined as follows. Let

$X_i \overset{\text{iid}}{\sim} \mathrm{f}_\mathrm{x}(\mathrm{x})$ with E(X) = μ , $Y_i \overset{\text{iid}}{\sim} \mathrm{f}_\mathrm{y}(\mathrm{y})$ with E(Y) = v, i = 1, 2, . . . , m and $X_{(m)} \sim \mathrm{g}_\mathrm{x}\left(x_{(m)}\right)$ with

$E\left(X_{(m)}\right) = \mu_\varepsilon$, $Y_{(m)} \sim \mathrm{g}_\mathrm{y}\left(\mathrm{y}_{(\mathrm{m})}\right)$ with $E\left(Y_{(m)}\right) = \nu_\varepsilon$. Then the ratio

$$\boldsymbol{r}_{\varepsilon} = \frac{\varepsilon \nu_{\varepsilon} / \nu}{\varepsilon \mu_{\varepsilon} / \mu} = \frac{\mu \nu_{\varepsilon}}{\nu \mu_{\varepsilon}} \quad (298)$$

is called the luck-coefficient ratio of X relative to Y where $\varepsilon = \frac{1}{m}$. Based on this definition, X is said to have a lower (higher) luck coefficient relative to Y if $r_{\varepsilon} > 1 \ \left(r_{\varepsilon} < 1\right)$. Otherwise, X and Y are said to be identical in terms of the luck coefficient.

However, to the best of our knowledge, the respective asymptotic standard normal test for the nul

$$\boldsymbol{H}_0 : \boldsymbol{r}_{\varepsilon} \leq 1 \quad (299)$$

is not available in the literature. Nevertheless, similar to the derivation of the test of the Sharpe ratio (292), it is not hard to cook up such a test by using Slutzky's theorem, the Cramer theorem, and the multivariate central limit theorem, which is given in Eq. 301.

$$\boldsymbol{Z}_r = \frac{\sqrt{\boldsymbol{n}}\left(\hat{\boldsymbol{r}}_{\varepsilon} - \boldsymbol{r}_{\varepsilon}\right)}{\hat{\omega}_2} \quad (300)$$

where

$$\hat{\boldsymbol{r}}_{\varepsilon} = \frac{\overline{\pi}^2 \overline{\pi}^1_{\boldsymbol{m}}}{\overline{\pi}^1 \overline{\pi}^2_{\boldsymbol{m}}} \quad (301)$$

and

$$\begin{aligned}
\hat{\omega}_2^2 = {} & \frac{\varepsilon\left(\overline{\pi}^1_m\right)^2}{\left(\overline{\pi}^1_m\right)^2\left(\overline{\pi}^2_m\right)^2}\left(\hat{\sigma}^2 + \frac{\left(\overline{\pi}^2\right)^2 \hat{\tau}^2}{\left(\overline{\pi}^2\right)^2}\right) + \frac{\left(\overline{\pi}^2\right)^2}{\left(\overline{\pi}^1\right)^2\left(\overline{\pi}^2_m\right)^2}\left(\hat{\tau}^2_{\varepsilon} + \frac{\left(\overline{\pi}^1_m\right)^2 \hat{\sigma}^2_{\varepsilon}}{\left(\overline{\pi}^2_m\right)^2}\right) \\
& - \frac{2\overline{\pi}^2 \overline{\pi}^1_m \hat{\tau}}{\left(\overline{\pi}^1\right)^3\left(\overline{\pi}^2_m\right)^2}\left(\varepsilon \overline{\pi}^1_m \hat{\rho}\hat{\sigma} + \overline{\pi}^2 \hat{\lambda}\hat{\tau}_{\varepsilon}\right) - \frac{2\overline{\pi}^2 \overline{\pi}^1_m \hat{\sigma}_{\varepsilon}}{\left(\overline{\pi}^1\right)^2\left(\overline{\pi}^2_m\right)^3}\left(\overline{\pi}^1_m \hat{o}\hat{\imath} + \overline{\pi}^2 \hat{\tau}_{\varepsilon}\hat{\zeta}\right) \quad (302) \\
& + \frac{2\overline{\pi}^2 \overline{\pi}^1_m}{\left(\overline{\pi}^1\right)^2\left(\overline{\pi}^2_m\right)^2}\left(\hat{\sigma}\hat{\kappa}\hat{\tau}_{\varepsilon} + \frac{\overline{\pi}^2 \overline{\pi}^1_m \hat{\sigma}_{\varepsilon}\hat{o}\hat{\tau}}{\overline{\pi}^1 \overline{\pi}^2_m}\right)
\end{aligned}$$

$\overline{\pi}^1_m, \overline{\pi}^2_m$ are the corresponding sample means of Y(m) and X(m). $\hat{\tau}^2_{\varepsilon}$ and $\hat{\sigma}^2_{\varepsilon}$ are the corresponding sample variances of Y(m) and X(m), and $\hat{\rho}, \hat{\zeta}, \hat{\kappa}, \hat{\imath}, \hat{\lambda},$ and $\hat{o}$ are the corresponding sample correlation coefficients as indicated in Eq. (304).

$$\begin{bmatrix} \boldsymbol{corr}\left(\boldsymbol{X}_{\boldsymbol{i}}, \boldsymbol{Y}_{\boldsymbol{i}}\right) \\ \boldsymbol{corr}\left(\boldsymbol{X}_{(\boldsymbol{m})}, \boldsymbol{Y}_{(\boldsymbol{m})}\right) \\ \boldsymbol{corr}\left(\boldsymbol{X}_{\boldsymbol{i}}, \boldsymbol{Y}_{(\boldsymbol{m})}\right) \end{bmatrix} = \begin{bmatrix} \rho \\ \zeta \\ \kappa \end{bmatrix}, \begin{bmatrix} \boldsymbol{corr}\left(\boldsymbol{X}_{\boldsymbol{i}}, \boldsymbol{X}_{(\boldsymbol{m})}\right) \\ \boldsymbol{corr}\left(\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{Y}_{(\boldsymbol{m})}\right) \\ \boldsymbol{corr}\left(\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}_{(\boldsymbol{m})}\right) \end{bmatrix} = \begin{bmatrix} \imath \\ \lambda \\ o \end{bmatrix} \quad (303)$$

11.5. Monte carlo simulation

Since it is hard to obtain analytical results of the performance of the GA in relation to various stochastic processes, Monte Carlo simulation methodology is used in this study. Each stochastic

process listed in Tables 1–5 and Eqs (278) to (282) is used to generate 1000 independent time series, each with 105 observations $\left(\{r_t\}_{t=1}^{105}\right)$.

For each series, the first 70 observations $\left(\{r_t\}_{t=1}^{70}\right)$are taken as the training sample, and the last 35 observations $\left(\{r_t\}_{76}^{105}\right)$ are used as the testing sample. The OGA are then employed to extract trading strategies from these training samples. These strategies are further tested by the testing samples, and the resulting accumulated returns (π) are calculated, i.e.

$$\pi = \prod_{t=76}^{105}(1 + r_t) \quad (304)$$

In the meantime, the accumulated returns of the benchmark are also calculated. In following convention, our choice of the benchmark is simply the buy-and-hold (B&H) strategy.

Let π_i^1 (i = 1, 2, . . ., 1000) be the accumulated returns of the B&H strategy when tested on the ith ensemble of a stochastic process, and ┌ 2i be the accumulated returns of the OGA when tested on the same ensemble. The issue which we shall address, given the set of observations $S\left(\equiv\{\pi_i^1, \pi_i^2\}_{i=1}^{1000}\right)$ is to decide whether the OGA-based trading strategies can statistically significantly outperform the B&H strategy under the stochastic process in question. The answers are given in the next section.

11.6. Test results

11.6.1 ARMA Processes

We start our analysis from the linear stochastic processes. Table 25 summarizes the statistics defined in Section 4. Several interesting features stand out. First, from the statistics $\hat{p}_w$ and z_w, it can be inferred that, in accumulated returns, the probability that the OGA-based trading strategies can beat the B&H strategy is significantly greater than 0.5. For the stochastic processes with linear signals (L-1–L-12), the winning probability $\hat{p}_w$ ranges from 0.713 (L-5) to 0.991 (L-12). What, however, seems a little puzzling is that, even in the case of white noise (L-13), theGAcan also beat B&H statistically significantly, while with much lower winning probabilities pw (0.606). This seemingly puzzling finding may be due to the fact that a pseudorandom generator can actually generate a series with signals when the sample size is small. For example, Chen and Tan (1999)[152] show that, when the sample size is 50, the probability of having signals in a series generated from a pseudo-random generator is about 5%, while that probability can go to zero when the sample size is 1000. Therefore, by supposing that the OGA-based trading strategies can win in all these atypical ensembles and get even with the B&H strategy in other normal ensembles, then ˆpw can still be significantly greater than 0.5. Second, by directly comparing $\bar{\pi}^1$ with $\bar{\pi}^2$, we can see that, except for the case of white noise, the OGA-based trading strategies unanimously outperform the B&H strategy numerically in all linear ARMA(p, q) processes. From

the $\dot{\pi}$ statistic (27), we see that the triumph of GA over B&H extends from a low of 19% (L-10) to a high of 916% (L-3). The z_{π} statistic, ranging from 2.12 to 47.39, signifies the statistical significance of these differences. Third, to see how the GA effectively exploited the excess potential returns earned by the omniscient trader, $\tilde{\pi}$ is also included in Table 25. There it is observed that the GA exploited 2–31% .

Table 26: Performance Statistics of the OGA and B&H – ARMA

Code	Model	$\overline{\pi}^1$	$\overline{\pi}^2$	$\overline{\pi}^*$	z_w	$\overline{\pi}(\%)$	$\dot{\pi}(\%)$	p_w	z_w
L-1	ARMA(1,0)	1.198	1.355	4.388	6.33	4	20	0.732	16.56
L-2	ARMA(1,0)	1.992	2.868	6.658	13.67	19	88	0.859	32.62
L-3	ARMA(2,0)	0.845	2.265	5.480	42.98	31	916	0.976	98.35
L-4	ARMA(2,0)	1.123	1.185	5.170	27.08	2	50	0.896	41.02
L-5	ARMA(0,1)	1.103	1.269	4.241	7.63	5	161	0.713	14.89
L-6	ARMA(0,1)	1.199	1.775	5.166	20.61	15	289	0.861	32.99
L-7	ARMA(0,2)	0.853	1.633	5.104	39.97	18	531	0.926	51.46
L-8	ARMA(0,2)	1.065	1.522	5.285	21.58	11	703	0.848	30.65
L-9	ARMA(1,1)	0.898	1.229	4.128	24.55	10	325	0.812	25.25
L-10	ARMA(1,1)	1.452	1.538	4.783	2.12	3	19	0.721	15.58
L-11	ARMA(2,2)	1.306	2.588	6.957	30.4	23	23	19 0.	51.90
L-12	ARMA(2,2)	0.721	2.167	6.189	47.3	26	26	18 0.	164.40
L-13	ARMA(0,0)	0.983	0.993	3.881	0.67	0	59	0.606	6.85

Code	Model	$\hat{s}_1$	$\hat{s}_2$	$\hat{d}$	z_d	$\gamma^1_{0.05}$	$\gamma^2_{0.05}$	$\tau_{0.05}$	z_r
L-1	ARMA(1,0)	0.166	0.438	0.272	11.74	0.179	0.126	1.416	3.32
L-2	ARMA(1,0)	0.236	0.526	0.290	8.40	.310 0	0.214	1.450	1.75
L-3	ARMA(2,0)	−0.342	1.181	1.523	32.14	0.115	0.106	1.087	1.68
L-4	ARMA(2,0)	0.111	0.877	0.767	24.53	0.182	0.114	1.594	4.45
L-5	ARMA(0,1)	0.110	0.419	0.309	13.40	0.169	0.117	1.449	4.23
L-6	ARMA(0,1)	0.135	0.602	0.467	5.02	.216 0	0.138	1.563	2.48
L-7	ARMA(0,2)	−0.353	0.948	1.301	27.67	0.108	0.099	1.092	1.68
L-8	ARMA(0,2)	0.065	0.624	0.559	18.18	0.181	0.120	1.509	4.18
L-9	ARMA(1,1)	−0.307	0.524	0.831	22.43	0.093	0.092	1.007	0.16
L-10	ARMA(1,1)	0.214	0.392	0.177	5.39	0.263	0.171	1.534	2.50
L-11	ARMA(2,2)	0.170	0.854	0.684	11.19	0.240	0.141	1.708	3.34
L-12	ARMA(2,2)	−1.363	1.224	2.587	36.46	0.083	0.105	0.795	−6.21
L-13	ARMA(0,0)	−0.025	−0.016	0.010	0.37	0.130	0.096	1.353	3.90

Note:[1]

[1] Note: $\overline{\pi}^1$, $\overline{\pi}^2$ and $\overline{\pi}^*$ are the respective sample mean return of OGA, B&H and the omniscient trader. $\tilde{\pi}$ is the exploitation ratio (Eq. (285)), and $\dot{\pi}$ is the relative superiority index (Eq. (286)). p_w is the sample winning probability of OGA over B&H (Eq. (287)). $\hat{s}_1$ and $\hat{s}_2$ are the corresponding sample Sharpe ratio of OGA and B&H (Eq. (290)). Their sample difference is $\hat{d}$ (Eq. (32)). $\hat{l}^1_{0.05}$ land $\hat{l}^2_{0.05}$ are the sample luck coefficient of OGA and B&H (Eq. (298)), and $\hat{r}_{0.05}$ is the sample luck coefficient ratio between the two (Eq. (299)). The z_{π}, z_w, z_d and z_r are the test statistics of the mean return difference, winning probability, Sharpe ratio differential, and luck coefficient ratio, respectively. The critical value of them is 1.28 at the 10% significance level, and is 1.64 at the 5% significance level.

of the potential excess returns. However, as we expect, it was to no avail when the scenario changed to white noise. As mentioned earlier, we should not judge the performance of the GA solely by the profitability criterion. The risk is a major concern in business practice. We, therefore, have also calculated the Sharpe ratio, a risk-adjusted profitability criterion. It is interesting to notice that in all cases the Sharpe-ratio differential (ˆd) is positive. In other words, the GA still outperforms B&H even after taking into account the risk. The test of this differential also lends support to its statistical significance.

Finally, we examine whether the GA wins just by luck in the sense that its return performance depends heavily on its best 5% trades. Based on the statistic of luck coefficient $\hat{r}_{0.05}$, it is found that in only one of the 13 cases, i.e. the case L-12, dose the GA have a higher luck coefficient; in the other 12 cases, the luck-coefficient ratios are larger than 1, meaning that the dominance of the GA over B&H cannot be attributed to the presence of a few abnormally large returns. From the test z_r, this result is again significant except for the case L-9. All in all, we can conclude that if the return follows a simple linear ARMA process, then the superior performance of the GA compared to B&H is expected.

11.6.2 Bilinear Processes

By moving into the bilinear processes, we are testing the effectiveness of the GA when the return series is nonlinear. Table 27 summarizes all the key statistics. Obviously, the performance of the GA is not as glamorous as before. Out of the eight battles, it loses twice (cases BL-1 and BL-2) to B&H (see z_r and z_w). Taking the risk into account would not help reverse the situation (see zd). It is, however, interesting to notice a unique feature shared by BL-1 and BL-2. As mentioned in Section 3.2, the two stochastic processes do not have any linear component (all ϕ_i and ϕ_j in Eq. (259) or Table 21 are zero). In other words, these two cases are pure nonlinear (pure bilinear). If some linear components are added back to the series, then the significant dominance of the GA does come back. This is exactly what happens in the other six cases (BL-3 to BL-8), which all have the ARMA component as a part (Table 21).

Even for the six cases where the GA wins, we can still observe some adverse impacts of nonlinearity on the GA. Roughly speaking, Table 26 shows that the distribution of both $\dot{\pi}$ and $\tilde{\pi}$ becomes lower as opposed to those items observed in the linear stochastic processes. So, not only does the advantage of the GA relative to B&H shrink, but its disadvantage relative to the omniscient also becomes larger.

However, nonlinearity does not change many of the results in relation to the luck coefficients. The luck-coefficient ratios are all higher than 1, and most of the results are statistically significant, indicating the relative stability of the GA.

Table 27: Performance Statistics of the OGA and B&H – Bilinear

Code	Model	$\overline{\pi}^1$	$\overline{\pi}^2$	$\overline{\pi}^*$	z_w	$\overline{\pi}(\%)$	$\dot{\pi}(\%)$	p_w	z_w
BL-1	BL(0,0,1,1)	1.253	1.126	4.398	−6.78	−4	−50	0.491	–0.57
BL-2	BL(0,0,1,1)	1.151	1.064	4.228	−4.66	−3	−58	0.517	1.08
BL-3	BL(0,1,1,2)	1.302	1.830	5.341	11.50	13	175	0.861	17.78
BL-4	BL(0,1,1,2)	1.186	1.356	4.449	6.95	5	91	0.745	17.78
BL-5	BL(1,0,2,1)	1.260	1.419	4.539	5.07	5	61	0.747	17.97
BL-6	BL(1,0,2,1)	2.292	3.143	7.226	9.89	17	66	0.877	36.30
BL-7	BL(1,1,2,2)	1.841	2.471	6.448	8.83	14	75	0.848	30.65
BL-8	BL(1,1,2,2)	1.602	2.287	5.894	19.57	16	114	0.870	34.79

Code	Model	$\hat{s}_1$	$\hat{s}_2$	$\hat{d}$	z_d	$\gamma^1_{0.05}$	$\gamma^2_{0.05}$	$\tau_{0.05}$	z_r
BL-1	BL(0,0,1,1)	0.316	0.251	–0.065	–3.29	0.132	0.105	1.256	3.30
BL-2	BL(0,0,1,1)	0.190	0.144	–0.046	–2.21	0.144	0.101	1.427	4.14
BL-3	BL(0,1,1,2)	0.167	0.425	0.259	7.31	0.182	0.124	1.793	3.08
BL-4	BL(0,1,1,2)	0.162	0.724	0.562	16.32	0.232	0.129	1.465	3.22
BL-5	BL(1,0,2,1)	0.178	0.465	0.287	13.53	0.211	0.138	1.531	3.54
BL-6	BL(1,0,2,1)	0.251	0.539	0.289	10.38	0.346	0.226	1.534	2.05
BL-7	BL(1,1,2,2)	0.285	0.711	0.426	9.29	0.270	0.168	1.603	2.67
BL-8	BL(1,1,2,2)	0.179	0.386	0.207	2.52	0.272	0.182	1.494	1.14

Note:1

11.6.3 ARCH and GARCH Processes

As we have already seen from the bilinear processes, nonlinearity can have some adverse effects on the performance of the GA. It would be imperative to know whether this finding is just restricted to a specific class of nonlinear processes or can be generalized to other nonlinear processes. In this and the next two sections, we shall focus on this question, and briefly mention other details when we see the necessity.

Let us first take a look at the results of the other two nonlinear stochastic processes, namely, ARCH and GARCH. Just like what we saw in the bilinear processes, these two classes of processes can become pure nonlinear stochastic if some specific coefficient values are set to zero. This is basically what we do

Table 28 : Performance Statistics of the OGA and B&H – ARCH

Code	Model	$\overline{\pi}^1$	$\overline{\pi}^2$	$\overline{\pi}^*$	z_w	$\overline{\pi}(\%)$	$\pi(\%)$	p_w	z_w
AH-1	AR(0)-ARCH(1)	1.038	1.013	3.195	−1.99	−1	−66	0.546	2.92
AH-2	AR(0)-ARCH(1)	1.001	1.005	4.251	0.19	0	400	0.592	5.92
AH-3	AR(0)-ARCH(2)	0.985	0.991	2.307	0.67	0	40	0.562	3.95
AH-4	AR(0)-ARCH(2)	1.007	0.997	2.268	−1.09	−1	−143	0.529	1.84
AH-5	AR(1)-ARCH(1)	1.175	1.509	2.187	22.88	33	191	0.862	33.1
AH-6	AR(1)-ARCH(2)	1.300	1.705	3.061	17.64	23	135	0.838	29.0
AH-7	MA(1)-ARCH(1)	0.869	1.551	3.602	44.12	25	521	0.959	73.2

Code	Model	$\hat{s}_1$	$\hat{s}_2$	$\hat{d}$	z_d	$\gamma^1_{0.05}$	$\gamma^2_{0.05}$	$\tau_{0.05}$	z_r
AH-1	AR(0)-ARCH(1)	0.170	0.038	–0.032	−1.33	0.117	0.091	1.285	4.53
AH-2	AR(0)-ARCH(1)	0.001	0.010	0.009	0.34	0.149	0.105	1.411	3.19
AH-3	AR(0)-ARCH(2)	−0.038	−0.035	0.002	0.09	0.100	0.079	1.269	4.03
AH-4	AR(0)-ARCH(2)	0.017	−0.012	−0.030	−1.22	0.099	0.080	1.246	3.24
AH-5	AR(1)-ARCH(1)	0.211	0.774	0.563	15.42	0.145	0.109	1.331	3.43
AH-6	AR(1)-ARCH(2)	0.221	0.605	0.384	10.79	0.187	0.140	1.332	2.15
AH-7	MA(1)-ARCH(1)	−0.641	1.126	1.766	35.75	0.076	0.086	0.889	−3.44

Note:1

in Tables 23 and 24. Notice that, based on these settings, AH-1 to AH-4 (ARCH) and GH-1 to GH-6 (GARCH) are all pure nonlinear stochastic processes, i.e. pure ARCH or pure GARCH without linear ARMA components. For the rest, they are a mixture of pure ARCH (GARCH) and linear ARMA processes. Tables 27 and 28 summarize the results of the two stochastic processes. A striking feature is that, in contrast to its performance in mixed processes, the GA performed dramatically worse in pure nonlinear ARCH and GARCH scenarios.

Let us take the ARCH processes as an illustration. In the mixed processes AH-5, AH-6 and AH-7, the GA has a probability of up to 80% or higher of beating B&H, and earned 135–521% more than B&H. The fact that these excess returns are not compensation for risk is further confirmed by the Sharpe-ratio differentials which are significantly positive. In addition, the GA exploited 23–33% of the potential returns earned by the omniscient trader. However, when coming to the pure

Table 29: Performance Statistics of the OGA and B&H – GARCH

Code	Model	$\overline{\pi}^1$	$\overline{\pi}^2$	$\overline{\pi}^*$	z_w	$\overline{\pi}(\%)$	$\dot{\pi}(\%)$	p_w	z_w
GH-1	AR(0)-GARCH(1,1)	0.987	0.983	2.457	−0.42	0	−31	0.539	2.47
GH-2	AR(0)-GARCH(1,1)	0.968	0.979	2.580	1.19	1	34	0.554	3.44
GH-3	AR(0)-GARCH(1,2)	1.008	1.007	2.474	−0.04	0	−13	0.544	2.79
GH-4	AR(0)-GARCH(1,2)	0.998	1.007	2.434	0.90	1	50	0.572	4.60
GH-5	AR(0)-GARCH(2,1)	0.978	1.001	2.637	2.24	1	05	0.584	5.39
GH-6	AR(0)-GARCH(2,1)	0.982	0.997	2.595	1.50	1	3 0	0.563	4.02
GH-7	AR(1)-GARCH(1,1)	1.428	1.926	3.511	18.40	24	116	0.856	32.07
GH-8	AR(1)-GARCH(1,2)	1.356	1.747	3.298	12.58	20	110	0.841	29.49
GH-9	AR(1)-GARCH(1,1)	1.378	1.934	3.616	19.20	25	147	0.872	35.21
GH-10	MA(1)-GARCH(1,1)	0.911	1.376	2.769	36.44	25	521	0.949	49 64

Code	Model	$\hat{s}_1$	$\hat{s}_2$	$\hat{d}$	z_d	$\gamma^1_{0.05}$	$\gamma^2_{0.05}$	$\tau_{0.05}$	z_r
GH-1	AR(0)-GARCH(1,1)	−0.030	−0.652	−0.035	−1.19	0.101	0.079	1.282	4.30
GH-2	AR(0)-GARCH(1,1)	−0.080	−0.076	0.004	0.17	0.098	0.081	1.202	4.08
GH-3	AR(0)-GARCH(1,2)	−0.005	0.020	0.024	1.05	0.094	0.081	1.166	3.32
GH-4	AR(0)-GARCH(1,2)	0.020	0.026	0.007	0.27	0.108	0.093	1.151	1.68
GH-5	AR(0)-GARCH(2,1)	−0.051	0.005	0.056	2.04	0.103	0.083	1.233	4.10
GH-6	AR(0)-GARCH(2,1)	−0.044	−0.012	0.032	1.23	0.097	0.083	1.178	3.50
GH-7	AR(1)-GARCH(1,1)	0.244	0.620	0.375	11.06	0.225	0.158	1.426	2.72
GH-8	AR(1)-GARCH(1,2)	0.231	0.614	0.383	14.52	0.201	0.143	1.405	2.59
GH-9	AR(1)-GARCH(1,1)	0.703	0.239	0.465	13.47	0.213	0.147	1.454	3.13
GH-10	MA(1)-GARCH(1,1)	−0.476	1.034	1.509	29.43	0.070	0.081	0.867	−3.90

Note 1

nonlinear processes AH-1 to AH-4, this dominance either disappears or becomes weaker. This can be easily shown by the sharp decline in the statistics z_{Γ}, zw and zd in Table 28 with an almost 0% exploitation ($\widetilde{\pi}$) of the maximum potential returns.

This discernible pattern also extends toTable 28. The double-digit z_π, z_w, z_d and z_r of the mixed processes (GH-7 to GH-10) distinguish themselves from the low, or even negative, single-digit ones of the pure nonlinear processes (GH-1 to GH-6). For the former, the GA has 84–95% chance of beating B&H and earned 110–521% more than B&H. Again, from zd, we know that the high returns are more than compensation for risk.Very similar to the case of ARCH, 20–25% of the maximum potential returns can be exploited by the GA, but that value $\widetilde{\pi}$ drops near to 0% when the underlying processes change to pure GARCH.

Despite the fact that pure nonlinear processes continue to deal the GA a hard blow, as far as the winning probability is concerned, its relative performance to B&H is overwhelmingly good. This

can be reflected by the z_w statistics which are consistently significantly positive in all cases. A similar property holds for the luck coefficient (see z_r). The only two exceptions are the cases AH-7 and GH-10, which, however, are not pure nonlinear. In fact, they both have MA(1) as their linear component.

11.6.4 Threshold Processes

The threshold process leads to a different kind of nonlinear process. While its global behavior is nonlinear, within each local territory, characterized by Ω_i, it can be linear. TH-1 and TH-2 in Table 25 are exactly processes of this kind. The former is switching between two AR(1) processes, whereas the latter is switching between two ARMA(1,1) processes. Since the GA can work well with linear processes, it would be interesting to know whether its effectiveness will extend to these local linear processes. Our results are shown in Table 30. The four statistics z_π, z_w, z_d and z_r all give positive results. The GA is seen to exploit 20–30% of the maximum potential returns, and the winning probabilities are greater than 90%. TH-4 and TH-5 are another kind of complication. TH-4 switches between two mixed processes, while TH-5 switches between a pure nonlinear process and a linear process. From previous experiences, we already knew that the GA can work well with the mixed process. Now, from Table 30, it seems clear that it can survive these two complications as well. Finally, we come to the most difficult one TH-5, i.e the one which switches between two pure nonlinear (bilinear) processes. Since the GA did not show its competence in the pure nonlinear process, at least from the perspective of the return criteria, one may conjecture that TH-5 will deal another hard blow to the

Table 30: Performance Statistics of the OGA and B&H – Threshold

Code	Model	$\bar{\pi}^1$	$\bar{\pi}^2$	$\bar{\pi}^*$	z_w	$\bar{\pi}(\%)$	$\dot{\pi}(\%)$	p_w	z_w
TH-1	TBL(2;1,0,0,0)	0.612	1.233	3.372	24.89	23	160	0.910	45.30
TH-2	TBL(2;1,1,0,0)	1.262	2.743	6.361	21.15	29	565	0.931	53.77
TH-3	TBL(2;0,0,1,1)	1.161	1.074	4.207	–4.38	–3	–54	0.502	0.13
TH-4	TBL(2;1,1,2,2)	1.271	1.406	4.497	5.41	4	50	0.717	15.23
TH-5	TBL(2;2,0,2,2)	0.654	1.236	3.890	37.38	18	168	0.919	48.56

Code	Model	$\hat{s}_1$	$\hat{s}_2$	$\hat{d}$	z_d	$\gamma^1_{0.05}$	$\gamma^2_{0.05}$	$\tau_{0.05}$	z_r
TH-1	TBL(2;1,0,0,0)	–0.398	0.374	0.772	9.33	0.267	0.119	2.252	4.30
TH-2	TBL(2;1,1,0,0)	0.093	0.727	0.634	11.86	0.329	0.163	2.012	2.95
TH-3	TBL(2;0,0,1,1)	0.208	0.176	–0.032	–1.42	0.136	0.098	1.394	3.72
TH-4	TBL(2;1,1,2,2)	0.208	0.426	0.219	10.41	0.192	0.140	1.379	2.97
TH-5	TBL(2;2,0,2,2)	–0.813	0.484	1.297	16.8	0.130	0.097	7 1.3	3.54

Note1

GA. Both z┌ and zd in Table 30 confirm this conjecture. Not just the returns, but zw shows that the winning probability is also not good, which is similar to what we experienced in BL-1 and BL-

2. The only criterion that remains unaffected by this complication is the luck coefficient. Furthermore, it turns out that z_r seems to give the most stable performance across all kinds of processes considered so far, except the MA process.

11.6.5 Chaotic Processes

Chaotic processes are also nonlinear, but they differ from the previous four nonlinear processes in that they are deterministic rather than stochastic. These processes can behave quite erratically without any discernible pattern. Can the GA survive well with this type of nonlinear process? The answer is a resounding yes.

All the statistics in Table 31 are sending us this message.

The winning probabilities are all higher than 85%. In the case of the Henon map (C-2), the GA even beats B&H in all of the 1000 trials. In addition, in this

Table 31: Performance Statistics of the OGA and B&H – Chaos

Code	$\overline{\pi}^1$	$\overline{\pi}^2$	$\overline{\pi}^*$	z_w	$\overline{\pi}(\%)$	$\tilde{\pi}(\%)$	p_w	z_w
C-1	1.019	5.664	21.876	31.15	22	24447	0.993	186.99
C-2	5.387	23.235	33.452	85.62	64	407	1.000	*
C-3	0.937	4.124	11.374	44.65	31	5059	0.990	352.49
C-4	1.188	3.066	25.563	22.91	8	999	0.950	65.29
C-5	0.928	1.790	23.172	17.18	4	1197	0.876	36.08

Code	$\hat{s}_1$	$\hat{s}_2$	$\hat{d}$	z_d	$\gamma^1_{0.05}$	$\gamma^2_{0.05}$	$\tau_{0.05}$	z_r
C-1	0.009	0.832	0.824	16.59	0.297	0.184	1.615	2.28
C-2	1.600	2.502	0.901	23.56	0.112	0.090	1.252	4.39
C-3	−0.075	1.160	1.235	28.92	0.153	0.127	1.207	2.75
C-4	0.074	0.627	0.554	10.39	0.348	0.200	1.739	2.66
C-5	−0.045	0.518	0.563	14.45	0.279	0.169	1.649	2.88

Note 1

.

map, the GA is seen to exploited 64% of the potential excess returns earned by the omniscient trader, which is the highest of all the processes tested in this chapter. One of the possible reasons why the GA can work well with these nonlinear deterministic processes is that they are not pure nonlinear. C-1, C-2 and C-4 have linear AR(1) or AR(2) components. C-3, like the threshold processes, switches between two linear processes. As already evidenced in Section 6.4, the GA can handle these types of processes effectively. So, the success is not totally unanticipated.

However, the explanation above does not apply to C-5, which has no linear component. Nonetheless, statistics such as z_π, $\tilde{\pi}$ and p_w all indicate that this process is not as easy as the other four. For example, only 4% of the potential excess returns are exploited in this process. Regardless of these weaknesses, the fact that the GA can dominate B&H in this case motivates us to ask the following question: Can the GA work better for the pure nonlinear deterministic

processes than the respective stochastic ones, and hence can it help distinguish the chaotic processes from the stochastic processes? This is a question to pursue in the future.

11.6.6 Summary

The Monte Carlo simulation analysis conducted above provides us with an underpinning of the practical financial applications of the GA. It pinpoints the kinds of stochastic processes which we may like to see fruitful results. We have found that the GA can perform well with all kinds of stochastic processes which have a linear process (signal) as a part of them. Preliminary studies also suggest that it may also work well with chaotic processes. However, the class of nonlinear stochastic processes presents a severe limitation for the GA. In the next section, we shall see the empirical relevance of these results by actually applying OGA-based trading strategies to financial data.

11.7. Empirical analysis

11.7.1 Data Description and Analysis

The empirical counterpart of this chapter is based on two sets of high-frequency time series data regarding foreign exchange rates, namely, the Euro dollar vs. the U.S. dollar EUR/USD and the U.S. dollar vs. the Japanese yen USD/JPY.6 The data is from January 11, 2010 to April 17, 2010. Data within this period are further divided into 12 sub-periods with roughly equal numbers of observations. Table 32 gives the details. Let $P_{i,t}^{U}\left(P_{i,t}^{P}\right)$ denote the tth (t = 1, 2, . . . , ni) observation of the ith sub-period (i = A, B, . . . , L) of the EUR/USD (USD/JPY) forex series. The price series is transformed into the return series by the usual logarithmic formulation,

$$r_{i,t}^{j} = \ln\left(P_{i,t}^{j}\right) - \ln\left(P_{i,t-1}^{j}\right) \quad (305)$$

where j = U, P. Tables 33 and 34 give some basic statistics of the returns of each sub-period. Both return series share some common features. From Tables 33 and 34, the mean, median and skewness of these two return series are all close to zero. The kurtosis is much higher than 3, featuring the well-known fat-tail property. The Jarque-Bera (1980)[153] test further confirms that these forex returns do not follow the normal distribution, and that is true for each sub-period. In addition, the series is not independent due to its significant negative first-order serial correlation ρ1.

However, there is no evidence of serial correlation in higher orders.To apply what we learned from the Monte Carlo simulation to predict the effectiveness of theGAover these series, we must first gauge their likely stochastic processes. Here we follow a standard procedure frequently used in econometrics

Table 32: Data Quotations – EUR/USD and USD/JPY

Sub-Period	A	B	C	D	E	F
EUR/USD						
Number	12000	12000	12000	12000	12000	12000
From (GMT)	2/25 7:59	3/1 0:59	3/3 15:3	3/8 6:43	3/10 6:53	3/12 7:26
To (GMT)	2/26 8:22	3/2 7:17	3/5 3:04	3/9 1:08	3/11 7:12	3/15 1:16
Sub-Period	G	H	I	J	K	L
Number	12000	12000	12000	12000	12000	12000
From (GMT)	3/17 7:36	3/19 0:19	3/24 15:0	3/26 15:4	3/31 7:32	4/15 6:14
To (GMT)	3/18 6:12	3/22 2:01	3/26 2:12	3/30 6:23	4/02 1:14	4/17 0:37
Sub-Period	A	B	C	D	E	F
USD/JPY						
Number	12000	12000	12000	12000	12000	10808
From (GMT)	1/11 6:11	1/15 0:00	1/27 15:14	2/04 8:47	2/17 7:20	2/23 6:10
To (GMT)	1/14 8:11	1/21 0:00	2/03 3:24	2/11 2:43	2/23 6:09	2/26 21:4
Sub-Period	G	H	I	J	K	L
Number	12000	12000	12000	12000	12000	12000
From (GMT)	2/28 18:15	3/04 10:02	3/09 21:52	3/15 5:25	3/18 6:07	3/24 13:00
To (GMT)	3/04 10:01	3/09 21:52	3/15 1:21	3/18 6:06	3/24 13:00	3/30 10:41

(Chen & Lu, 1999)[154]. First, notice that all series used in our Monte Carlo simulation are stationary. To make sure that the forex returns are stationary, the Augmented Dickey-Fuller (ADF) test is applied (Dickey & Fuller, 1979)[155]. From Table 35, the null hypothesis that $r_{i,t}^{j}$ contains a unit root is rejected at the 1% significance level, meaning that the $r_{i,t}^{j}$ are stationary.

Second, since our Monte Carlo simulations demonstrate the effectiveness of the GA over the linear stochastic processes, it is important to know whether the forex returns have a linear component. To do so, the famous Rissanen's predictive stochastic complexity (PSC) as a linear filter is taken.Table 35 gives the ARMA(p, q) process extracted from the forex return series.AMA(1) linear process is founded for both forex returns in each sub-period. In fact, it re-confirms the early finding that the high-frequency forex returns follow a MA(1) process (Moody & Wu, 1997; Zhou, 1996)[156]. Third, it should be not surprising if none of these series is just linear. To see whether nonlinear dependence exists, one of the most frequently used statistics, the BDS test, is applied to the residuals filtered through the PSC filter.9 There Table 33.

Table 33 :Basic Statistics of the Return Series – EUR/USD

Sub-Period	A	B	C	D	E	F
Mean	–2.56E –07	–8.13E –07	–7.37E –07	5.39E –07	5.63E –07	–7.49E –07
Media	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Std. Dev.	0.000252	0.000252	0.000213	0.000191	0.000238	0.000264
Skewness	–0.01583	0.007214	–0.034436	0.002017	–0.001071	–0.009908
Kurtosis	5.606484	5.558600	5.636056	5.976148	6.136196	5.757020
Jarque-Bera	3397.10	3273.05	3476.48	4428.37	4917.45	3800.46
P-value	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ρ_1	–0.513935	–0.503725 I	–0.494695	–0.504014	–0.486925	–0.509612

Sub-Period	G	H	I	J	K	L
Mean	3.81E –07	8.00E –07	7.48E –07	5.64E –08	2.37E –07	–1.13E –06
Media	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Std. Dev.	0.000225	0.000217	0.000184	0.000241	0.000292	0.000219
Skewness	0.011155	–0.050369	–0.119412	0.007646	–0.021431	–0.203838
Kurtosis	6.512019	5.435495	6.226714	5.337107	8.780986	10.97326
Jarque-Bera	6166.88	2970.40	5233.92	2730.92	16708.03	31861.55
P-value	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ρ_1	–0.493223	–0.505528	–0.480500	–0.498232	–0.475452	–0.464571

Note: ρ_1 is the first-order autocorrelation coefficient. Jarque-Bera statistic converges to a chi-square distribution with two degrees of freedom under the normality assumption.

Table 34: Statistics of the Return Series – USD/JPY

Sub-Period	A	B	C	D	E	F
Mean	3.97E –07	–5.16E –07	–2.01E –06	2.54E –07	1.69E –06	–1.44E –06
Media	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Std. Dev.	0.000413	0.002108	0.001853	0.000332	0.000311	0.000363
Skewness	0.008135	0.080038	–0.018340	–0.057694	0.022959 –	–0.003358
Kurtosis	6.769064	6.711594	6.854310	7.170642	6.757800	6.374525
Jarque-Bera	7091.806	6898.478	7426.049	8700.883	7059.230	5123.885
P-value	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ρ_1	–0.343317	–0.338790	–0.370748	–0.362052	–0.360786	–0.335953

Sub-Period	G	H	I	J	K	L
K L						
Mean	2.53E –06	–1.09E –06	–2.54E –06	–2.75E –07	–7.87E –07	1.90E –06
Media	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Std. Dev.	0.000301	0.000279	0.000322	0.000287	0.000265	0.000247
Skewness	0.080100	0.019734	0.079313 0	0.002414	–0.019244	0.213584
Kurtosis	5.597214	6.763973	6.747828	8.198238	7.650768	6.701801
Jarque-Bera	3385.029	7083.936	6459.934	13508.60	10811.96	6941.746
P-value	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ρ_1	–0.436860	–0.396329	–0.344660	–0.348622	–0.361993	–0.364189

Note: ρ_1 is the first-order autocorrelation coefficient. Jarque-Bera statistic converges to a chi-square distribution with two degrees of freedom under the normality assumption

Table 35: . Basic Econometric Properties of the Return Series – EUR/USD and USD/JPY.

Sub-Period	A	B	C	D	E	F
EUR/USD						
ADF	–74.9502	–76.4264	–74.0755	–76.6226	–77.4292	–79.1714
Critical Value	–3.4341	–3.4341	–3.4341	–3.4341	–3.4341	–3.4341
PSC	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)
Sub-Period	G	H	I	J	K	L
ADF	–74.7427	–74.7053	–68.8254	–73.4958	–72.3726	–67.6148
Critical Value	–3.4341	–3.4341	–3.4341	–3.4341	–3.4341	–3.4341
PSC	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)
Sub-Period	A	B	C	D	E	F
USD/JPY						
ADF	–57.1573	–55.2394	–56.0518	–56.8433	–55.0202	–51.1507
Critical Value	–2.5660	–2.5660	–2.5660	–2.5660	–2.5660	–2.5660
PSC	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)
Sub-Period	G	H	I	J	K	L
ADF	–59.3422	–57.4123	–55.5809	–58.0822	–57.5485	–59.5623
Critical Value	–3.4341	–3.4341	–3.4341	–3.4341	–3.4341	–3.4341
PSC	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)

Note: The “Critical Value” indicates the critical value of the ADF test that is taken from the table provided by Dickey and Fuller at the 1% significance level.

are two parameters used to conduct the BDS test. One is the distance measure (ε standard deviations), and the other is the embedding dimension. The parameter “ε” considered here is equal to one standard deviation. (In fact, other are also tried, but the results are not sensitive to the choice of ε.) The embedding dimensions considered range from 2 to 5. Following Barnett et al. (1997)[157], if the absolute values of all BDS statistics under various embedding dimensions are greater than 1.96, the null hypothesis of an identical independent distribution (IID) is rejected. From Table 36, the BDS statistics for the EUR/USD and USD/JPY are all large enough to reject the null hypothesis, i.e. nonlinear dependence is detected.

Fourth, given the existence of the nonlinear dependence, the next step is to identify its possible form, i.e. by modelling nonlinearity. While there is no standard answer as to howthis can be done, the voluminous(G)ARCHliterature over the past two decades has proposed a second-moment connection (Bollerslev et al., 1992)[158]. In order to see whether (G)ARCH can successfully capture nonlinear signals, we Table 36. The BDS Test of the PSC-filtered Return Series – EUR/USD and USD/JPY.

Table 36: The BDS Test of the PSC-filtered Return Series – EUR/USD and USD/JPY.

Sub-Period	A		B		C		D		E		F	
Part	I	II	I	II	I	II	I	II	I	II	I	II
EUR/USD												
DIM = 2	20.47	26.82	22.58	26.56	13.60	20.25	17.15	14.66	18.23	18.09	18.03	19.37
DIM = 3	27.57	34.17	30.61	34.72	19.44	26.84	22.50	20.12	22.78	23.48	24.63	26.43
DIM = 4	33.60	40.03	37.25	40.81	23.80	31.27	26.80	24.22	25.68	27.63	30.21	32.09
DIM = 5	38.50	45.80	43.40	46.75	27.43	35.23	30.38	27.40	28.54	31.23	35.26	37.94

Sub-Period	G		H		I		J		K		L	
	I	II	I	II	I	II	I	II	I	II	I	II
DIM = 2	12.04	16.97	23.90	19.45	13.06	12.40	20.13	13.41	35.69	19.74	8.18	22.23
DIM = 3	17.84	22.20	30.02	25.59	17.30	17.31	26.84	18.79	46.83	24.39	10.98	27.08
DIM = 4	21.09	26.34	34.39	30.41	20.35	20.57	31.24	22.98	56.42	27.22	12.97	30.22
DIM = 5	24.08	30.18	39.31	35.47	23.29	23.40	35.39	26.48	66.58	29.79	14.20	33.13

Sub-Period	A		B		C		D		E		F	
	I	II	I	II	I	II	I	II	I	II	I	II
USD/JPY												
DIM = 2	15.36	23.15	15.68	13.41	12.00	16.63	14.76	20.44	12.98	17.84	17.88	16.61
DIM = 3	17.89	28.38	18.83	16.04	14.54	20.02	17.11	23.15	16.08	20.87	21.35	18.94
DIM = 4	20.03	31.37	20.17	17.89	15.32	22.24	18.72	24.27	17.49	22.82	23.35	20.44
DIM = 5	22.30	34.58	21.57	19.13	16.07	24.42	20.28	25.43	18.52	24.56	24.43	22.16

Sub-Period	G		H		I		J		K		L	
	I	II	I	II	I	II	I	II	I	II	I	II
DIM = 2	15.65	11.34	15.56	16.84	16.44	15.51	20.98	17.79	19.41	15.51	15.28	15.61
DIM = 3	17.64	13.92	18.57	18.91	18.50	18.68	25.07	21.84	21.94	16.84	16.32	17.87
DIM = 4	19.30	15.35	20.86	19.45	19.78	21.02	27.72	24.43	23.23	17.52	17.21	19.34
DIM = 5	20.82	16.49	23.10	19.73	20.95	22.76	30.10	26.45	24.15	18.56	18.14	20.62

carry out the Lagrange Multiplier (LM) test for the presence of ARCH effects. The LM test for ARCH effects is a test based on the following model:

$$\sigma_t^2 = h\left(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_p \varepsilon_{t-p}^2\right) \quad (306)$$

where h is a differential function. The null hypothesis that the ARCH effect does not exist is

$$\alpha_1 = ... = \alpha_p \quad (307)$$

By taking p = 1, 2, . . . , 4, the LM test results are given in Table 37. It is found that the ARCH effect does exist in both return series.

Table 37: The LM Test of the ARCH Effect in the Return Series – EUR/USD and USD/JPY.

Sub-Period	A	B	C	D	E	F
EUR/USD						
p = 1	1029.94	821.665	681.92	560.27	463.98	401.08
p = 2	1572.34	1191.26	998.22	1094.7	960.83	585.88
p = 3	2030.32	1501.74	1202.15	1320.58	1052.5	705.17
p = 4	2169.98	1731.33	1295.77	1471.40	1195.9	871.73

Sub-Period	G	H	I	J	K	L
p = 1	275.07	797.26 4	411.61	390.94	1584.30	1571.04
p = 2	423.33	1168.19	689.02 5	553.11	1668.88	1587.53
p = 3	493.11	1262.87	1001.22	678.90	1714.39	1640.60
p = 4	551.99	1354.28	1050.53	715.68	2036.42	1641.41

Sub-Period	A	B	C	D	E	F
USD/JPY						
p = 1	533.15	411.35	479.80 7	769.49	550.15	685.34
p = 2	639.75	490.58	6018.02	849.31	604.18	752.71
p = 3	677.49	531.78	667.50	854.11	614.26	821.85
p = 4	709.00	559.97	687.09	923.01	636.99	854.71

Sub-Period	G	H	I	J	K	L
p = 1	600.528	545.791	696.185	749.650	883.107	795.762
p = 2	648.101	656.653	758.918	1094.82	926.127	929.618
p = 3	695.639	727.043	811.000	1101.78	939.221	1059.00
p = 4	726.942	764.836	844.766	1103.08	951.489	1109.23

Note: The LM test is asymptotically distributed as $\chi 2$ with p degrees of freedom when the null hypothesis is true. There is no need to report the p values here because they are all 0.0000.

After these series of statistical tests, we may conclude that basically both the EUR/USD and the USD/JPY return series have MA(1) as a linear component and ARCH as a part of its nonlinear components. In Section 11.6.3, the Monte Carlo simulation analysis already indicated that the GA can work well with MA(1) plus (G)ARCH processes. To see the empirical relevance of the simulation study, in the next sections, the GA is applied to the two return series.

11.7.2 Experimental Design

In order to compare the empirical results with our earlier simulation analysis, the experiments are designed in a similar fashion to the one which our Monte Carlo simulation follows. Specifically, many “ensembles” are generated from the original series to evaluate the performance of the GA. Of course, rigorously speaking, they are not the “ensembles” defined in the stochastic process. They are just subseries taken from the original return series. Each subseries has 105

observations. The first 70 observations are treated as the training sample, and the last 35 observations are used as the testing sample.

Nonetheless, to make the tests we developed in Section 11.4 applicable, we cannot just continuously chop the return series into subseries, because doing so will not make the sampling process independent, and hence will violate the fundamental assumption required for the central limit theorem. One solution to this problem is to leave an interval between any two consecutive subseries so that they are not immediately connected. The purpose in doing this is hopefully to make them independent of each other as if they were sampled independently. However, how large an interval would suffice? To answer this question, we take a subsequence with a fixed number of lags, say, $\{r_{i,t}^j, r_{i,t+k}^j, r_{i,t+2k}^j\}$ from the original return series, where k varies from 40, 60, . . . , to 300.We then apply the BDS test to each of these subsequences.

Table 38 summarizes the BDS test results. For the EUR/USD case, it is found that when k is greater than 100, the null hypothesis that the subsequence $\{r_{i,t}^j, r_{i,t+k}^j, r_{i,t+2k}^j\}$is IID is not rejected. In other words, leaving an interval of 100 observations between each of two consecutive subseries would suffice. For the EUR/USD case, k can even be smaller than 60. To ensure the quality of the sampling process, we, however, take an even larger number of lags, i.e. k = 200.

This choice leaves us with a total of 720 subseries from the EUR/USD and 709 subseries from the USD/JPY. The GA is then employed to extract trading strategies from the training samples of these subseries, and the strategies extracted are further applied to the respective testing samples. The resulting accumulated returns (┌) are then compared with that of the B&H strategy.

11.7.3 Results of the Experiments

Since the analysis of the data shows that the two forex returns are mixtures of MA(1) and (G)ARCH processes, our previous results of Monte Carlo simulations may provide a good reference for what one can expect from such empirical applications. Both Tables 38 and 39 indicate the superior performance of the GA over B&H, except in relation to the criterion for the luck coefficient, when the underlying stochastic processes areMAplus (G)ARCH.Will the dominance carry over?

Table 38: The BDS Test of the Lag Period in the Return Series – EUR/USD and USD/JPY.

Lag	DIM=2	DIM=3	DIM=4	DIM=5
EUR/USD				
40	2.94	3.45	3.86	4.18
60	0.72	1.20	1.27	1.38
80	1.11	1.21	1.38	1.50
100	0.66	0.66	0.69	0.69
120	0.61	0.66	0.79	0.88
140	0.45	0.52	0.54	0.58
160	0.30	0.43	0.46	0.54
180	0.21	0.30	0.42	0.49
200	–0.01	0.08	0.12	0.11
220	0.11	0.14	0.13	0.13
240	0.25	0.24	0.27	0.24
260	–0.02	–0.04	–0.04	–0.01
280	0.10	0.11	0.14	0.14
300	0.06	0.07	0.05	0.01
USD/JPY				
40	1.39	1.50	1.50	1.57
60	0.53	0.69	0.75	0.89
80	0.56	0.63	0.72	0.80
100	–0.08	–0.12	–0.12	–0.16
120	0.13	0.22	0.19	0.20
140	0.01	–0.13	–0.14	–0.09
160	0.05	0.09	0.09	0.12
180	–0.01	–0.07	0.01	0.06
200	–0.04	–0.08	–0.08	–0.06
220	0.21	0.29	0.30	0.32
240	0.15	0.13	0.11	0.12
260	0.05	0.12	0.09	0.07
280	–0.14	–0.09	–0.11	–0.10
300	0.06	0.02	0.05	0.04

Note: The BDS statistic follows an asymptotically standard normal distribution.

Table 39 is the kind of table which we have presented many times in Section 11.6. All the key statistics z_π, z_w, and z_d are consistent with those of AH-7 (Table 28) and GH-10 (Table 29). So, in both forex return series, the dominance of the GA over B&H is statistically significant. The consistency continues even to a finer level of the results: $\bar{\pi}^1 < 1$ and $\bar{\pi}^2 > 1$. As already seen,

B&H earned negative profits in both of the cases AH-7 and GH-10, while the GA earned positive profits in both cases. In addition, both the winning probability and the exploitation ratio are also Table 37.

Table 39: . Performance Statistics of the OGA and B&H – EUR/USD and USD/JPY

	$\bar{\pi}^1$	$\bar{\pi}^2$	$\bar{\pi}^*$	z_w	$\bar{\pi}(\%)$	$\dot{\pi}(\%)$	p_w	z_w
EUR/USD	0.9999	1.0012	1.0028	38.58	43	9257	0.972	77.10
USD/JPY	0.9999	1.0010	1.0039	23.70	27	11462	0.850	26.17

	$\hat{s}_1$	$\hat{s}_2$	$\hat{d}$	z_d	$\gamma^1_{0.05}$	$\gamma^2_{0.05}$	$\tau_{0.05}$	z_r
EUR/USD	–0.0338	1.4193	1.4532	18.32	0.0812	0.0933	0.8710	–1.69
USD/JPY	–0.0086	0.8786	0.8873	20.64	0.0826	0.0948	0.8713	–1.66

Note: $\bar{\pi}^1$, $\bar{\pi}^2$ and $\bar{\pi}^*$ are the respective sample mean return of OGA, B&H and the omniscient trader. $\tilde{\pi}$ is the exploitation ratio (Eq. (285)), and $\dot{\pi}$ is the relative superiority index (Eq. (286)). p_w is the sample winning probability of OGA over B&H (Eq. (287)). $\hat{s}_1$ and $\hat{s}_2$ are the corresponding sample Sharpe ratio of OGA and B&H (Eq. (290)). Their sample difference is $\hat{d}$ (Eq. (291)). $\hat{l}^1_{0.05}$ land $\hat{l}^2_{0.05}$ are the sample luck coefficient of OGA and B&H (Eq. (298)), and $\hat{r}_{0.05}$ is the sample luck coefficient ratio between the two (Eq. (299)). The z_π, z_w, z_d and z_r are the test statistics of the mean return difference, winning probability, Sharpe ratio differential, and luck coefficient ratio, respectively. The critical value of them is 1.28 at the 10% significance level, and is 1.64 at the 5% significance level.

comparable. p_w is around 95% for both AH-7 and GH-10, and $\tilde{\pi}$ is about 25%. The value of p_w remains as high for the EUR/USD series, while it drops a little to 85% for the USD/JPY series. As to $\tilde{\pi}$, it is also about 25% for the USD/JPY series, but is greater than 40% for the EUR/USD series.

Notice that our earlier simulation result already indicated that, for some reason unknown to us, the MA component when combined with the ARCH or GARCH component may bring a negative impact to the luck coefficient. This has been already shown in the cases AH-7 and GH-10. What interests us here is that this observation repeats itself in our empirical results. The statistic z_r is statistically negative in both return series. As a result, to a large extent, what we have found from the early Monte Carlo simulations applies quite well to the real data. Hence, theGA can be useful in extracting information to develop trading strategies involving these high-frequency financial data because the underlying stochastic process, based on the Monte Carlo simulation analysis, is not a hard one for the GA.

11.8. Concluding remarks

The literature on financial data mining, driven by the rapid development and applications of computational intelligence tools, are frequently clothed with a “magic house” notoriety. Unlike in mainstream econometrics, users are usually not well informed of the stochastic properties of these tools, which in turn makes it difficult to grasp the significance of the result obtained from one specific application, be it positive or negative. An essential question is how we can know that what happens in one specific application can or cannot extend to the other one. Will we still be so “lucky” next time?

By using the Monte Carlo simulation methodology, a statistical foundation for using the GA in market-timing strategies is initiated. This foundation would allow us to evaluate how likely the GA will work given a time series whose underlying stochastic process is known. This helps us to distinguish the luck from normal expectations. We believe that this is a major step toward lightening the black box. We emphasize that this work provides a statistical foundation, not the statistical foundation, because there are many other ways of enriching the current framework and of making it more empirically relevant.

First, different benchmarks may replace the B&H strategy. This is particularly so given a series of articles showing that simple technical analysis can beat B&H. However, since we can never run out of interesting benchmarks, the exploitation ratio $\widetilde{\pi}$ introduced in this PhD thesis will always be a good reference. For example, in this chapter, we can hardly have a $\widetilde{\pi}$ of 30% or higher. Consequently, the 70% left there may motivate us to try more advanced version of the GA or different computational intelligence algorithms.

Second, financial time series are not just restricted to the six stochastic processes considered in this chapter, but introducing new stochastic processes causes no problems for the current framework. Third, different motivations may define different evaluation criteria. The four criteria used in this chapter are by no means exhausted. For example, the downside risk or VaR (Value at Risk) frequently used in current risk management can be another interesting criterion. However, again, it is straightforward to add more criteria to the current framework as long as one is not bothered by deriving the corresponding statistical tests. Fourth, the focus of this chapter is to initiate a statistical foundation. Little has been addressed regarding the practical trading behavior or constraints. Things like transaction costs, non-synchronous trading, etc., can be introduced to this framework quite easily. Fifth, our framework is also not restricted to just the ordinary GA, for the general methodology applies to other machine learning tools, including the more advanced versions of the GA.

Finally, while, in this PhD Thesis, we are only interested in the statistical foundation, we do not exclude the possibilities of having other foundations. As a matter of fact, we believe that a firm statistical foundation can show us where to ask the crucial questions, and that will help build a more general mathematical foundation. For example, in this chapter, we have been already well motivated by the question as to why the GA performed quite poorly in the pure nonlinear

stochastic processes, but may need more work before coming to its maturity. However, the point here is that theoretical questions regarding the GA's performance cannot be meaningfully answered unless we have firmly grasped their behavior in a statistical way.
This chapter provides a comprehensive example on how to tackle an investment problems using several techniques.

12. Using GA's + ANN's and SVM for financial time series forecasting – 2 applications

In this PhD thesis, a new approach for time series forecasting is presented. The forecasting activity results from the interaction of a population of experts, each integrating genetic and neural technologies. An expert of this kind embodies a genetic classifier designed to control the activation of a feedforward artificial neural network for performing a locally scoped forecasting activity.
Genetic and neural components are supplied with different information: The former deal with inputs encoding information retrieved from technical analysis, whereas the latter process other relevant inputs, in particular past stock prices. To investigate the performance of the proposed approach in response to real data, a stock market forecasting system has been implemented and tested on two stock market indexes, allowing for account realistic trading commissions.
In this thesis, we present a hybrid approach to stock market forecasting that integrates both GAs and ANNs and cooperatively exploits them to forecast the next-day price of stock market indexes. In particular, we use an extended classifier system (XCS), which relies on this first application in technical-analysis indicators to determine the current market status, in conjunction with feedforward ANNs explicitly designed for financial time series prediction. We have called the proposed approach NXCS, standing for neural XCS, and customized it for financial time series prediction. A forecasting system based on an NXCS module has been tested on financial time series showing the trend of some major stock market indexes on a fairly large observational window.
In particular, about 72 years of data of S&P500 stock market indexes have been used to train and test the system, being very careful to avoid any form of data snooping. In both cases, the first 1000 data were used to train and tune the NXCS module, and the resulting system was tested on the subsequent 250 data, leaving its overall configuration unchanged.
We compared the system's forecasting capabilities with the ''Buy and Hold'' (B&H) strategy, considering realistic transaction costs. The results are encouraging, demonstrating the validity of the approach
In recent years, advances in both analytical and computational methods have led to a number of interesting new approaches to financial time series forecasting, based on non-linear and non-stationary models. In the following, we focus the review of previous work on GAs and ANNs applied to stock market prediction, as our proposal is based on the integration of such techniques. In addition, as the resulting framework is also an implementation of the general concepts known as mixture of experts, the related work on this topic will be briefly described.

12.1. GAs for financial time series forecasting

GAs are a family of computational models inspired by natural evolution [159],[160]as we have seen before In a broader usage of the term, a genetic algorithm is any population based model that uses selection and recombination operators to generate new sample points in a search space. An implementation of a GA deals with a population of ‘‘chromosomes’’, each representing a potential solution of a target problem, usually in form of binary strings. In particular, for classification tasks, a chromosome can be represented by a condition–action pair, where the condition is a string of values in {0,1,#} and the action is a label that denotes a class. Thus, each chromosome is specialized on a subset of the input space; i.e., some inputs activate the chromosome and other inputs exclude it from the decision process.

Usually, these chromosomes are randomly created, and undergo reproductive opportunities in such a way that better solutions are given more chances to reproduce than poorer ones. Although GAs have been adopted in a multitude of different tasks, in this chapter we focus on proposals that address only the problem of financial time series forecasting. Noever and Baskaran [161]investigated the task of predicting trends and prices in financial time series, making experiments on the S&P500 stock market. Mahfoud and Mani [162] addressed the general problem of predicting future performances of individual stocks.

Their work is particularly relevant, as they make a comparison between GAs and ANNs applied to financial forecasting. According to their experiments—repeated on several stock markets—both approaches outperform the B&H strategy. A combined approach, obtained by averaging out GAs and ANNs outputs, is also experimented with positive results. GAs have also been used in a variety of hybrid approaches to financial time series prediction. For example, Muhammad and King [163] devised evolutionary fuzzy networks to forecast the foreign exchange market, whereas Kai and Wenhua [164] exploited GAs to train ANNs for predicting a stock price index.

12.1.1 ANNs for financial time series forecasting

ANNs appear to be particularly suited for financial time series forecasting, as they can learn highly non-linear models, have effective learning algorithms, can handle noisy data, and can use inputs of different kinds (see [165] for a survey). Furthermore, complex non-linear models based on exponential GARCH processes [166] show similar results (in terms of out-of-sample prediction performance) to those obtained by much simpler ANNs based on multilayer perceptron (MLP) architectures [167]. A major weakness of MLPs, from a time-series forecasting perspective, is the absence of an internal state, making it difficult to capture the dynamics of the given data series. Even if standard MLPs can still be used [168], due to their simplicity and flexibility, a variety of more complex architectures with some form of internal memory has been proposed (see [169] for a survey). In particular, Recurrent ANNs (RANNs) have proved capable of outperforming stateless

architectures in financial time series forecasting [170]. Nevertheless, it is well known [171] that out-of-sample results have a strong dependence on the time period used to train the network. In other words, ANNs trained on data belonging to a specific period perform reasonably well only if the test period is relatively similar to the one used for training.
Furthermore, the similarity between two periods of time is guaranteed by sharing some kind of economic condition (e.g., bullish or bearish period), instead of being caused by temporal contiguity. This problem is related to the presence of the so-called regime-shifting phenomenon, which occurs when the underlying process can be thought of as being multistationary. In this case, several periods of stationarity (i.e., regimes) hold, separated by usually rapid transitions. Under this hypothesis, obtaining a single model that holds for different regimes can be extremely difficult.

12.1.2 Multiple experts for financial time series forecasting

As previously pointed out, GAs deal with a population of chromosomes, each specialized on a subset of the input space. Hence, nothing prevents us from considering them as an implementation of the multiple experts concept, which received a great deal of attention—though in forms substantially different from the GA proposal—from both the connectionist and the time series communities. As we believe that a multiple experts centered perspective could be useful for dealing with the problem of regime-shifting, below we also briefly recall some related work in this specific area. Note that, in this case, the problem of how to combine experts outputs in order to obtain a response from the overall system becomes a crucial aspect. In the time series community, Weigend et al.—starting from Jacobs and Jordan's mixtures of experts [172]— devised non-linear gated experts [173],[174] applying them to several time series domains, including financial ones. The key elements characterizing these experts are the non-linearity of local experts and gates, the experts' capability of separately adapting themselves to noisy data, and a soft-partitioning mechanism (i.e., softmax [175]) for blending outputs. Moreover, Shi and Weigend [176]proposed a statistical method for experts selection based on Hidden Markov models.

12.2. A hybrid approach for dealing with stock market forecasting

In this section, the hybrid approach previously summarized is described with more detail, from both a conceptual and a technical perspective. As for conceptual issues, a novel kind of model identification is introduced, originated by the need of dealing with multistationary processes. In addition, the idea of partitioning the input space starting from suitable technical-analysis domain knowledge is illustrated and framed in a multiple-experts perspective. To this end, the underlying framework is briefly introduced, to give the reader the ability to perceive the general characteristics of the proposed approach. As for technical issues, some basic notions about the particular kind of GAs (i.e., XCSs) adopted for evolving a population of hybrid experts is given.

Finally, the particular kind of ANN that has been designed and adopted is described, with all customizations devised to deal with the task of financial time series forecasting.

12.2.1 Context-based identification of multistationary models

The basic idea that lies behind the proposed approach is simple: rather than trying to identify a global model, an attempt to identify different local models is performed, according to the hypothesis that financial time series are multistationary.
Let us point out in advance that, no matter whether the task is prediction or classification, our approach to model identification differs very much from the ones based on a state description — e.g., Hidden Markov models [177] and RANNs. To stress the difference with existing approaches and to give a flavor of the underlying assumptions, we decided to call ‘‘context-based’’ our approach to model identification. In fact, it applies the divide-and-conquer strategy by first identifying the current context and then using one or more local experts acknowledged as able to deal with it. Here, different sources of information are entrusted with different tasks, i.e., performing the identification of the current context vs. performing the intended classification or prediction. As a particular case of context-based model identification, let us consider a population of experts, each embedding two components: a context selector and a classifier/predictor. The selector is aimed at controlling the activation of the associated classifier/predictor, depending on the current input. In the next session, we briefly outline the corresponding framework, named guarded experts.

12.2.2 The guarded experts framework

Let us assume that an input and an output space exist (i.e., I and O, respectively), and that an oracle K holds that, for each $\boldsymbol{x} \in \boldsymbol{I}$, maps x to a value in O. Given an input vector, a globally scoped expert $\widetilde{\boldsymbol{H}}$ is a total function that approximates Λ on the whole input space, whereas a guarded expert $\widetilde{\Gamma}$ is a partial function that approximates Λ on a proper subset of the input space. In its simplest form, $\widetilde{\Gamma}$ can be represented by a pair $\left\langle \boldsymbol{g}, \widetilde{\boldsymbol{h}} \right\rangle$, where g is a contextselector (i.e., a guard) devoted to accept or reject any given input x as belonging to the domain of the function $\widetilde{\boldsymbol{h}}$.
. Without loss of generality, let us separate the inputs used for performing context selection from the ones used for making a prediction or classification. Thus, $\boldsymbol{I} \equiv \boldsymbol{I}_g \boldsymbol{I}_h$, where $\boldsymbol{g} : \boldsymbol{I}_g \rightarrow \boldsymbol{B}$ maps a non-empty subset of input features to $\boldsymbol{B} = \{\boldsymbol{false}, \boldsymbol{true}\}$, and $\widetilde{\boldsymbol{h}} : \boldsymbol{I}_h \rightarrow \boldsymbol{O}$ approximates the oracle on the subset of inputs for which the property g holds. Applying a guarded expert $\widetilde{\Gamma} = \left\langle \boldsymbol{g}, \widetilde{\boldsymbol{h}} \right\rangle$ to an input $\boldsymbol{x} \equiv \boldsymbol{x}_g \boldsymbol{x}_h$ returna a value that is described by the following semi-fomal semantics:

$$\widetilde{\Gamma}(\boldsymbol{x}) = \boldsymbol{if}\ \boldsymbol{g}(\boldsymbol{x}_g)\ \text{then}\ \widetilde{\boldsymbol{h}}(\boldsymbol{x}_h)\ \text{else}\ \perp \quad (308)$$

namely, given a guarded expert $\widetilde{\Gamma} = \langle \boldsymbol{g}, \widetilde{\boldsymbol{h}} \rangle$, the classifier/predictor $\widetilde{\boldsymbol{h}}$ is activated by an input $\boldsymbol{x} \equiv \boldsymbol{x}_g \boldsymbol{x}_h$ only if $\boldsymbol{g}(\boldsymbol{x}_g) = \boldsymbol{true}$ (i.e., if the input x matches the guard g). The subset of inputs covered by a guarded expert $\widetilde{\Gamma} = \langle \boldsymbol{g}, \widetilde{\boldsymbol{h}} \rangle$ hi strictly depends on its guard g, according to the definition:

$$\boldsymbol{Covers}(\Gamma) = \left\{ \boldsymbol{x}_g \boldsymbol{x}_h \in \boldsymbol{I} \middle| \boldsymbol{g}(\boldsymbol{x}_g) = \boldsymbol{true} \right\} \quad (309)$$

Let us now briefly depict the characteristics of a population Ω of guarded experts, starting from the following definition:

$$\Omega = \left\{ \widetilde{\Gamma}_i \middle| \widetilde{\Gamma}_i = \langle \boldsymbol{g}_i, \widetilde{\boldsymbol{h}}_i \rangle, \boldsymbol{i} = 0,1,...\boldsymbol{n} \right\} \quad (310)$$

The set of inputs covered by X can be defined on top of Eq. (2) as follows:

$$\boldsymbol{Covers}(\Omega) = \bigcup_{\widetilde{r}_i \in \Omega} \boldsymbol{Covers}(\widetilde{\Gamma}_i) \quad (311)$$

Furthermore, given an input $\boldsymbol{x} \equiv \boldsymbol{x}_g \boldsymbol{x}_h$ and a population of guarded experts Ω, the function $\boldsymbol{Select}(\Omega | \boldsymbol{x})$ is responsible for the match-set formation; in other words, it returns the set of guarded experts whose guards are matched by x. In symbols:

$$\boldsymbol{Select}(\Omega | \boldsymbol{x}) = \boldsymbol{Select}(\Omega | \boldsymbol{x}_g \boldsymbol{x}_h) = \left\{ \Gamma = \langle \boldsymbol{g}, \boldsymbol{h} \rangle \in \Omega \middle| \boldsymbol{g}(\boldsymbol{x}_g) = \boldsymbol{true} \right\} \quad (312)$$

We are interested in investigating a particular class of populations of experts, able to cover the whole input space. We say that a population Ω of guarded experts covers the whole input space I when the following constraint holds:

$$\boldsymbol{Covers}(\Omega) \equiv \boldsymbol{I} \quad (313)$$

This property allows us to guarantee that a population of guarded experts can substitute a globally scoped expert on the whole input space. In the following, we assume that such a property holds; i.e. that a non-empty match set can be formed for any input $\boldsymbol{x} \in \boldsymbol{I}$:

$$\forall \boldsymbol{x} \in \boldsymbol{I} : \left| \boldsymbol{Select}(\Omega | \boldsymbol{x}) \right| > 0 \quad (314)$$

Note that the definitions given by Eqs. (182) and (183) do not rule out the case of multiple experts being selected for dealing with a given input x. Without going into unnecessary details, let us assume that a voting policy or an outputsblending mechanism (for classification and prediction tasks, respectively) should be supplied for evaluating the overall response of the experts enabled by the current input. As for classification tasks, the most widely acknowledged voting policy is the so-called majority rule, which consists of adopting the response that reaches the highest score among the selected experts. As for prediction tasks, the most widely acknowledged outputs-

blending mechanism is the average rule, which averages out the outputs of the selected experts. A‘‘weighted’’ version of both the majority and average rule can be enforced by taking into account each single expert according to a different degree of reliability (e.g., its degree of adaptation to the given environment).

12.2.3 Neural XCS

Bearing in mind that here the term GAs is used in a broad sense, denoting systems where a population of classifiers evolve according to Darwinian selection, we decided to implement the guarded experts framework by hybridizing Wilson's XCSs with feedforward ANNs explicitly designed for financial time series prediction. The resulting approach has been called NXCS, standing for Neural XCS. As for XCSs (briefly outlined in Appendix A), they have been devised by Wilson based on Holland's Learning Classifier Systems [178]. Since then, further related work has provided a deep theoretical understanding of the XCS approach [179,180,181] and many extensions have been proposed [182,183,184,185]. For the sake of simplicity, in the following we restrict our attention to the prediction task only, being interested in investigating how NXCS experts can be customized for financial time series forecasting. To relate NXCS with the general framework of guarded experts, let us point out that XCS classifiers play the role of guards, whereas ANNs are used for implementing predictors. In particular, given an input vector $\boldsymbol{x} \in \boldsymbol{I}$, let us separate some of its binary $(\boldsymbol{x}_b)$ features from the rest of the input vector $(\boldsymbol{x}_r)$, yielding $\boldsymbol{x} \equiv \boldsymbol{x}_b \boldsymbol{x}_r$. Thus, the guard of an NXCS expert $\widetilde{\Gamma} = \langle \boldsymbol{g}, \widetilde{\boldsymbol{h}} \rangle$ maps any binary input to {false, true}, whereas its predictor $\widetilde{\boldsymbol{h}}$ is an ANN trained and activated only on a subset of the input space (according to the selection performed by the guard g). Furthermore, let us assume that an NXCS expert can have multiple outputs. It is worth pointing out that an NXCS expert can be described from both an evolutionary and a neural perspective. According to the former interpretation, it can be considered an extension of an XCS classifier, whose action part has been replaced with a suitable ANN. According to the latter interpretation, it can be considered an extension of a neural predictor, equipped with an additional guard that acts as a selector, thus preventing the neural predictor from working outside the context defined by the guard itself. In any case, NXCSs are able to perform a synergistic integration between two sources of information (i.e., binary and numerical), adopting a suitable technology for each one. Fig. 1 shows the structure of a generic NXCS expert, pointing to its guard and the corresponding predictor (the capability of the genetic guard to enable the neural predictor is also put into evidence).

12.2.4 Handling a population of NXCS experts

An NXCS is an evolutionary system where a population of NXCS experts, each characterized by a genetic guard and the corresponding neural predictor, is raised in a typical XCS-like environment.

Figure 69: The structure of an NXCS expert

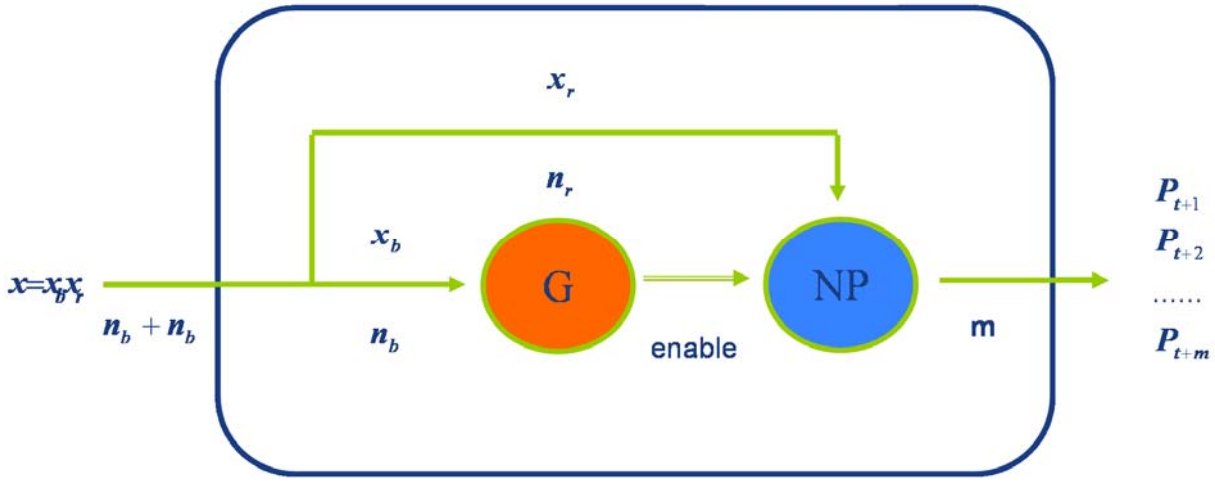


Fig. 56. The structure of an NXCS expert, which embeds a genetic guard and a neural predictor.

Each NXCS expert has a set of (neural predictor)associated parameters, to be used and updated while evaluating, step-by-step, the evolution of the population. In particular, for each NXCS expert $\widetilde{\Gamma}$, let usrecall the following parameters:

• the fitness (f), i.e., the degree of adaptation of $\widetilde{\Gamma}$ to the environmental constraints;• the prediction or strength (p), i.e., the expected reward to the system for using $\widetilde{\Gamma}$.

Given a population of NXCS experts, its dynamics is basically conformant to the one that characterizes XCS classifier systems. Of course, some differences exist, due to the fact that the action part of an XCS classifier has been replaced with an ANN. The remainder of this subsection is devoted to illustrate the major differences that hold between an XCS and an NXCS. In particular, the mechanisms for (i) generating and maintaining experts, as well as for (ii) performing experts' selection and outputs blending will be briefly described.

12.3. Generating and maintaining NXCS experts

Given an NXCS expert, its genetic guard supports covering, crossover, and mutation operations, whereas its neural predictor is trained once, immediately after being created, and left unaltered while the system is running. The fitness f of an NXCS expert is updated according to the default XCS policy. On the other hand, the prediction p is dealt with in a non-standard way: Having to cope with prediction tasks, we let p coincide with th primary ANN output.

Furthermore, when a new NXCS expert has to be inserted into the population, its predictor is generated with initial random weights and then trained using a set of examples obtained by selecting the inputs that match its guard on a fixed-length window of past prices. In this way, each

predictor is trained on a different set of examples, according to the current time and to the corresponding guard.

12.3.1 NXCS mechanisms for experts selection and outputs blending

In an XCS classifier system, experts' selection and outputs combination occur as follows (see also Appendix A). Given a message, first the match set (M) is created gathering all XCS classifiers whose condition is matched by the current input, then the adopted voting policy is applied. The default voting policy is the so-called fitness-weighted majority rule. Being $\boldsymbol{M_i} \subseteq \boldsymbol{M}$ the set of XCS classifiers that support the action i, the reward the system can achieve for choosing it (i.e., Pi) is estimated according to the formula:

$$\boldsymbol{P_i} = \frac{\sum_{c \in M_i} \boldsymbol{p_c} \cdot \boldsymbol{f_c}}{\sum_{c \in M_i} \boldsymbol{f_c}} \quad (315)$$

where pc and fc are the prediction and the fitness of each XCS classifier $\boldsymbol{c} \in \boldsymbol{M_i}$

Note that each supporting classifier contributes to strengthen the hypothesis on the corresponding action according to its current fitness. The winning action (i*) is the one that maximizes the estimated reward. In an NXCS used for prediction tasks, we decided to leave unaltered the selection mechanism and to define an outputs-blending mechanism suited to match the characteristics of the given task. In particular, to forecast the next value of a time series, a fitness weighted average rule is enforced by blending the outputs of the experts that belong to the match set according to their fitness. In this way, the more reliable is an expert the more its ANN output contributes to the overall prediction. In particular, given a match set M, the overall primary output, say p(t+1), is evaluated according to the formula:

$$\boldsymbol{p}(\boldsymbol{t}+1) = \frac{\sum_{c \in M_i} \boldsymbol{p_c} \cdot \boldsymbol{f_c}}{\sum_{c \in M_i} \boldsymbol{f_c}} \quad (316)$$

where pc and fc are the prediction—which coincides with the primary ANN output—and the fitness of each NXCS expert that belongs to the match set M.

12.4. Customizing NXCS experts for stock market forecasting

According to the general choice made for NXCSs, guards and the neural predictors have been supplied with different information, to better exploit their capabilities of dealing with binary and real inputs, respectively. As technical analysis indicators are commonly used by financial traders to predict transitions between different market regimes, we assumed they could be sensible inputs for the partitioning system, whereas the forecasting of future prices clearly requires knowledge

about past prices. In agreement with these informative concepts, NXCS experts have been tailored to deal with stock market forecasting as follows: (i) some technical-analysis domain knowledge has been embodied into the guard of each expert; (ii) a novel kind of feedforward ANN, able to deal with short-term dynamics on a weekly base (i.e., 5-days), has been defined and adopted to perform the prediction task.

12.4.1 Embodying technical-analysis domain knowledge into NXCS guards

$$\boldsymbol{DOA}_{N1,N2}(\boldsymbol{t}) = \boldsymbol{MA}_{N1}[\boldsymbol{q}](\boldsymbol{t}) - \boldsymbol{MA}_{N2}[\boldsymbol{q}](\boldsymbol{t}) \quad (317)$$

$$\boldsymbol{ROC}_{N}(\boldsymbol{t}) = \frac{\boldsymbol{q}(\boldsymbol{t}) - \boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})}{\boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})} \quad (318)$$

$$\boldsymbol{RSI}_{N}(\boldsymbol{t}) = \left(1 + \frac{\boldsymbol{MA}^{+}_{N}[\boldsymbol{neg}](\boldsymbol{t})}{\boldsymbol{MA}^{+}_{N}[\boldsymbol{pos}](\boldsymbol{t})}\right)^{-1}$$

$$\boldsymbol{where}: \quad (319)$$

$$\boldsymbol{pos}(\boldsymbol{t}) = \max\left(0, \frac{\boldsymbol{q}(\boldsymbol{t}) - \boldsymbol{q}(\boldsymbol{t}-1)}{\boldsymbol{q}(\boldsymbol{t})}\right)$$

$$\boldsymbol{neg}(\boldsymbol{t}) = \max\left(0, \frac{\boldsymbol{q}(\boldsymbol{t}-1) - \boldsymbol{q}(\boldsymbol{t})}{\boldsymbol{q}(\boldsymbol{t})}\right)$$

$$\boldsymbol{CV}_{N}(\boldsymbol{t}) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N}+\boldsymbol{i}) - \boldsymbol{r}(\boldsymbol{t}-\boldsymbol{N}+\boldsymbol{i}) \quad (320)$$

$\boldsymbol{where}$ r is the connecting $\boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})$ and $\boldsymbol{q}(\boldsymbol{t})$

$\boldsymbol{UP}_{N} = 1$ if local minima in $[\boldsymbol{t}-\boldsymbol{N}, \boldsymbol{t}]$ form an increasing sequence; 0 otherwise.

Being $\boldsymbol{q}_{k} = \boldsymbol{q}(\boldsymbol{t}-\boldsymbol{k})$, a local minimum occurs if $\boldsymbol{q}_{k} = \min(\boldsymbol{q}_{k-1}, \boldsymbol{q}_{k}, \boldsymbol{q}_{k+1})$ for $0 < \boldsymbol{k} < \boldsymbol{N}$. (321)

Note that $\boldsymbol{q}(\boldsymbol{t})$ and $\boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})$ are considered minima.

$\boldsymbol{DW}_{N} = 1$ if local maxima in $[\boldsymbol{t}-\boldsymbol{N}, \boldsymbol{t}]$ form an increasing sequence; 0 otherwise.

Being $\boldsymbol{q}_{k} = \boldsymbol{q}(\boldsymbol{t}-\boldsymbol{k})$, a local maximum occurs if $\boldsymbol{q}_{k} = \max(\boldsymbol{q}_{k-1}, \boldsymbol{q}_{k}, \boldsymbol{q}_{k+1})$ for $0 < \boldsymbol{k} < \boldsymbol{N}$. (322)

Note that $\boldsymbol{q}(\boldsymbol{t})$ and $\boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})$ are considered maxima

As a kind of relevant domain knowledge, some technical analysis indicators have been considered. They are reported in in the previous formulas, under the assumption that q(t) is the

stock price at day t, and that MA (i.e., Moving Average) is a generic ‘‘low-pass’’ filtering operator defined as:

$$MA_N[x](t) = \frac{1}{N}\sum_{i=0}^{N-1} x(t-i) \quad (323)$$

The input of a NXCS guard is a vector of eight binary features, i.e., $x_b = (b_1, b_2, ..., b_8)$

Table 40: Binary inputs, to be , matched by the guard of an XCS expert

Binary Inputs	Note
DOA (1,30) > 30 and DOA (1,30) < 30	We’ve just crossed a valley
DOA (1,30) < 30 and DOA (1,30) > 30	We’ve just surmounted a peak
RSI (15, t) < 0,3	Too many sales vs purchases
RSI (15, t) > 0,7	Too many purchases vs sales
UP (5,t) = 1	Bullish period
DW (5,t) = 1	Bearish period
UP (5,t-1) = 1 and UP (5,t) not = 1	Current Bullish period has finished
DW (5,t-1) = 1 and DW (5,t) not = 1	Current Bearish period has finished

. As shown in Table 2, the meaning of each input feature directly stems from a technical-analysis indicator. In particular, in its basic form, a feature has been obtained by adding a relational constraint (e.g., b3; b4). More complex features have been obtained composing basic features (e.g., b1; b2).

Accordingly, an NXCS guard consists of a vector of eight values in {0,1,#}. It can either take into account or disregard a given input feature by suitably setting the corresponding matching condition. In particular, a matching condition set to either ‘‘0’’ or ‘‘1’’ enables a feature, whereas a matching condition set to ‘‘#’’ (I don’t care) allows to disregard it. The overall matching between an input vector and an NXCS guard occurs when all ‘‘non-#’’ values coincide.

Each resulting feature has a very simple interpretation, together with a default action to be undertaken in absence of other relevant information. As an example, let us consider the feature b3, obtained by adding the relational constraint ‘‘<0.3’’ to the indicator ‘‘relative strength index’’, i.e., RSIn (t). Such an indicator basically accounts for the number of purchases vs. the number of sales in a given observational window of length N. Hence, the relational expression RSI15(t) < 0:3 tries to capture the informal semantics ‘‘heavy price falls’’. In this case, the associated default

action (i.e., the action to be undertaken in absence of further information) would be ‘‘take a long position’’.

Thus, the feature b3 could be the precondition of a technical-analysis rule that—although controversial—can be summarized as ‘‘when there is a heavy price fall, then take a long position’’. Unfortunately, due to the fact that an NXCS guard usually takes into account several features together, it is not common the case that all involved features suggest taking the same position. In fact, it is far more common the case of contradictory default suggestions about the action to be undertaken. That is why, in the proposed system, no particular emphasis is given on extracting technical-analysis rules embedded within genetic guards.

Note that indicators used to prepare inputs for the genetic guards are evaluated on windows that range from 2to 6 weeks (i.e., usually 10–30 days of actual trading), depending on the selected indicator. The reason why there is no common observational window depends only loosely on the experiments performed while tuning the system. In fact, the decision about which window has to be used for a given indicator results from a priori, thus subjective, appraisals made on the underlying dynamics of each indicator. Notwithstanding this arbitrary degree of freedom, such a choice should not be considered as a drawback, since ‘‘carving up’’ indicators with the aim of optimizing the behavior of the system might introduce an unwanted correlation between the system’s parameters and the time series used for tuning it. In addition, given the great number of indicators that have been exploited, an attempt to find the best window for each indicator would end up to a long tuning activity.

12.4.2 Devising a feedforward ANN for stock market forecasting

Two different requirements influenced the design and implementation of the feedforward ANN used to perform predictions on a local basis; i.e., limiting the overfitting phenomenon and the influence of noise. Overfitting training data while performing poorly on test data is one of the major problems of learning algorithms, in particular when noisy data are involved. This problem is particularly relevant for economic time series, which are very noisy and require complex learning algorithms. Overfitting is strongly related to non-stationarity: data may appear to be noisy when using inadequate models. On the other hand, partitioning the input space reduces the number of learning examples for each model, thus increasing the risk of overfitting each smaller training set. Two approaches are commonly adopted to reduce overfitting while using ANNs: stopping the learning process when the network has attained good performance on a separate validation set, and adopting some form of weight pruning (e.g., [186,187]) as a means of reducing the network’s ability to learn the least significant patterns. Furthermore, it is common practice to pre-process data for reducing the effect of outliers—often a result of exogenous shocks—and to compress the dynamic range, for instance using logarithms as in [188]. One of the major problems that arises when using a single-output MLP to forecast time series is that the only information used by the backpropagation (BP) algorithm is the derivative of the cost function—usually the squared error

between desired and actual output.With noisy data, the small amount of information contained in a single output is typically inadequate to distinguish non-linearities from noise. A possible solution could be to predict more than one output, to provide more information aimed at characterizing the output dynamics. However, as pointed out by Weigend and Zimmermann [189], this approach only results in a slight performance increase over single-output networks. For this reason, the authors propose a multilayer architecture explicitly designed to utilize all available information and to minimize the risk of overfitting. The network forecasts dynamic variables, such as derivatives and curvatures of prices on different time spans, instead of directly forecasting future prices. These variables are subsequently combined in an interaction layer (with constant weights), which outputs several estimates that, properly averaged, give the final prediction.

In this PHD thesis, we agree with the conjecture of Weigend and Zimmermann, aimed at increasing the information flow by predicting multiple interacting outputs. However, the approach adopted here is much simpler, as it does not rely on numerous layers or on explicit forecast of derivatives and curvatures of the price. The feedforward ANN defined and used by NXCS to tackle the 3 In fact, the BP algorithm on a standard MLP keeps the information relative to each point separate, so that the derivatives of the cost function relatively to each output are independent, and outputs can interact only through the hidden units. problem of stock market forecasting is shown in Fig. 2. The first block (Block1) of this architecture is a standard MLP, whereas the second block (Block2), whose structure is similar to the one adopted on cascade correlation architectures [190], has been designed to enforce the desired outputs interaction. Note that the resulting ANN is still a feedforward network, so that backpropagation can be used. The input layer of Block1 is denoted by X, whereas the input layer of Block2 (which corresponds to the outputs of Block1) is denoted by Y. The overall ANN is used to predict the prices for the next three days, 4 represented by the output vector $\boldsymbol{p} = (\boldsymbol{p}_{t+1}, \boldsymbol{p}_{t+2}, \boldsymbol{p}_{t+3})$. The primary output is associated with the prediction of the next-day price of the stock index (i.e., p t+1). Note that p t+2 is an input for p t+1, and that p t+3 is an input for both p t+1 and p t+2; hence, the primary output takes as input every other forecasted price. In this way,derivatives of the standard Mean Square Error cost function are not independent with respect to each output, thus helping to reduce the influence of noise—according to the cited Weigend and Zimmermann's conjecture.

The input $\boldsymbol{X} \equiv (\boldsymbol{r}_1, \boldsymbol{r}_2, ...\boldsymbol{r}_{10})$ is a vector of 10 numerical features (see Table 3). In particular, the input of a single ANN is represented by the value of 5 technical-analysis indicators, together with the prices of the last 5 days, suitably filtered to reduce the effect of outliers. The following equation has been used for pre-processing prices:

$$\boldsymbol{dl}_N(\boldsymbol{t}) = \boldsymbol{sign}[\boldsymbol{q}(\boldsymbol{t}) - \boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})] \cdot \ln\left(\left|\frac{\boldsymbol{q}(\boldsymbol{t}) - \boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})}{\boldsymbol{q}(\boldsymbol{t}-\boldsymbol{N})}\right| + 1\right) \quad (324)$$

where q(t) and q(t-N) are the prices at day t and t-N, respectively, whereas $\boldsymbol{dl}_N(\boldsymbol{t})$is the corresponding input to the ANN. The equation used for preprocessing is similar to the one used

by Giles et al. [191], and to the one proposed in [192], the most notable difference being that we set N =5 instead of N =1.

Figure 70: A feedforward ANN for stock market forecasting, with 10 numerical inputs and 3 outputs.

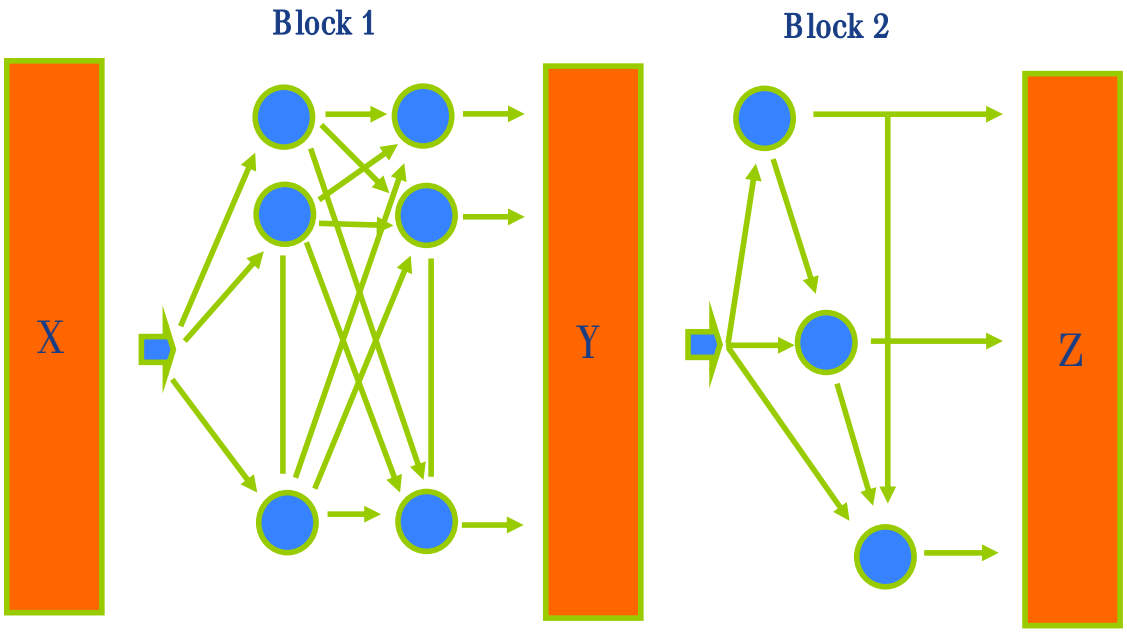


Table 41: Inputs to the ANN

Numerical Inputs	Note
DOA (1,30) / MA (30,t)	Difference of averages (normalized with MA30)
MA 3 [ROC15(T)]	3 weeks rate of change (averaged with MA3)
CV(10,T)	2 weeks convexity
RSI (15, t)	3 weeks relative strength index
MA 10 [SD30(T)]	Monthly Standard Deviation of relative price (averaged with MA 10)
DL (5,T)	Last quotations (k = 0)
DL (5,T-1)	Last quotations (k = 1)
DL (5,T-2)	Last quotations (k = 2)
DL (5,T-3)	Last quotations (k = 3)
DL (5,T-4)	Last quotations (k = 4)

Note that setting N = 5 enforces a non-linear transformation of the current week's prices, which allows to embody within dl inputs also information about the previous week. The

quantity $\sigma_{30}(t)$ represents the standard deviation of the relative price variations, computed for the last 30 days.

12.5. Experimental results

Figure 71: SP500

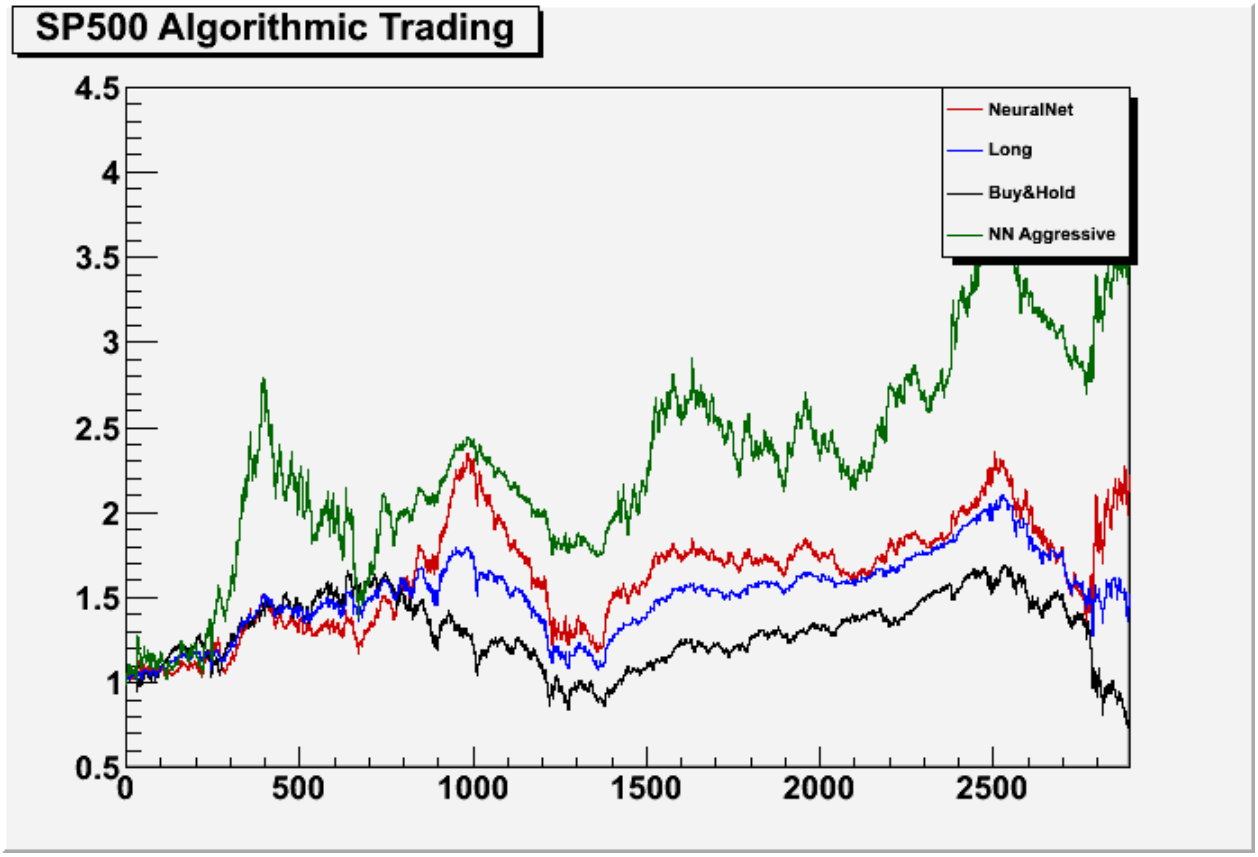


Table 42 : Perfomance metrics model 12/9/97-12/3/2009

SP500 Model	An. Return	St Dev	SR
1 long / 1 short	7.30%	20%	0.2650
Benchmark	-1.85%	22%	-0.1750
1 long / Cash	3.80%	15%	0.1200
2 long / 1 short	8.30%	33%	0.1909
2 long / 2 short - 0,5 MA	11.60%	20%	0.4800

Table 43: EUR/USD model 29/11/90-11/3/10

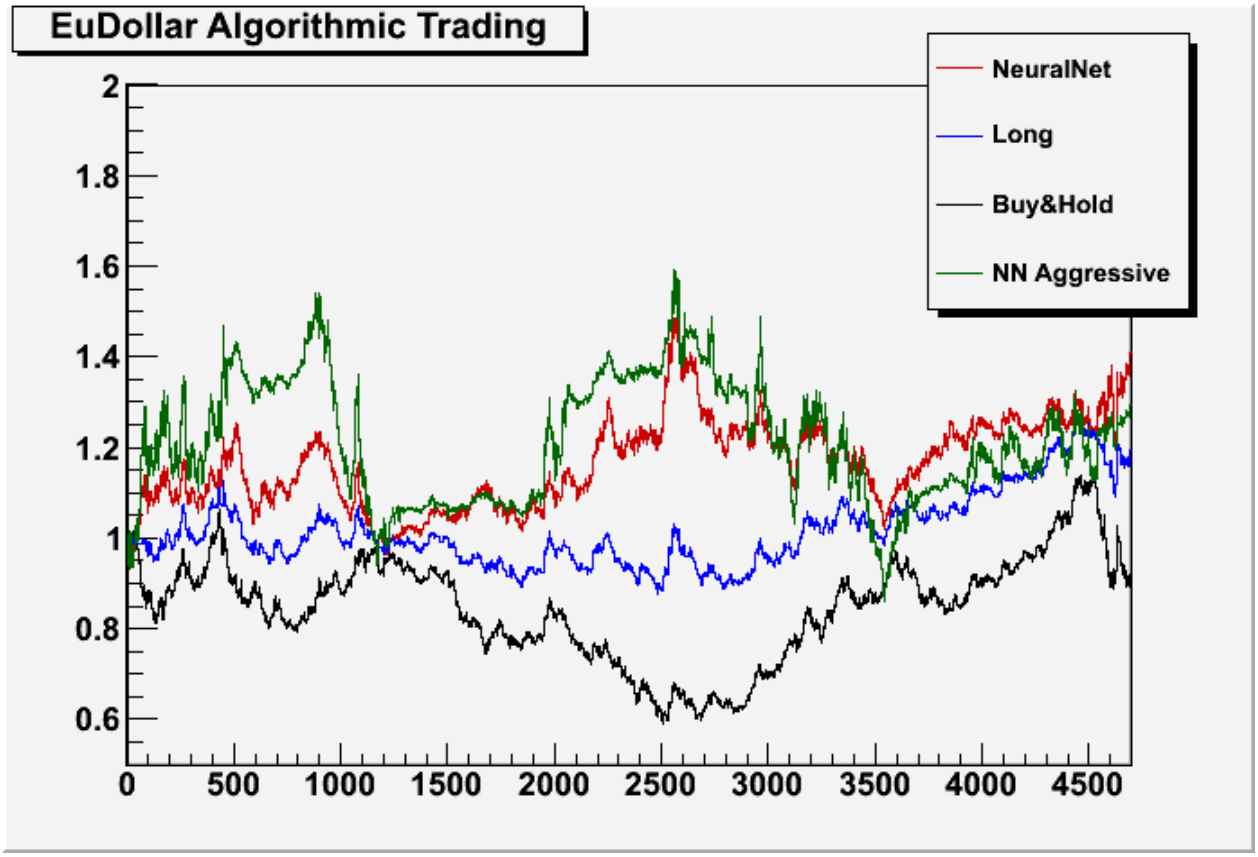


Table 44: Perfomance metrics model 29/11/90-11/3/2009

EUR/USD Model	An. Return	St Dev	SR
1 long / 1 short	1.87%	10%	-0.0130
Benchmark	-0.44%	10%	-0.2440
1 long / Cash	0.95%	7%	-0.1500
2 long / 1 short	2.27%	15%	0.0180
2 long / 2 short - 0,5 MA	1.44%	13%	-0.0431

12.6. Forecasting stock market direction with Support Vector Machines

12.6.1 Experiment design

In our empirical analysis, we set out to examine the weekly changes of the NIKKEI 225 Index. The NIKKEI 225 Index is calculated and disseminated by Nihon Keizai Shinbum Inc. It measures the composite price performance of 225 highly capitalized stocks trading on the Tokyo Stock Exchange (TSE), representing a broad cross-section of Japanese industries. Trading in the index has gained unprecedented popularity in major financial markets around the world. Futures and options contracts on the NIKKEI 225 Index are currently traded on the Singapore International Monetary Exchange Ltd (SIMEX), the Osaka Securities Exchange and the Chicago Mercantile

Exchange. The increasing diversity of financial instruments related to the NIKKEI 225 Index has broadened the dimension of global investment opportunity for both individual and institutional investors. There are two basic reasons for the success of these index trading vehicles. First, they provide an effective means for investors to hedge against potential market risks. Second, they create new profit making opportunities for market speculators and arbitrageurs. Therefore, it has profound implications and significance for researchers and practitioners alike to accurately forecast the movement direction of NIKKEI 225 Index.

12.6.2 Model inputs selection

Most of the previous researchers have employed multivariate input. Several studies have examined the cross-sectional relationship between stock index and macroeconomic variables. The potential macroeconomic input variables which are used by the forecasting models include term structure of interest rates (TS), short-term interest rate (ST), long-term interest rate (LT), consumer price index (CPI), industrial production (IP), government consumption (GC), private consumption (PC), gross national product (GNP) and gross domestic product (GDP). However, Japanese interest rate has dropped down to almost zero since 1990. Other macroeconomic variables weekly data are not available for our study.

Japanese consumption capacity is limited in the domestic market. The economy growth has a close relationship with Japanese export. The largest export target for Japan is the United States of America (USA), which is the leading economy in the world. Therefore, the economic condition of USA influences Japan economy, which is well represented by the NIKKEI 225 Index. As the NIKKEI 225 Index to Japan economy, the S& P 500 Index is a well-known indicator of the economic condition in USA. Hence, the S& P 500 Index is selected as model input. Another import factor that aspects the Japanese export is the exchange rate of US Dollars against Japanese Yen (JPY), which is also selected as model input. The prediction model can be written as the following function:

$$Direction_t = F\left(S_{t-1}^{S\&P500}, S_{t-1}^{JPY}\right) \quad (325)$$

where $S_{t-1}^{S\&P500}$ and S_{t-1}^{JPY} are 1rst order difference natural logarithmic transformation to the raw S& P 500 index and JPY at time t−1, respectively. Such transformations implement an effective detrending of the original time series. Direction t is a categorical variable to indicate the movement direction of NIKKEI 225 Index at time t. If NIKKEI 225 Index at time t is larger than that at time t − 1, Direction t is 1. Otherwise, Direction t is −1.

12.6.3 Comparisons with other forecasting methods

To evaluate the forecasting ability of SVM, we use the random walk model (RW) as a benchmark for comparison. RW is a one-step-ahead forecasting method, since it uses the current actual value to predict the future value.

We also compare the SVM's forecasting performance with that of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and elman backpropagation neural networks (EBNN). LDA can handle the case in which the within-class frequencies are unequal and its performance has been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set, thereby guaranteeing maximal separability. QDA is similar to LDA, only dropping the assumption of equal covariance matrices. Therefore, the boundary between two discrimination regions is allowed to be a quadratic surface (for example, ellipsoid, hyperboloid, etc.) in the maximum likelihood argument with normal distributions. Interested readers should refer to [193] or some other statistical books for a more detailed description. In this paper, we derive a linear discriminant function of the form:

$$L\left(S_{t-1}^{S\&P500}, S_{t-1}^{JPY}\right) = a_0 + a_1 S_{t-1}^{S\&P500} + a_2 S_{t-1}^{JPY} \quad (326)$$

and a quadratic discriminant function of the form:

$$Q\left(S_{t-1}^{S\&P500}, S_{t-1}^{JPY}\right) = a + P\left(S_{t-1}^{S\&P500}, S_{t-1}^{JPY}\right)^T + \left(S_{t-1}^{S\&P500}, S_{t-1}^{JPY}\right) T \left(S_{t-1}^{S\&P500}, S_{t-1}^{JPY}\right)^T \quad (327)$$

where a0,a1,a2,a,P,T are coefficients to be estimated.

Elman Backpropagation Neural Network is a partially recurrent neural network. The connections are mainly feed-forward but also include a set of carefully chosen feedback connections that let the network remember cues from the recent past. The input layer is divided into two parts: the true input units and the context units that hold a copy of the activations of the hidden units from the previous time step. Therefore, network activation produced by past inputs can cycle back and affect the processing of future inputs. For more details about Elman Backpropagation Neural Network, refer to [194]and[195].

The whole data set covers the period from January 1, 1996 to December 31, 2008, a total of 676 pairs of observations. The data set is divided into two parts. The first part (640 pairs of observations) is used to determine the specifications of the models and parameters. The second part (36 pairs of observations) is reserved for out-of-sample evaluation and comparison of performances among various forecasting models.

12.6.4 Experiment results

Table 45: Forecasting performance of different classification methods

Classification Method	Hit Ratio %
RW	50
LDA	55
QDA	69
EBNN	69
SVM	73

SVM has the highest forecasting accuracy among the individual forecasting methods. One reason that SVM performs better than the earlier classification methods is that SVM is designed to minimize the structural risk, whereas the previous techniques are usually based on minimization of empirical risk. In other words, SVM seeks to minimize an upper bound of the generalization error rather than minimizing training error. So SVM is usually less vulnerable to the over-fitting problem.

QDA out-performs LDA in term of hit ratio, because LDA assumes that all the classes have equal covariance matrices, which is not consistent with the properties of input variable belonging to different classes as shown in Tables 2 and 3. In fact, the two classes have different covariance matrices. Heteroscedastic models are more appropriate than homoscedastic models.

The integration of SVM and the other forecasting methods improves the forecasting performance. Different classification methods typically have access to different information and therefore produce different forecasting results. Given this, we can combine the individual forecaster's various information sets to produce a single superior information set from which a single superior forecast could be produced.

13. Other non-linear Techniques

13.1. Takens' theorem - delay embedding theorem – State space

In mathematics, a delay embedding theorem gives the conditions under which a chaotic dynamical system can be reconstructed from a sequence of observations of the state of a dynamical system. The reconstruction preserves the properties of the dynamical system that do not change under smooth coordinate changes, but it does not preserve the geometric shape of structures in phase space.

Takens' theorem is the 1981 delay embedding theorem of Floris Takens[196]. It provides the conditions under which a smooth attractor can be reconstructed from the observations made with a generic function. Later results replaced the smooth attractor with a set of arbitrary box counting dimension and the class of generic functions with other classes of functions.

Delay embedding theorems are simpler to state for discrete-time dynamical systems. The state space of the dynamical system is a v-dimensional manifold M. The dynamics is given by a smooth map

$$\boldsymbol{f} : \boldsymbol{M} \rightarrow \boldsymbol{M} \quad (328)$$

Assume that the dynamics f has a strange attractor A with box counting dimension dA. Using ideas from Whitney's embedding theorem, A can be embedded in k-dimensional Euclidean space with

$$\boldsymbol{k} > 2\boldsymbol{d}_{\boldsymbol{A}} \quad (329)$$

That is, there is a diffeomorphism φ that maps A into Rk such that the derivative of φ has full rank. A delay embedding theorem uses an observation function to construct the embedding function. An observation function α must be twice-differentiable and associate a real number to any point of the attractor A. It must also be typical, so its derivative is of full rank and has no special symmetries in its components. The delay embedding theorem states that the function

$$\phi_{\boldsymbol{T}}(\boldsymbol{x}) = \left(\alpha(\boldsymbol{x}), \alpha(\boldsymbol{f}(\boldsymbol{x})), \ldots\ldots, \alpha\left(\boldsymbol{f}^{\boldsymbol{k}-1}(\boldsymbol{x})\right)\right) \quad (330)$$

is an embedding of the strange attractor A.

Figure 72: 1-D signal

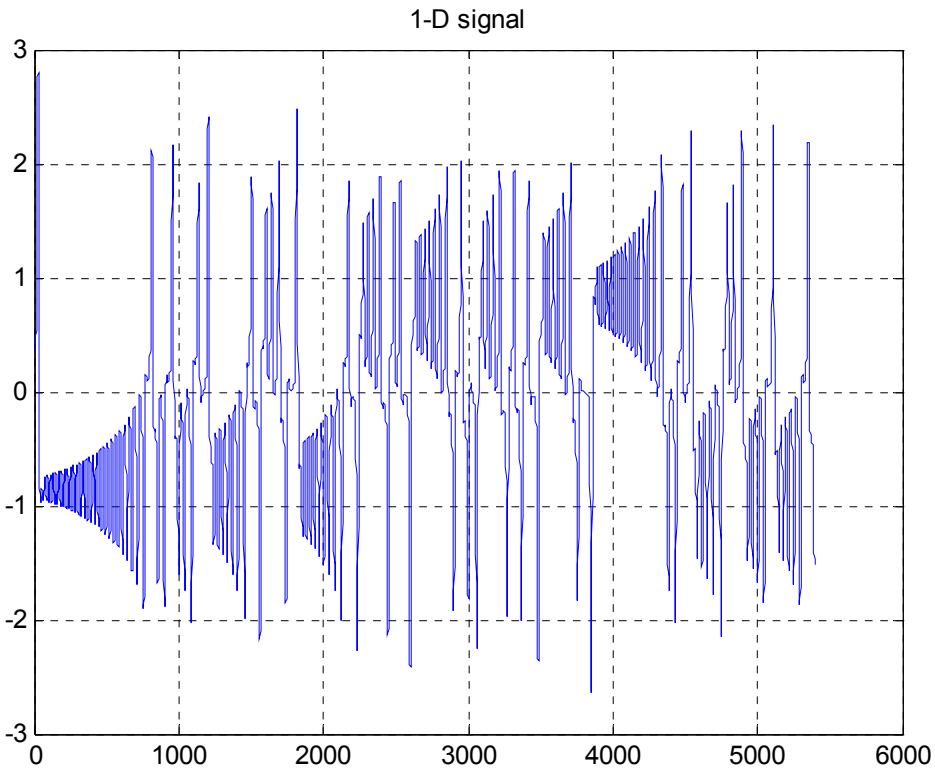
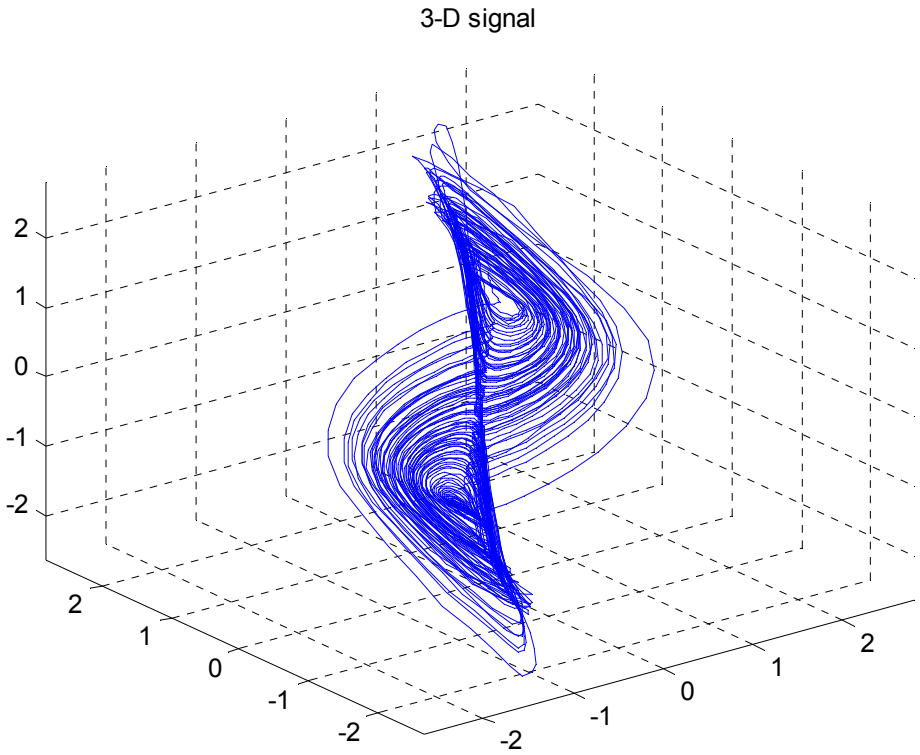
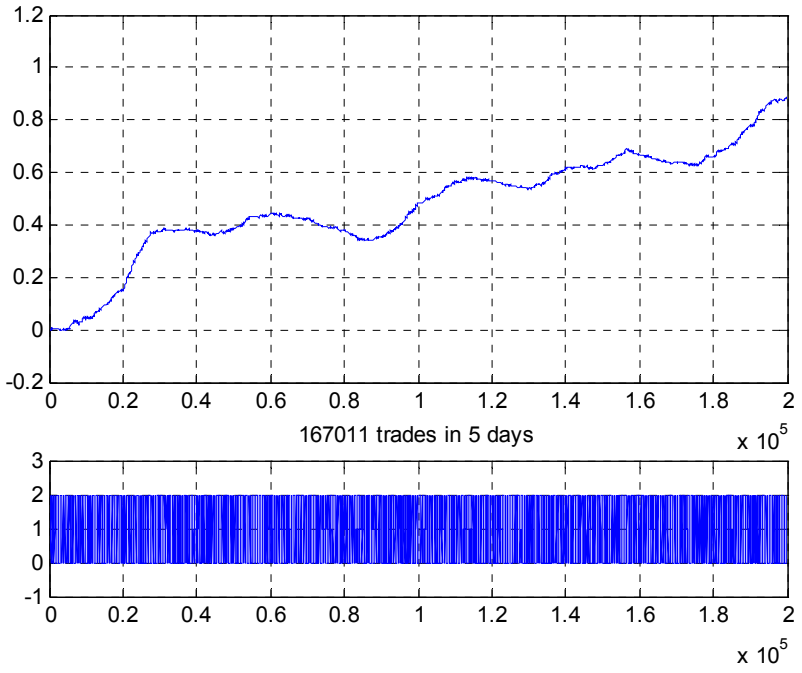


Figure 73: 3-D signal



13.1.1 FX application - Takens' theorem - delay embedding theorem

Figure 74: Results GBP/USD 1 second



13.2. Macroeconomics Forecasting – From VARMA to Reduced form VAR

Smets and Wouters(2004[197], 2007[198])

- Linearized model for the United States economy
- System of 14 equations with a companion system to model potential output
- 14 endogenous variables
- 7 exogenous disturbances
- Equivalent to a structural VARMA(3,2) model with 7 equations and numerous parameter restrictions
- Ignoring restrictions, you can invert the MA lag operator to obtain an unrestricted reduced-form VAR(3) model

Here we do the operations to transform to VAR reduced form

$$A_0 Y_t = a'' + \sum_i A_i'' Y_{t-i} + \sum_i B_i'' U_{t-i} + B_0 U_t \quad (331)$$

$$B_0 U_t = A_0 W_t \quad (332)$$

$$U_t = B_o^{-1} A_0 W_t \quad (333)$$

$$A_0 Y_t = a'' + \sum_i A_i'' Y_{t-i} + \sum_i B_i'' B_o^{-1} A_0 W_t + A_0 W_t \quad (334)$$

$$Y_t = A_0^{-1} a'' + \sum_i A_0^{-1} A_i'' Y_{t-i} + \sum_i A_0^{-1} B_i'' B_o^{-1} A_0 W_t + W_t \quad (335)$$

$$Y_t = a' + \sum_i A_i' Y_{t-i} + \sum_i B_i' W_{t-i} + W_t \quad (336)$$

$$Y_t - \sum_i A_i' Y_{t-i} = a' + \sum_i B_i' W_{t-i} + W_t \quad (337)$$

$$A'(L) Y_t = a' + B'(L) W_t \quad (338)$$

$$B'(L)^{-1} A'(L) Y_t = B'(L)^{-1} a' + W_t \quad (339)$$

$$A(L) Y_t = a + W_t \quad (340)$$

$$Y_t - \sum_i A_i Y_{t-i} = a + W_t \quad (341)$$

$$Y_t = a + \sum_i A_i Y_{t-i} + W_t \quad (342)$$

Series	Description	FRED Series
rGDPt	Change in log nominal GDP	GDP
rDEFt	Change in logGDP deflator	GDPDEF
rWAGESt	Change in log total paid compensation	COE
rHOURSt	Change in log non-farm business sector hours worked	HOANBS
rTB3t	Interestrate for 3-month Treasury bill	TB3MS
rCONSt	Change in log personal consumption expenditures	PCEC
rINVt	Change in log gross private domestic investment	GPDI
rUNEMPt	Rate of unemployment	UNRATE

$$\mathbf{Y_t} = \mathbf{a} + \sum_{\mathbf{i}} \mathbf{A_i} \mathbf{Y_{t-i}} + \mathbf{W_t}$$
$$\mathbf{W_t} \sim \mathbf{N}(\mathbf{0}, \mathbf{Q})$$
$$\mathbf{Y_{t-i}} = \begin{bmatrix} \text{rGDPt} \\ \text{rDEFt} \\ \text{rWAGESt} \\ \text{rHOURSt} \\ \text{rTB3t} \\ \text{rCONSt} \\ \text{rINVt} \\ \text{rUNEMPt} \end{bmatrix} \qquad (343)$$

Differences with Smets-Wouters Model

Time series choices

- Replace Fed Funds rate with 3-month Treasury Bills
- Replace hourly wages paid with total wages paid
- Added unemployment since a psychological indicator of economic health

Time series formats

- Smets-Wouters mixes integrated and differenced data
- We will use differenced (or rate) data exclusively

Nominal versus real

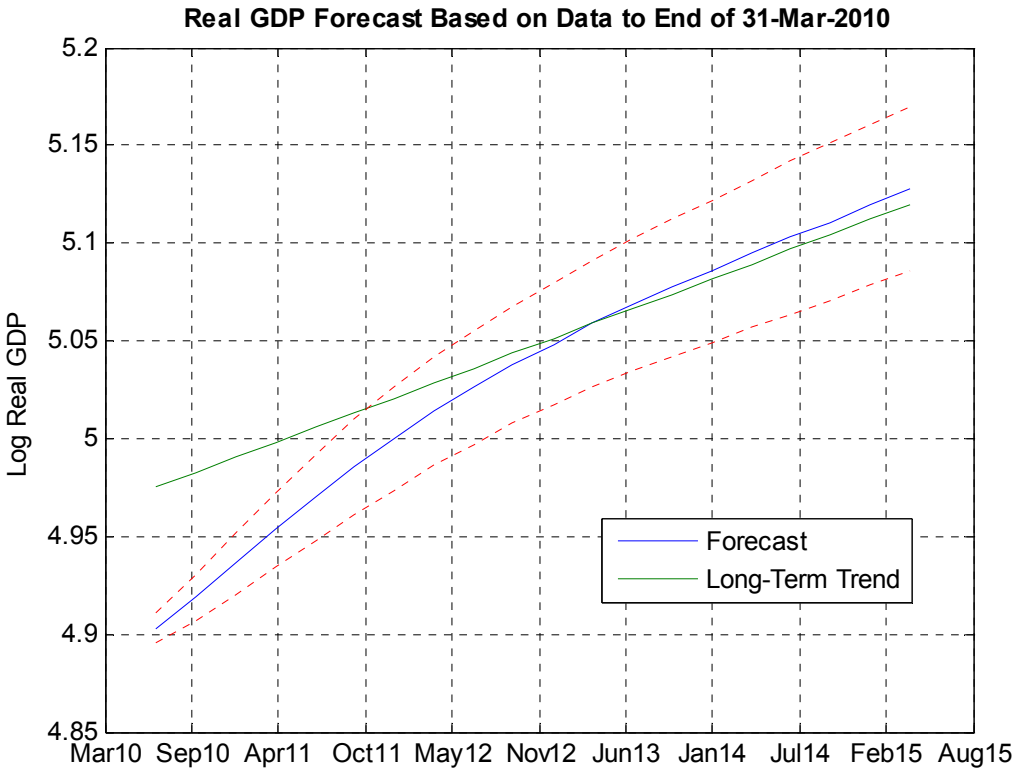
- Use nominal data since GDP deflator is in the model

Detrending

- Smets-Wouters detrends some series with a common GDP growth rate trend
- Rate data induces stationarity by differencing instead of detrending

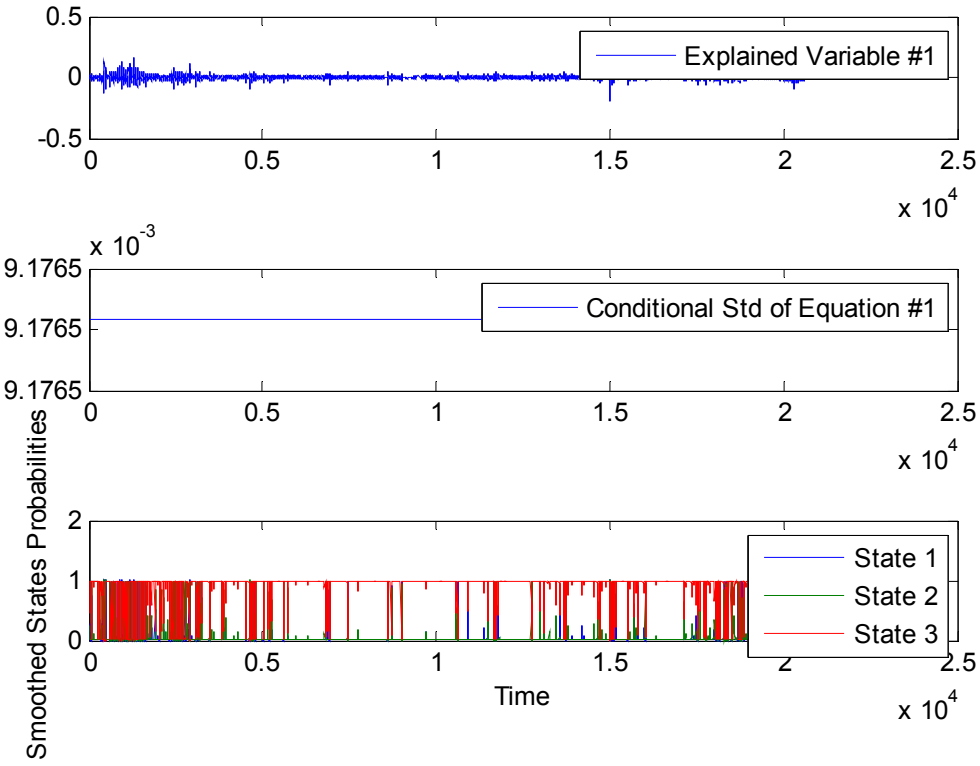
Order of auto-regression

- Theory says order of model is 3 although lag 3 matrix is extremely sparse
- Let data determine the optimal lag order for the model



13.3. Markov switching SP500 / VIX 3 states

Figure 75: SP500 Markov Switching 3 states 1928-2010



***** Numerical Optimization Converged *****

Final log Likelihood: 64951.4601
Number of estimated parameters: 13
Type of Switching Model: Univariate
Distribution Assumption -> Normal
Method SE calculation -> 1

***** Final Parameters for Equation #1 *****

---> Non Switching Parameters <---

There was no Non Switching Parameters for Indep matrix of Equation #1. Skipping this result

Non Switching Variance of model
  Value: 0.000084
  Std Error (p. value): 0.0000 (0.00)

---> Switching Parameters (Regressors) <---

Switching Parameters for Equation #1 - Indep column 1

State 1
  Value: 0.0421
  Std Error (p. value): 0.0014 (0.00)
State 2
  Value: -0.0384
  Std Error (p. value): 0.0012 (0.00)
State 3
  Value: 0.0004
  Std Error (p. value): 0.0001 (0.00)

---> Transition Probabilities Matrix (std. error, p-value) <---

0.20 (0.03,0.00)	0.23 (0.04,0.00)	0.01 (0.00,0.00)
0.19 (0.04,0.00)	0.22 (0.03,0.00)	0.01 (0.00,0.00)
0.61 (0.06,0.00)	0.54 (0.05,0.00)	0.98 (0.01,0.00)

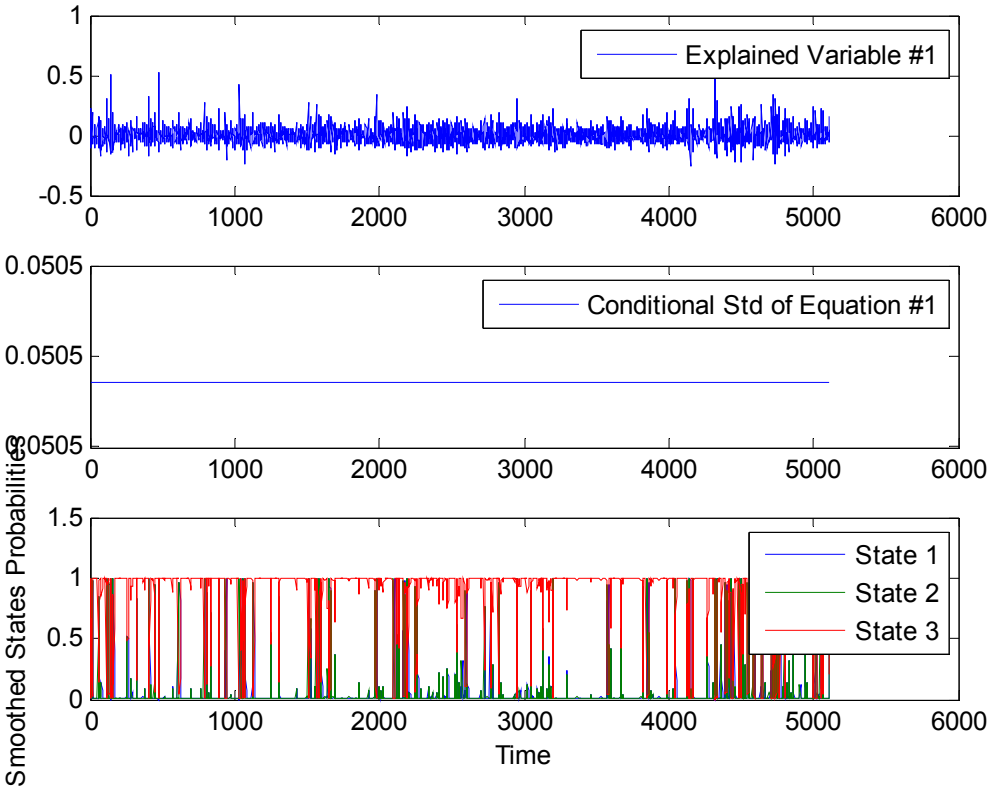
---> Expected Duration of Regimes <---

Expected duration of Regime #1: 1.25 time periods

Expected duration of Regime #2: 1.29 time periods

Expected duration of Regime #3: 46.47 time periods

Figure 76: VIX Switching 3 states 1980-2010



**** Numerical Optimization Converged *****

Final log Likelihood: 7467.6832

Number of estimated parameters: 13

Type of Switching Model: Univariate

Distribution Assumption -> Normal

Method SE calculation -> 1

***** Final Parameters for Equation #1 *****

---> Non Switching Parameters <---

There was no Non Switching Parameters for Indep matrix of Equation #1. Skipping this result

Non Switching Variance of model

Value: 0.002551

Std Error (p. value): 0.0001 (0.00)

---> Switching Parameters (Regressors) <---

Switching Parameters for Equation #1 - Indep column 1

State 1
Value: -0.0548
Std Error (p. value): 0.0044 (0.00)
State 2
Value: 0.1821
Std Error (p. value): 0.0071 (0.00)
State 3
Value: -0.0006
Std Error (p. value): 0.0008 (0.49)

---> Transition Probabilities Matrix (std. error, p-value) <---

0.49 (0.06,0.00)	0.94 (0.10,0.00)	0.00 (0.00,1.00)
0.15 (0.03,0.00)	0.06 (0.03,0.09)	0.02 (0.00,0.00)
0.36 (0.05,0.00)	0.00 (Inf,1.00)	0.98 (0.01,0.00)

---> Expected Duration of Regimes <---

Expected duration of Regime #1: 1.95 time periods
Expected duration of Regime #2: 1.06 time periods
Expected duration of Regime #3: 49.50 time periods

14. Asymmetric Algorithmic trading strategies

By asymmetric algorithmic trading strategies we mean the techniques used to tackle assets that show asymmetric behaviour on the upside and on the downside movement. We build algorithms that may have different buying than selling signals, valid in momentum or mean-reverting assets and asymmetry on the upside/downside.

14.1. Commodity Algo

In this section we show how the extract good sharpe ratios from commodities that show mean-reverting / momentum and asymmetry the algorithms are optimized for every commodity.

14.1.1 Commodity market characteristics

- Some commodities trend while others counter-trend.

- Commodities have asymmetric returns: Fast upward movements versus slow downward movements, for instance (gasoline).
- High volatility.

The strategy takes advantage of these characteristics:

- The algorithm uses Exponential Moving Average (EWMA) indicators designed to extract value from these characteristics.
- Algorithm parameters for each commodity are chosen independently, leading to a higher individual performance, a lower correlation between the returns of individual underlyings, and a higher Sharpe ratio for the overall portfolio.
- Applies a strategy that can be long or short, potentially generating positive returns in both bull and bear markets.
- Generally, high commodity price volatility drives positive strategy returns.

Other features designed to avoid the pitfalls of passive or other active commodity trading strategies:

- An algorithmic investment approach ensures transparency and avoids the style drift of active managers.
- Positions are adjusted based on signals generated from current market conditions in order to constantly capture new market opportunities.
- Monthly rebalancing of commodity exposure maintains portfolio diversification and generates rebalancing returns (Erb and Harvey, 2006)[199].

Table 46: Commodity weights and EWMA

Commodity	EMA 1,A	EMA2	EMA 1,B	EMA2	Weights
WTI Crude Oil	70	99	115	163	22,90%
Brent Crude Oil	30	42	110	154	12,50%
Gasoil	35	49	70	98	4,50%
RBOB Gasoline	25	35	115	161	2,80%
Heating oil	25	34	95	127	4,30%
HHUB Natural Gas	55	74	40	54	4,00%
Gold	40	28	65	46	4,70%
Silver	60	44	70	51	1,40%
Copper	20	24	50	59	7,00%
Aluminum	25	18	65	46	6,60%
Zinc	95	127	40	54	2,40%
Nickel	55	77	105	147	1,70%
Lead	105	147	120	168	1,10%
Sugar #11	100	130	40	52	3,50%
Corn	100	135	20	27	5,10%
Soybean	100	120	30	36	5,00%
Wheat	95	68	30	22	3,60%
Coffee	70	55	25	20	2,60%
Live Cattle	90	77	65	55	4,30%

Table 47: Strategy perfomance

	10/97-4/2010
Annualised Excess Returns	20.60%
Volatility of Excess Returns	15.46%
Sharpe Ratio	1,43
Maximum Drawdown	17%

15. Integrated Algorithmic Trading Strategy Theory - Putting it all together

From the perspective of statistical arbitrage, the primary weakness of existing methodology is the lack of an integrated modelling framework which addresses the three modelling stages of intelligent pre-processing, predictive modelling, and decision implementation. This is a crucial failing because the practical advantage to be gained from any one of these stages is highly dependent upon the other two aspects: “statistical mispricings” are only useful if they are in some sense predictable, and “predictive information” is only useful if it can be successfully exploited. Conversely, the trading strategy is dependent upon the quality of the predictive information and the predictive model is reliant upon the (continuing) properties of the mispricing dynamics.
I now introduce an Integrated Algorithmic Trading Strategy Theory a framework on how to extract value from a given liquid market. We have shown in the previous chapters the theory and some applications using the techniques used in IATST.

15.1. Integrated Algorithmic Trading Strategy Theory – Definitions and techniques

Integrated Algorithmic Trading Strategy Theory is a modelling framework and a set of techniques with the ultimate goal of achieving best results in predictive modelling.
In the following sections we provide the modelling framework and the techniques.

15.2. Modelling Framework

1. Modelling the dynamics – Invariance. Advanced dynamics. Discrete / Continuous.
2. Estimation
3. Forecasting – Dynamics projection to the investment horizon. Pricing at the investment horizon.
4. Portfolio optimization

15.3. Techniques

1. Times series analysis
    - Statistics: Perform an exhaustive statistics descriptive analysis to the time series, in different frequencies. Chapters 3,4 and 5.
        - Mean, Median, standard deviation, kurtosis , asymmetry
        - Test the existence of regimes

        - Stationarity
        - Test econometric models : Univariate time series techniques
    - Test simple mean-reverting / trend following strategies in different frequencies. As we have seen so far there is not a simple unique definition of momentum, so try all the definitions
    - Non-linear techniques
2. Technical analysis indicators. Chapter 8.
3. Chaos. Chapter 7.
4. Multivariate techniques: Related assets Test cointegration and exploitable relationships with other assets. Chapter 5.
5. Microstructure / Macrostructure – Model the microstructure of the chosen market. Crowded / Non crowded strategies.
6. Event analysis: Earnings announcements, macroeconomic data, etc. Model the reaction to these events.
7. Derivatives: perform the same tests for derivatives linked to this asset: implied volatilities, futures, ADR, etc…Modelling techniques used for the volatility.
8. Factor analysis:
    1. External observable variables that may affect the price: valuation ratios, traded volume, volatility, order book etc..
    2. External non observable variables: PCA, Markov hidden models
9. Use the set of forecasting techniques given the characteristics Backtest. Understand the backtest. Start first with linear techniques, then non-linear techniques.
10. Use the set of strategies, optimizing the weights through mean variance, Conditional VAR. See Section 17.15.6 and 17.15.7.

15.4. Perfomance measurement

We use in this framework a the winning probability, luck coefficient and the Sharpe ration .
Defined in Section 11.4.

15.5. GOLD IATST

We show here an example of the application of the framework to gold.

- Gold appears to undertake “regime shifts” – periods of stable gains are followed by steep declines
- Gold is effectively a financial asset: storage costs are negligible - Physical commodities tend to have substantial storage costs, whereas financial assets have none
- Gold shows mean-reverting behaviour as we have seen in Chapter 13
- Gold price is affected by many independent drivers:
    - Macroeconomic outlook

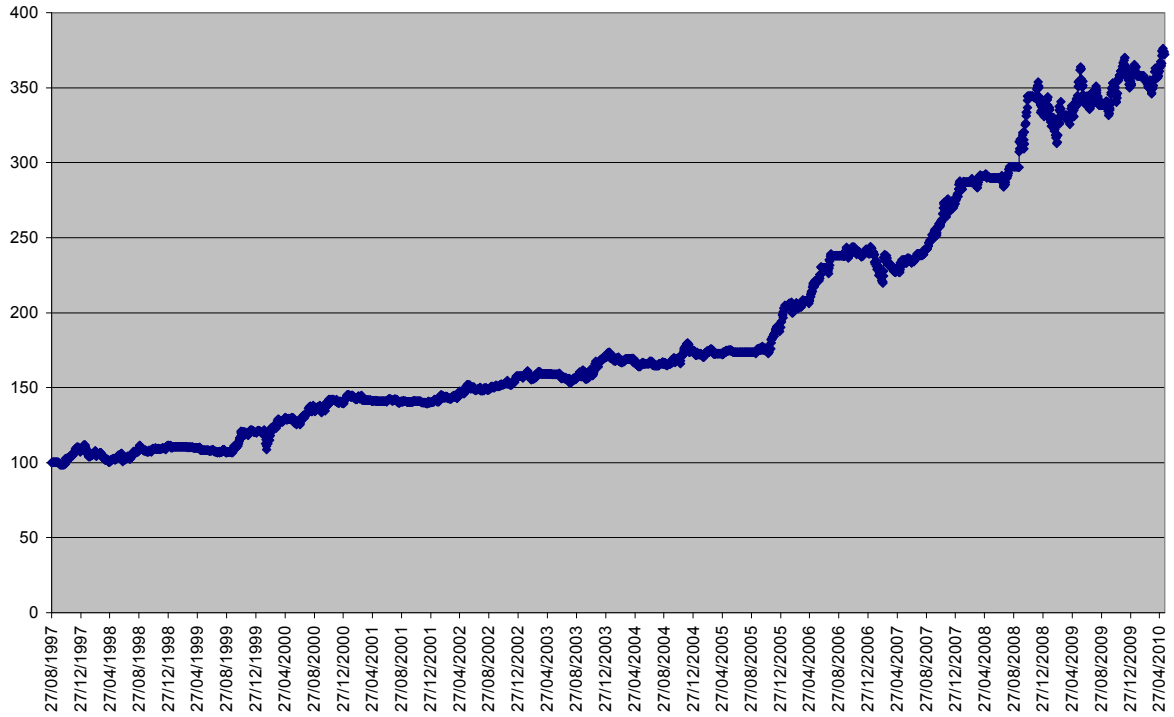
- Inflation
- Supply and Demand

- GOLD IATST Variables
  - Real Interest Rate- When real rates of return decrease or even fall negative, savers lose wealth and purchasing power. During these periods of high inflation and falling demand for the dollar and other major fiat currencies, gold acts as a safe haven and a store of value and often rises in price.
  - Indian Gold Premium- India is world's largest gold consumer due to its high demand for jewellery, which makes Mumbai's gold market very important as a whole (typically comprising over 20% of world demand). That's why large positive (negative) premiums in Mumbai's gold compared to COMEX gold are a good indicator of an increase (decrease) in gold price.
  - Open Interest- There exists a fixed quantity of gold in the world almost all of which is held by its owners as a tangible store of value. Therefore any fluctuation in open interest (total number of futures contracts not yet expired) is a very good indicator for moves in gold price. Rising open interest tends to precede rising gold prices and vice-versa.
  - Gold Mining Stocks - Investors usually take advantage of expected large movements in gold by investing in gold stocks which provide them with higher leverage. These large movements in Gold Mining stocks tend therefore to be leading indicators for moves in gold price.

Table 48: Backtesting all strategies and GOLD AITST – 8/1997-12/2009

	Real Interest Rate	Open Interest	Indian Premium	Gold Mining Stocks	GOLD AITST	Long Only Gold
Excess Return (p.a.)	5.23%	3.70%	4.05%	2.73%	10.73%	6.66%
Volatility	9.20%	8.20%	8.10%	9.00%	8.80%	17.10%
Sharpe Ratio	0.57	0.45	0.50	0.30	1.22	0.39
Calmar Ratio	0.32	0.19	0.29	0.17	1.38	0.20
Drawdown	16.23%	19.21%	13.92%	15.79%	7.77%	33.89%

Figure 77: GOLD IATST



15.6. Multi Strategy IATST

The Multi Strategy IATST consists of a portfolio of systematic algorithm based strategies designed to capture uncorrelated trends or risk premium across various financial markets.

Strategy weights have been determined to offer consistent risk profiles across the global strategy.

The underlying strategies are anticipated to have low risk profiles and are completely transparent in terms of rules and mechanisms, as well as being liquid and cost-efficient by virtue of being implemented through liquid underlying futures contracts in each asset class.

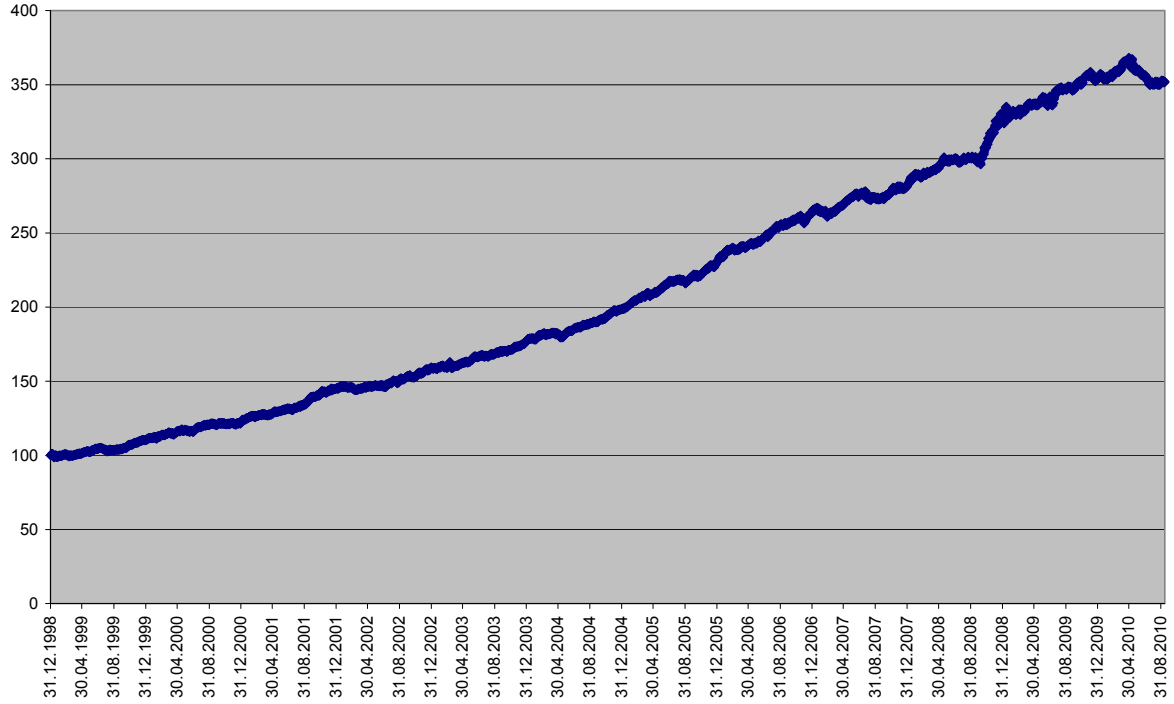
Combining these uncorrelated strategies with an already successful live track record and back-test that spans multiple business cycles, results in a portfolio expected to have strong risk/return characteristics.

I combine in these application 6 different strategies and several asset classes: 2 using momentum, 2 using risk premia and 2 using factors (economic indicators).

Table 49: Perfomance measures jan 1999-15/10/2010

	Multi Strategy IATST
Excess Return (p.a.)	11.30%
Volatility	3.70%
Sharpe Ratio	3.04
Calmar Ratio	2.24%
Drawdown	5.10%

Figure 78: Multi Strategy IATST



16. Conclusions

Markets are a complex system: not deterministic, not completely random and not chaotic either, we may fall into the curse of dimensionality in the multivariate context, so we face a formidable adversary.

One of the fundamental laws of the markets may be that the market dynamics are non stationary. Statistical behaviour changes in such a way that financial data from the last five years might have no bearing on the next five minutes. The markets are a moving target. They keep remaking themselves, switching gears, altering their premises.

The difference between systems that are predictable and systems that are not lies mostly in the numbers of degrees of freedom they possess. Markets' prices might never be released from human influence, and humans are herd animals who tend to follow relatively few fads and fashions. Investor preferences, trading lore, and human foibles appear to be one main source of both the pattern and the non stationarity in financial markets.

One of the conclusions of this PhD thesis is that there is no perfect application in financial forecasting.

There is no single best model for financial forecasting as we have seen in the different experiments: law of reversion, ARIMA, macroeconomic indicators and in the references (Chatfield

2000 , Weigend and Gershenfeld 1994 , Bass 1999), all techniques depend very much on the asset and the context and the features of the time series I are forecasting.

However there are good models for certain markets and contexts. We've seen in this PhD thesis some good models for some asset classes in some contexts. The markets might have stationary regimes where some structure appears and sits in a statistically stationary pattern but after some time it invariably disappears.

So the most important task is finding models that have the strongest signal and persist the longest like value and momentum.

Nonstationarity is present in different forms, the best way around the problem is to build ensembles / combinations of models, that keep generating/adapting models and selecting the "fittest" among them to track nonstationary targets.

They have to work on the principle of overlapping predictions which are "voted" together before deciding wether to buy or sell. Best practices in the real world (Renaissance, DE Shaw, MAN AHL) are the best proof of that.

I have introduced the Integrated Algorithmic Trading Strategy Theory a systematic framework to address the quantitative investing challenge. I propose a rigorous process testing the most appropriate techniques given an asset and a context (frequency and bet structure).

I show an application to Gold metal commodity with high ex-ante Sharpe ratios of 1.22. The Multistrategy IATST is another application of an algorithmic trading strategy combining 6 different stand-alone algorithmic strategies provides an excellent historical Sharpe ratio of 3.

The framework really helps us to identify good ensembles of models.

The contributions of this thesis are new theoretical concepts: the law of reversion, the integrated econometric / statistical framework (IATST), showing 2 interesting applications and the machine learning technique combining Genetic Algorithms and Artificial Neural Networks and support vector machines.

I also present a whole set of new experiments, regarding momentum, value and econometric tests to well known markets, some new models using moving averages showing excellent results in asymmetric markets, and two new machine learning experiments (GA+ANN , SVM).

Algorithmic Trading Strategy Theory integrates all this new concepts in a unified framework to address any market.

17. Appendix

17.1. Forecast error

The forecast error is the difference between the actual value and the forecast value for the corresponding period.

$$E_t = Y_t - F_t \quad (344)$$

where E is the forecast error at period t, Y is the actual value at period t, and F is the forecast for period t.

Measures of aggregate error:

Mean Absolute Error (MAE) $$MAE = \frac{\sum_{t=1}^{N} |E_t|}{N} \quad (345)$$

Mean Absolute Percentage Error (MAPE) $$MAPE = \frac{\sum_{t=1}^{N} \left| \frac{E_t}{Y_t} \right|}{N} \quad (346)$$

Percent Mean Absolute Deviation (PMAD) $$PMAD = \frac{\sum_{t=1}^{N} |E_T|}{\sum_{t=1}^{N} |Y_t|} \quad (347)$$

Mean squared error (MSE) $$MSE = \frac{\sum_{t=1}^{N} E_t^2}{N} \quad (348)$$

Root Mean squared error (RMSE) $$RMSE = \sqrt{\frac{\sum_{t=1}^{N} E_t^2}{N}} \quad (349)$$

17.2. Analytical proof of the law of reversion

Write $\boldsymbol{X} = \boldsymbol{P}_t$ (350) and $\boldsymbol{Y} = \boldsymbol{P}_{t-1}$ (351) to simplify notation. The two events:

$$\{\boldsymbol{X} < \boldsymbol{Y} \cap \boldsymbol{Y} > \boldsymbol{m}\} \text{and} \{\boldsymbol{X} > \boldsymbol{Y} \cap \boldsymbol{Y} < \boldsymbol{m}\} \quad (352)$$

are disjoint (easily seen from the fact that Y > m and Y < m cannot occur simultaneously: On the graph, regions (a) and (b) do not overlap), so the probability of the disjunction is simply the sum of the individual probabilities. Consider the first part of the disjunction:

$$\Pr[\boldsymbol{X} < \boldsymbol{Y} \cap \boldsymbol{Y} > \boldsymbol{m}] = \int_{\mathrm{m}}^{\infty} \int_{-\infty}^{\mathrm{y}} \mathrm{f}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{dxdy} \quad (353)$$

where fXY(x, y) denotes the joint density function of X and Y. By the assumption of independence, the joint density is just the product of the individual marginal densities, which in this case are identical (also by assumption). Denoting the marginal density generically by f (.) and its corresponding distribution by F(.), proceed as follows:

$$\begin{aligned} &\int_{\mathrm{m}}^{\infty} \int_{-\infty}^{\mathrm{y}} \mathrm{f}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{dxdy} = \int_{\mathrm{m}}^{\infty} \int_{-\infty}^{\mathrm{y}} \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{y}) \boldsymbol{dxdy} = \int_{\mathrm{m}}^{\infty} \boldsymbol{F}(\boldsymbol{y}) \boldsymbol{f}(\boldsymbol{y}) \boldsymbol{dy} \\ &= \int_{\mathrm{m}}^{\infty} \boldsymbol{F}(\boldsymbol{y}) \boldsymbol{dF}(\boldsymbol{y}) \end{aligned} \quad (354)$$

The last step is simply recognition that the density function of a random quantity is the analytic derivative of the corresponding distribution function. The remaining steps are trivial:

$$\begin{aligned} &\int_{\mathrm{m}}^{\infty} \boldsymbol{F}(\boldsymbol{y}) \boldsymbol{dF}(\boldsymbol{y}) = \frac{1}{2} \boldsymbol{F}(\boldsymbol{y})^2 \Big|_{\boldsymbol{m}}^{\infty} = \frac{1}{2} \left[\left(\lim_{\boldsymbol{t} \to \infty} \boldsymbol{F}(\boldsymbol{t}) \right)^2 - \boldsymbol{F}(\boldsymbol{m})^2 \right] \\ &= \frac{1}{2} \left[1 - \left(\frac{1}{2} \right)^2 \right] = \frac{3}{8} \end{aligned} \quad (355)$$

For the second part of the disjunction, the result follows from a similar argument after an initial algebraic simplification. First, note that the event Y < m may be expressed as the union of two disjoint events:

$$\boldsymbol{Y} < \boldsymbol{m} \equiv \{\boldsymbol{X} > \boldsymbol{Y} \cap \boldsymbol{Y} < \boldsymbol{m}\} \cup \{\boldsymbol{X} < \boldsymbol{Y} \cap \boldsymbol{Y} < \boldsymbol{m}\} \quad (356)$$

By definition (recall that m is the median of the distribution), the probability of the event Y < m is one half. Therefore, using the fact that probabilities for disjoint events are additive, we may write:

$$\Pr[\boldsymbol{X} > \boldsymbol{Y} \cap \boldsymbol{Y} > \boldsymbol{m}] = -\frac{1}{2} \Pr[\boldsymbol{X} < \boldsymbol{Y} \cap \boldsymbol{Y} < \boldsymbol{m}] \quad (357)$$

Now, proceeding much as for the first part:

$$\begin{aligned} &\Pr[\boldsymbol{X} > \boldsymbol{Y} \cap \boldsymbol{Y} > \boldsymbol{m}] = \frac{1}{2} - \int_{-\infty}^{\boldsymbol{m}} \int_{-\infty}^{\mathrm{y}} \mathrm{f}_{\mathrm{xy}}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{dxdy} = \frac{1}{2} - \int_{-\infty}^{\boldsymbol{m}} \boldsymbol{F}(\boldsymbol{y}) \boldsymbol{dF}(\boldsymbol{y}) \\ &= \frac{1}{2} - \frac{1}{2} \left[\boldsymbol{F}(\boldsymbol{m})^2 - \left(\lim_{\boldsymbol{t} \to \infty} \boldsymbol{F}(\boldsymbol{t}) \right)^2 \right] = \frac{1}{2} - \frac{1}{2} \left[\left(\frac{1}{2} \right)^2 \right] = \frac{3}{8} \end{aligned} \quad (358)$$

17.3. The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck process is defined by the dynamics:

$$dX_t = -\theta(X_t - \mu)dt + dL_t \quad (359)$$

We introduce the integrator

$$\boldsymbol{Y_t \equiv e^{\theta t}(X_t - \mu)} \quad (360)$$

Then from Ito's lemma

$$\boldsymbol{dY_t = \theta e^{\theta t}(X_t - \mu)dt + \theta e^{\theta t}(-\theta(X_t - \mu)dt + \sigma dB_t) = \sigma e^{\theta t} dB_t} \quad (361)$$

Integrating we obtain

$$\boldsymbol{Y_t = Y_0 + \sigma \int_o^t e^{\theta s} dB_s} \quad (362)$$

which from (73) becomes

$$\boldsymbol{X_t = (1 - e^{\theta t})\mu + e^{\theta t} X_0 + \sigma \int e^{-\theta(t-s)} dB_s} \quad (363)$$

Therefore

$$\boldsymbol{X_t | x_0 \sim N(\mu_t | x_0, \sigma_t^2 | x_0)} \quad (364)$$

where

$$\boldsymbol{\mu_t | x_0 \equiv (1 - e^{\theta t})\mu + e^{\theta t} x_0} \quad (365)$$

To compute $\sigma_t^2 | x_0$ we use Ito's isometry

$$\boldsymbol{\sigma_t^2 | x_0 = \sigma^2 \int_0^t e^{-2\theta(t-s)} ds = \sigma^2 \int_0^t e^{-2\theta u} du = \sigma^2 \frac{e^{-2\theta u}}{-2\theta u}\Big|_0^t = \frac{\sigma^2}{2\theta}(1 + e^{-2\theta t})} \quad (366)$$

Similarly for the autocovariance

$$\sigma_t^2|x_0 \equiv Cov\{X_t, X_{t+\tau}|x_0\} = E\left\{\left(e^{-2\theta t}\int_o^t e^{-\theta s}dB_s\right)\left(\sigma e^{-\theta(t+\tau)}\int_0^{t+\tau} e^{\theta s}dBs\right)\right\}$$
$$= \sigma^2 e^{-2\theta t}e^{-\theta t}\int_0^t e^{2\theta s}ds$$
$$= \sigma^2 e^{-2\theta t}e^{-\theta t}\frac{e^{2\theta u}}{2\theta u}\Big|_0^t \quad (367)$$
$$= \frac{\sigma^2}{2\theta}e^{-\theta t}\left(1+e^{-2\theta t}\right)$$

Therefore the autocorrelation reads

$$Cor\{X_t, X_{t+\tau}|x_0\} = \frac{\sigma_{t,\tau}^2|x_0}{\sqrt{\sigma_t^2|x_0}\sqrt{\sigma_{t-\tau}^2|x_0}} = e^{-\theta\tau}\sqrt{\frac{1-e^{-2\theta t}}{1-e^{-2\theta(t+\tau)}}} \quad (368)$$

which in the limit t → ∞ becomes the unconditional autocorrelation

$$Cor\{X_t, X_{t+\tau}\} = e^{-\theta\tau} \quad (369)$$

17.4. Relation between CIR and OU processes

Consider the OU process with null unconditional expectation

$$dX_t = -\theta X_t dt + \sigma dB_t \quad (370)$$

Using Ito's rule

$$d\left(X_t^2\right) = 2X_t dX_t + \left(dX_t\right)^2 \quad (371)$$

we obtain

$$dY_t = 2\sqrt{Y_t}\left(-\theta\sqrt{Y_t}dt + \sigma dB_t\right) + \sigma^2 dt = \left(-2\theta Y_t + \sigma^2\right)dt + 2\sqrt{Y_t}\sigma dB_t \quad (372)$$

which is a specific instance of the general CIR process (59). More in general, following Vicente, de Toledo, P., and Caticha (2005), we consider several independent OU processes

$$dX_{j,t} = -\theta X_{j,t}dt + \sigma dB_{j,t} \quad (373)$$

whose square evolves as follows:

$$d\left(X_{j,t}\right)^2 = 2X_{j,t}\left(-\theta X_{j,t}dt + \sigma dB_{j,t}\right) + \sigma^2 dt = 2\theta X_{j,t}^2 dt + 2\sigma X_{j,t}dB_{j,t} + \sigma^2 dt \quad (374)$$

Then $v_t \equiv \sum_{j=1}^{J} X_{j,t}^2$ evolves as a CIR process

$$dv_t = \left(J\sigma^2 - 2\theta v_t\right)dt + 2\sigma\sqrt{v_t}dB_{j,t} \quad (375)$$

where in the last step we used the isometry

$$\sum_{j=1}^{J} X_{j,t}dB_{j,t} \overset{d}{=} \sqrt{\sum_{j=1}^{J} X_{j,t}^2 dB_t} \quad (376)$$

17.5. The Levy-Khintchine representation of a Levy process

First we compute the characteristic function of the Poisson process (41), which follows from (14):

$$\begin{aligned} \phi_{\varepsilon,t}(w) &\equiv E\left\{e^{iw\varepsilon t}\right\} = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{-(\lambda t)} e^{iwk\Delta} \\ &= e^{-(\lambda t)} \sum_{k=0}^{\infty} \frac{\left(\lambda t e^{iw\Delta}\right)^k}{k!} = e^{-(\lambda t)} \lambda t e^{iw\Delta} = e^{-(\lambda t)}\left(e^{iw\Delta} - 1\right) \end{aligned} \quad (377)$$

Now the generic representation of a Levy process (47), which we report here:

$$X_t = B_t^{\pi,\sigma^2} + \int_{-\infty}^{+\infty} P_t^{\Delta,\lambda(\Delta)} d\Delta \quad (378)$$

Its characteristic function reads

$$E\left\{e^{iwX_t}\right\} = E\left\{e^{iw\left(\pi t + \sigma B_t + \int_{-\infty}^{+\infty} P_t^{\Delta,\lambda(\Delta)} d\Delta\right)}\right\} \quad (379)$$

Approximating the integral with sums we obtain

$$\ln\left(E\left\{e^{iwX_t}\right\}\right)=iw\pi t-\frac{1}{2}\sigma^2 tw^2+\ln\left(E\left\{e^{iw\left(\sum_k P_t^{\Delta,\lambda(\Delta)}\right)}\right\}\right)$$
$$=iw\pi t-\frac{1}{2}\sigma^2 tw^2+\ln\prod_k E\left\{e^{iwP_t^{\Delta,\lambda(\Delta)}}\right\}$$
$$=iw\pi t-\frac{1}{2}\sigma^2 tw^2+\sum_k \ln e^{\lambda_k t\left(e^{iw\Delta k}-1\right)} \quad (380)$$
$$=iw\pi t-\frac{1}{2}\sigma^2 tw^2+t\sum_k {}^{\lambda_k\left(e^{iw\Delta k}-1\right)}$$
$$=iw\pi t-\frac{1}{2}\sigma^2 tw^2+t\int_{-\infty}^{+\infty}\lambda(\Delta)\left(e^{iw\Delta k}-1\right)d\Delta$$

Reverting from the sum to the integral notation we obtain

$$\ln\left(E\left\{e^{iwX_t}\right\}\right)=\left(iw\pi-\frac{1}{2}\sigma^2 w^2+\int_{-\infty}^{+\infty}\lambda(\Delta)\left(e^{iw\Delta k}-1\right)d\Delta\right)t \quad (381)$$

where the intensity λ (Δ) determines the relative weight of the discrete Poisson jumps on the grid ΔN.

17.6. Representation of compound Poisson process

Consider the general representation (50) of a Poisson process. We do not need to worry about the diffusive portion, so we only consider the jump portion

$$X_t=\sum_{n=1}^{P_t^\lambda} Z_n \quad (382)$$

The characteristic function reads

$$E\left\{e^{iwX_t}\right\}=E_P\left\{E_Z\left\{e^{\sum_{n=1}^{P_t^\lambda} iwZ_n}\right\}\right\}$$
$$=E_P\left\{\theta_Z(w)^{P_t^\lambda}\right\}=E_P\left\{e^{iyP_t^\lambda}\right\}\Big|_{\theta_{z(w)}=e^{i\lambda}} \quad (383)$$
$$=e^{t\lambda(\phi_Z(w)-1)}$$

where in the last step we used (90). Suppose now that in (94) we set μ ≡ σ2 ≡ 0 and λ (Δ) ≡ λfZ (Δ). Then (94) becomes

$$\ln\left(\boldsymbol{E}\left\{\boldsymbol{e}^{\boldsymbol{iwX}_t}\right\}\right)=\left(\int_{-\infty}^{+\infty}\boldsymbol{f}_{\boldsymbol{Z}}(\Delta)\boldsymbol{e}^{\boldsymbol{iw}\Delta}\boldsymbol{d}\Delta+\int_{-\infty}^{+\infty}\boldsymbol{f}_{\boldsymbol{Z}}(\Delta)\boldsymbol{d}\Delta\right)\boldsymbol{t}\lambda \quad (384\text{-}385)$$
$$=\boldsymbol{t}\lambda\left(\phi_{\boldsymbol{Z}}(\Delta)-1\right)$$

which is (96).

17.7. Deterministic linear dynamical system

Consider the dynamical system

$$\begin{pmatrix}\dot{z}_1\\ \dot{z}_2\end{pmatrix}=\begin{pmatrix}a & -b\\ b & a\end{pmatrix}\begin{pmatrix}z_1\\ z_2\end{pmatrix} \quad (386)$$

To compute the solution, consider the auxiliary complex problem

$$\dot{z}=\mu z \quad (387)$$

where $z\equiv z_1+\boldsymbol{i}z_2$ and $\mu\equiv\boldsymbol{a}+\boldsymbol{ib}$. By isolating the real and the imaginary parts we realize that (143) and (144) coincide.
The solution of (144) is

$$z(t)=e^{\mu t}z(0) \quad (388)$$

Isolating the real and the imaginary parts of this solution we obtain:

$$z_1(t)=\mathrm{Re}\left(e^{\mu t}z^0\right)=\mathrm{Re}\left(e^{(a+ib)t}\left(z_1(0)+iz_2(0)\right)\right) \quad (389)$$
$$=\mathrm{Re}\left(e^{at}\left(\cos bt+i\sin bt\right)\left(z_1(0)+iz_2(0)\right)\right)$$

$$z_2(t)=\mathrm{Im}\left(e^{\mu t}z^0\right)=\mathrm{Re}\left(e^{(a+ib)t}\left(z_1(0)+iz_2(0)\right)\right) \quad (390)$$
$$=\mathrm{Im}\left(e^{at}\left(\cos bt+i\sin bt\right)\left(z_1(0)+iz_2(0)\right)\right)$$

or

$$z_1(t)=e^{at}\left(z_1(0)\cos bt-z_2(0)\sin bt\right) \quad (391)$$

$$z_2(t)=e^{at}\left(z_1(0)\cos bt-z_2(0)\sin bt\right) \quad (392)$$

The trajectories (148)-(149) depart from (shrink toward) the origin at the exponential rate a and turn counterclockwise with frequency b.

17.8. OU (auto)covariance: general expression

We recall that the OU process integrates as follows

$$X_t = e^{-\Theta t} X_0 + \left(I - e^{-\Theta t}\right)m + \int_0^t e^{\Theta(u-t)} S dB_u \quad (393)$$

To compute the covariance we use Ito's isometry:

$$Cov\left\{X_t \middle| x_0\right\} = \int_0^t e^{\Theta(u-t)} \Sigma e^{\Theta'(u-t)} du \quad (394)$$

where $\boldsymbol{\Sigma} \equiv \mathbf{SS}'$. We can simplify further this expression as in Van der Werf (2007). Using the identity

$$\mathbf{vec}(\mathbf{ABC}) \equiv \left(\mathbf{C}' \otimes \mathbf{A}\right)\mathbf{vec}(\mathbf{B}) \quad (395)$$

where vec is the stack operator and $\otimes$ is the Kronecker product. Then

$$vec\left(e^{\Theta(u-t)} \Sigma e^{\Theta'(u-t)}\right) = \left(e^{\Theta(u-t)} \otimes e^{\Theta(u-t)}\right) vec(\Sigma) \quad (396)$$

Now we can use the identity

$$e^{A \oplus B} = e^A \otimes e^B \quad (397)$$

where $\oplus$ is the Kronecker sum

$$\mathbf{A}_{\mathbf{MxM}} \oplus \mathbf{B}_{\mathbf{NxN}} \equiv \mathbf{A}_{\mathbf{MxM}} \otimes \mathbf{I}_{\mathbf{NxN}} + \mathbf{I}_{\mathbf{MxM}} \otimes \mathbf{B}_{\mathbf{NxN}} \quad (398)$$

Then we can rephrase the term in the integral (153) as

$$vec\left(e^{\Theta(u-t)} \Sigma e^{\Theta'(u-t)}\right) = \left(e^{(\Theta \oplus \Theta)(u-t)}\right) vec(\Sigma) \quad (399)$$

Substituting this in (151) we obtain

$$
\begin{aligned}
vec\left(Cov\left\{X_t \middle| x_0\right\}\right) &= \left(\int_0^t e^{(\Theta\oplus\Theta)(u-t)}du\right)vec(\Sigma) \\
&= (\Theta\oplus\Theta)^{-1}e^{(\Theta\oplus\Theta)(u-t)}\Big|_0^t vec(\Sigma) \\
&= (\Theta\oplus\Theta)^{-1}\left(I - e^{-(\Theta\oplus\Theta)(u-t)}\right)vec(\Sigma)
\end{aligned}
\qquad (400)
$$

Similar steps lead to the expression of the autocovariance.

17.9. OU (auto)covariance: explicit solution

For t ≤ z the autocovariance is

$$
\begin{aligned}
&\boldsymbol{Cov}\left\{\boldsymbol{Z}_t, \boldsymbol{Z}_{t+\tau} \mid \boldsymbol{Z}_0\right\} = \boldsymbol{E}\left\{\left(\boldsymbol{Z}_t - \boldsymbol{E}\left\{\boldsymbol{Z}_t \mid \boldsymbol{Z}_0\right\}\right)\left(\boldsymbol{Z}_{t+\tau} - \boldsymbol{E}\left\{\boldsymbol{Z}_{t+\tau} \mid \boldsymbol{Z}_0\right\}\right)' \middle| \boldsymbol{Z}_0\right\} \\
&= \boldsymbol{E}\left\{\left(\int_0^t \boldsymbol{e}^{\Gamma(u-t)}\boldsymbol{V}d\boldsymbol{B}_u\right)\left(\int_0^{t+\tau} d\boldsymbol{B}_u\boldsymbol{V}'\boldsymbol{e}^{\Gamma'(u-t)}\right)'\right\}\boldsymbol{e}^{-\Gamma'\tau} \\
&= \boldsymbol{E}\left\{\left(\int_0^t \boldsymbol{e}^{\Gamma(u-t)}\boldsymbol{V}d\boldsymbol{B}_u\right)\left(\int_0^{t} d\boldsymbol{B}_u\boldsymbol{V}'\boldsymbol{e}^{\Gamma'(u-t)}\right)'\right\}\boldsymbol{e}^{-\Gamma'\tau} \\
&= \left(\int_0^t \boldsymbol{e}^{\Gamma(u-t)}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{e}^{\Gamma'(u-t)}\boldsymbol{du}\right)\boldsymbol{e}^{-\Gamma'\tau} \\
&= \left(\int_0^t \boldsymbol{e}^{-\Gamma s}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{e}^{-\Gamma' s}\boldsymbol{du}\right)\boldsymbol{e}^{-\Gamma'\tau} \\
&= \left(\int_0^t \boldsymbol{e}^{-\Gamma s}\Sigma\boldsymbol{e}^{-\Gamma' s}\boldsymbol{ds}\right)\boldsymbol{e}^{-\Gamma'\tau}
\end{aligned}
\qquad (401)
$$

$$
\Sigma \equiv VV' = A^{-1}SSA'^{-1} \qquad (402)
$$

In particular, the covariance reads

$$
\Sigma(t) \equiv Cov\left\{Z_t \mid Z_0\right\} = \int_0^t e^{-\Gamma s}\Sigma e^{-\Gamma' s}ds \qquad (403)
$$

To simplify the notation, we introduce three auxiliary functions:

$$E_\gamma(t) \equiv \frac{1}{\gamma} - \frac{e^{-\gamma t}}{\gamma} \quad (404)$$

$$D_{\gamma,\omega}(t) \equiv \frac{\omega}{\gamma^2+\omega^2} - \frac{e^{-\gamma t}(\omega\cos\omega t + \gamma\sin\omega t)}{\gamma^2+\omega^2} \quad (405)$$

$$D_{\gamma,\omega}(t) \equiv \frac{\omega}{\gamma^2+\omega^2} - \frac{e^{-\gamma t}(\omega\cos\omega t + \gamma\sin\omega t)}{\gamma^2+\omega^2} \quad (406)$$

Using (161)-(163), the conditional covariances among the entries $\boldsymbol{k} = 1,...\boldsymbol{K}$ relative to the real eigenvalues read:

$$\begin{aligned} &\sum_{k,\tilde{k}}(t) = \int_0^t e^{-\lambda_k} \sum_{k,\tilde{k}} e^{-\lambda_{\tilde{k}} s} ds \\ &= \sum_{k,\tilde{k}} E(t;\lambda_k + \lambda_{\tilde{k}}) \end{aligned} \quad (407)$$

Similarly, the conditional covariances (160) of the pairs of entries $\boldsymbol{j} = 1,...\boldsymbol{J}$ relative to the complex eigenvalues with the entries $\boldsymbol{k} = 1,...\boldsymbol{K}$ relative to the real eigenvalues read:

$$\begin{aligned} &\begin{pmatrix} \sum_{j,k}^{(1)}(t) \\ \sum_{j,k}^{(2)}(t) \end{pmatrix} = \int_0^t e^{-\Gamma_{jS}} \begin{pmatrix} \sum_{j,k}^{(1)} \\ \sum_{j,k}^{(1)} \end{pmatrix} e^{-\lambda_k s} ds \\ &= \int_0^t e^{-(\gamma_j+\lambda_k)s} \begin{pmatrix} \sum_{j,k}^{(1)} \cos(\omega_j s) - \sum_{j,k}^{(2)} \sin(\omega_j s) \\ \sum_{j,k}^{(1)} \sin(\omega_j s) + \sum_{j,k}^{(2)} \cos(\omega_j s) \end{pmatrix} ds \end{aligned} \quad (408)$$

which implies

$$\begin{aligned} &\sum_{j,k}^{(1)}(t) \equiv \sum_{j,k}^{(1)} \int_0^t e^{-(\gamma_j+\lambda_k)^s} \cos(\omega_j s) ds \\ &- \sum_{j,k}^{(2)} \int_0^t e^{-(\gamma_j+\lambda_k)^s} \sin(\omega_j s) ds \\ &= \sum_{j,k}^{(1)} C(t;\gamma_j+\lambda_k,\omega_j) - \sum_{j,k}^{(2)} S(t;\gamma_j+\lambda_k,\omega_j) \end{aligned} \quad (409)$$

$$\begin{aligned} &\sum_{j,k}^{(2)}(t) \equiv \sum_{j,k}^{(1)} \int_0^t e^{-(\gamma_j+\lambda_k)^s} \sin(\omega_j s) ds \\ &+ \sum_{j,k}^{(2)} \int_0^t e^{-(\gamma_j+\lambda_k)^s} \cos(\omega_j s) ds \\ &= \sum_{j,k}^{(1)} S(t;\gamma_j+\lambda_k,\omega_j) + \sum_{j,k}^{(2)} C(t;\gamma_j+\lambda_k,\omega_j) \end{aligned} \quad (410)$$

Finally, using again (161)-(163), the conditional covariances (160) among the pairs of entries $j = 1,...J$ relative to the complex eigenvalues read:

$$\begin{pmatrix} \sum_{j,\tilde{j}}^{(1,1)}(t) & \sum_{j,\tilde{j}}^{(1,2)}(t) \\ \sum_{j,\tilde{j}}^{(2,1)}(t) & \sum_{j,\tilde{j}}^{(2,2)}(t) \end{pmatrix} = \int_0^t e^{-\Gamma_j s} \begin{pmatrix} \sum_{j,\tilde{j}}^{(1,1)} & \sum_{j,\tilde{j}}^{(1,2)} \\ \sum_{j,\tilde{j}}^{(2,1)} & \sum_{j,\tilde{j}}^{(2,2)} \end{pmatrix} e^{-\Gamma_{\tilde{j}}' s} ds \,|$$
$$= \int_0^t e^{-(\gamma_j + \gamma_{\tilde{j}})s} \begin{pmatrix} \cos\omega_j s & -\sin\omega_j s \\ \sin\omega_j s & \cos\omega_j s \end{pmatrix} \tag{411}$$
$$\begin{pmatrix} \sum_{j,\tilde{j}}^{(1,1)} & \sum_{j,\tilde{j}}^{(1,2)} \\ \sum_{j,\tilde{j}}^{(2,1)} & \sum_{j,\tilde{j}}^{(2,2)} \end{pmatrix} \begin{pmatrix} \cos\omega_{\tilde{j}} s & \sin\omega_{\tilde{j}} s \\ -\sin\omega_{\tilde{j}} s & \cos\omega_{\tilde{j}} s \end{pmatrix} ds$$

Using the identity

$$\begin{pmatrix} \tilde{a} & \tilde{b} \\ \tilde{c} & \tilde{d} \end{pmatrix} \equiv \begin{pmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{pmatrix} \tag{412}$$

where

$$\tilde{a} \equiv \frac{1}{2}(a+d)\cos(\alpha-\omega) + \frac{1}{2}(c-b)\sin(\alpha-\omega)$$
$$+ \frac{1}{2}(a-d)\cos(\alpha+\omega) - \frac{1}{2}(b+c)\sin(\alpha+\omega) \tag{413}$$

$$\tilde{b} \equiv \frac{1}{2}(b-c)\cos(\alpha-\omega) + \frac{1}{2}(a+d)\sin(\alpha-\omega)$$
$$+ \frac{1}{2}(b+c)\cos(\alpha+\omega) + \frac{1}{2}(a-d)\sin(\alpha+\omega) \tag{414}$$

$$\tilde{c} \equiv \frac{1}{2}(c-b)\cos(\alpha-\omega) - \frac{1}{2}(a+d)\sin(\alpha-\omega)$$
$$+ \frac{1}{2}(b+c)\cos(\alpha+\omega) + \frac{1}{2}(a-d)\sin(\alpha+\omega) \tag{415}$$

$$\tilde{d} \equiv \frac{1}{2}(a+d)\cos(\alpha+\omega) + \frac{1}{2}(c-b)\sin(\alpha-\omega)$$
$$+ \frac{1}{2}(d-a)\cos(\alpha+\omega) + \frac{1}{2}(b+c)\sin(\alpha+\omega) \tag{416}$$

we can simplify the matrix product in (167). Then, the conditional covariances among the pairs of entries $j = 1,...J$ relative to the complex eigenvalues read

$$
\begin{aligned}
S^{(1,1)}_{j,\tilde{j};t} \equiv & \frac{1}{2}\left(S^{(1,1)}_{j,\tilde{j}} + S^{(2,2)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(2,1)}_{j,\tilde{j}} + S^{(1,2)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(1,1)}_{j,\tilde{j}} - S^{(2,2)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t) \\
& - \frac{1}{2}\left(S^{(1,2)}_{j,\tilde{j}} + S^{(2,1)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t)
\end{aligned}
\tag{417}
$$

$$
\begin{aligned}
S^{(1,2)}_{j,\tilde{j};t} \equiv & \frac{1}{2}\left(S^{(1,2)}_{j,\tilde{j}} - S^{(2,1)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(1,1)}_{j,\tilde{j}} + S^{(2,2)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(1,2)}_{j,\tilde{j}} + S^{(2,1)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(1,1)}_{j,\tilde{j}} - S^{(2,2)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t)
\end{aligned}
\tag{418}
$$

$$
\begin{aligned}
S^{(2,1)}_{j,\tilde{j};t} \equiv & \frac{1}{2}\left(S^{(2,1)}_{j,\tilde{j}} - S^{(1,2)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& - \frac{1}{2}\left(S^{(1,1)}_{j,\tilde{j}} + S^{(2,2)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(1,2)}_{j,\tilde{j}} + S^{(2,1)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(1,1)}_{j,\tilde{j}} + S^{(2,2)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t)
\end{aligned}
\tag{419}
$$

$$
\begin{aligned}
S^{(2,2)}_{j,\tilde{j};t} \equiv & \frac{1}{2}\left(S^{(1,1)}_{j,\tilde{j}} + S^{(2,2)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(2,1)}_{j,\tilde{j}} - S^{(1,2)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}-\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(2,2)}_{j,\tilde{j}} - S^{(1,1)}_{j,\tilde{j}}\right) C_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t) \\
& + \frac{1}{2}\left(S^{(1,2)}_{j,\tilde{j}} + S^{(2,1)}_{j,\tilde{j}}\right) D_{\gamma_j+\gamma_{\tilde{j}},\omega_{\tilde{j}}+\omega_j}(t)
\end{aligned}
\tag{420-421}
$$

Similar steps lead to the expression of the autocovariance

17.10. The sup-Wald test

A common test of parameter stability is the Chow test. It assumes that at most one structural break occurred in the data generating process (i.e. multiple breaks are not explicitly taken into account) and that the potential break date is known in advance. In most financial applications, the second assumption is very unlikely to hold. An exception may be represented by a genuinely exogenous change, for example an abrupt change in regulation or in the infrastructure of a market.

In the case of the simple regression model considered in the text, the equations can be written as

$$\begin{aligned} r_t &= \alpha_1 + \beta_1 r_{Mt} + \varepsilon_t \quad t < \tau \\ r_t &= \alpha_2 + \beta_2 r_{Mt} + \varepsilon_t \quad t \geq \tau \end{aligned} \tag{422}$$

The test consists simply of an F test of the joint hypothesis that α1=α2 and β1=β2. Andrews (1993)[200] suggested an extension of this testing methodology which entails computing the F test statistic for any potential break date τ =τ ,τ +1,K,τ where τ and τ should leave enough observations at the beginning and at the end of the sample period to identify the change in parameter. Typically τ = [0.15T] and τ = [0.85T], where T is the sample size and [x] is the integer that is nearest to x. The test statistic advocated by Andrews is just the maximum of all F-stats, which follows under the null hypothesis a non-standard distribution, tabulated in Andrews (2003)[201].

As we state in the text, the test is known to have good power properties against the alternative hypotheses of multiple breaks and gradual change in the parameter values.

17.11. Power of the stability test

We ran a small Monte Carlo study to evaluate the power of the sup Wald test against a specific alternative: α dropping to zero in January 2001.

In particular, we simulated 100 series which were identical to the actual return data up to December 2000 and drawn from a normal distribution thereafter. The expected return for the simulated observations was set to zero, while the volatility was chosen to match the real data in the final part of the sample period.

Table 50 displays the results. For the value – momentum combination the test is very powerful: we would correctly reject the null 91% of the time at a significance level of 10% or 79% of the time at a significance level of 5%. Even better power properties are obtained when working with sector neutral returns: at the 5% confidence level the power goes up to almost 90%.

The intuition behind these results is that a breakdown in a strong signal can be easily detected. The Sharpe ratios displayed in Table 5 suggest that the combined value – momentum strategies generate indeed a strong signal (the mean is large compared to the variability of the return series). As a result, if the alpha of simple quant strategies had disappeared in 2001 then the available data would have been sufficient to identify the break through the sup Wald test.

Table 50: Power of the test against the hypothesis that alpha shrinks to zero

	Significance level		
	10%	**5%**	**1%**
50-50 Combination	**0.91**	**0.26**	**0.51**
50-50 Combination (SN)	**0.94**	**0.89**	**0.72**

17.12. Hedge fund returns and market neutrality

The fund data consisted of just over 3300 funds, of which 95 explicitly claimed to be market neutral, and monthly returns over the ten years from July 1999 to June 2009. Given the high data requirements of these tests we restricted our analysis to funds that have at least 24 monthly returns. All of this data came from HFR. The market returns consisted of two time series: the MSCI developed world index monthly returns were used as the market returns for non- US funds, and the S&P500 monthly returns for the US funds.

The market returns and many of the fund returns exhibited auto-correlation which would have invalidated the tests that we undertook. To avoid these issues we replaced the fund returns with the residuals from an MA(2) model and similarly for the market returns throughout our analysis.

Given the limited number of months we have data for, and the non-Gaussian distribution of the fund returns, we decided to use bootstrapping to try to apply Patton's tests meaningfully.

 For each fund, we considered the months for which we had returns data and looked at the returns to the market (either the US for the US funds, or the MSCI world index for funds outside the US) over the same months.

 We then used Politis and White's14 algorithm to choose the most appropriate length of bootstrap samples for the fund and for the market.

 We randomly and independently chose samples with lengths geometrically distributed around the corresponding lengths for the market and the fund returns until we had two samples, one for the market and one for the fund, with the same length as the original set of fund returns. These two samples should be independent, but have, on average, very similar distributions to the original fund and market returns.

 We calculated the relevant test statistic between the two samples (please see below for the definitions of the test statistic for each version of neutrality).

 We repeated this 1000 times, to get 1000 statistics showing the relationship between the market samples and the fund samples. We computed the 95^{th} percentile of these statistics and used this figure as our critical value.

 We calculated the statistic for the original, complete fund and market returns and compared it to the 95th percentile. If this statistic was higher than the percentile, we rejected our null hypothesis of neutrality.

Description of the data

Bootstrapping technique

We used the t-statistic of the correlation between the fund returns and the market returns as our statistic in our test for correlation neutrality.

The statistic we used to test for downside mean neutrality took several steps to compute:

i) We regressed the fund returns onto the market returns, their square and their cube, to calculate the coefficients β1, β2 or β3.

$$\mu_i(r_{mt}) = \beta_o + \beta_1 r_{mt} + \beta_2 r_{mt}^2 + \beta_3 r_{mt}^3 \quad (423)$$

ii) We evaluated the expression below with these coefficients:

$$E\left[\frac{\partial \mu_i(r_{mt})}{\partial r_{mt}}\right] = \hat{B}_1 + 2\hat{B}_2 E\left[r_{mt} | r_{mt} \le 0\right] + 3\hat{B}_3 E\left[r_{mt}^2 | r_{mt} \le 0\right] \quad (424)$$

iii) We calculated the standard error of the expression using the covariance matrix for the βi.

iv) We divided the expression through by the standard error.

Our results are in tables Table 51 and Table 52 below. A high proportion of funds that claim market neutrality as an explicit part of their strategy still fail one of these two tests of neutrality.

Table 51: Correlation neutrality

	Reject neutrality	Fail to reject neutrality	Overall
Fund does not claim neutrality	**2070**	**1179**	**3249**
Funds does claim neutrality	**26**	**69**	**95**
Overall	**2096**	**1248**	**3344**

Source: Own calculations

Table 52: Downside mean neutrality

	Reject neutrality	Fail to reject neutrality	Overall
Fund does not claim neutrality	**1743**	**1506**	**3249**
Funds does claim neutrality	**22**	**73**	**95**
Overall	**1765**	**1579**	**3344**

Source: Own calculations

17.13. Akaike information criterion

Akaike's information criterion, developed by Hirotsugu Akaike under the name of "an information criterion" (AIC) in 1971 and proposed in Akaike (1974) , is a measure of the goodness of fit of an estimated statistical model. It is grounded in the concept of entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality and can be said to describe the tradeoff between bias and variance in model construction, or loosely speaking that of accuracy and complexity of the model.

The AIC is not a test of the model in the sense of hypothesis testing; rather it is a test between models - a tool for model selection. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best. From the AIC value one

may infer that e.g. the top three models are in a tie and the rest are far worse, but it would be arbitrary to assign a value above which a given model is 'rejected'.

Definition

In the general case, the AIC is

$$AIC = 2k - 2\ln(L) \quad (425)$$

where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model.

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \quad (426)$$

Over the remainder of this entry, it will be assumed that the model errors are normally and independently distributed. Let n be the number of observations and RSS be the residual sum of squares. We further assume that the variance of the model errors is unknown but equal for them all. Maximizing the likelihood with respect to this variance, the AIC becomes

$$AIC = 2k + n\left[\ln(2\pi RSS / n) + 1\right] \quad (427)$$

This can be simplified by factoring out the term n * ln(2π). This is a constant term added to the AIC value of all the competing models. Therefore it can't affect the order in which we rank them and we can safely remove this term. When we also factor out the constant n, AIC simplifies to:

$$AIC = 2k + n\left[\ln(RSS)\right] \quad (428)$$

Increasing the number of free parameters to be estimated improves the goodness of fit, regardless of the number of free parameters in the data generating process. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting. The preferred model is the one with the lowest AIC value. The AIC methodology attempts to find the model that best explains the data with a minimum of free parameters. By contrast, more traditional approaches to modelling start from a null hypothesis. The AIC penalizes free parameters less strongly than does the Schwarz criterion.

AIC judges a model by how close its fitted values tend to be to the true values, in terms of a certain expected value. But it is important to realize that the AIC value assigned to a model is only meant to rank competing models and tell you which is the best among the given alternatives. The absolute values of the AIC for different models have no meaning; only relative differences can be ascribed meaning.

Relevance to χ2 fitting (maximum likelihood)

Often, one wishes to select amongst competing models where the likelihood function assumes that the underlying errors are normally distributed. This assumption leads to χ2 data fitting.

For any set of models where the data points are used, one can use a slightly altered AIC. For the purposes of this article, this will be called . It differs from the AIC only through an additive constant, which is a function only of the data points and not of the model. As only differences in the AIC are relevant, this constant can be ignored.

For χ2 fitting, the likelihood is given by

$$L = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma_i^2} \right)^{\frac{1}{2}} \exp\left(-\sum_{i=1}^{n} \frac{\left(y_i - f(x)^2\right)}{2\sigma_i^2} \right)$$

$$\therefore \quad \ln L = \ln\left(\prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma_i^2} \right)^{\frac{1}{2}} \right) - \frac{1}{2} \sum_{i=1}^{n} \frac{\left(y_i - f(x)^2\right)}{2\sigma_i^2} \quad (429)$$

$$\therefore \quad \ln L = C - \chi^2 / 2$$

, where C is a constant independent of the model used, and dependent only on the use of particular data points. i.e. it does not change if the data do not change.

The AIC is therefore given by $AIC = 2k - 2\ln(L) = 2k - 2\left(C - \chi^2 / 2\right) = 2k - 2C + \chi^2$ (430)

. As only differences in AICc are meaningful, the constant C can be omitted provided the same data points are used, giving

$$AIC_{\chi^2} = \chi^2 + 2k \quad (431)$$

This form is often convenient in that data fitting programs produce χ2 as a statistic for the fit. For models with the same number of data points, the one with the lowest should be preferred.

Similarly, if one has available the statistic R2 ("Variance Explained"), one may write

$$AIC_{R^2} = n \ln \frac{1 - R^2}{n} + 2k \quad (432)$$

The Pearson correlation r = R is a special case of this. Here, independence of the observations is assumed.

17.14. Genetic algorithms and technical analysis

17.14.1Coding Trading Strategies

Based on the trading formulation (261), to encode a trading strategy, we only need to encode the CONDS with three primitive predicates, which means the following three parts:

- $\vec{a} = (a_1, a_2, a_3)$
- $\vec{\oplus} = (\oplus_1, \oplus_2, \oplus_3)$
- the logical combination of the three predicates Cond(rt−i) (i = 1, 2, 3).

To encode $\vec{a}$, we first transform the range of the variable Z, [Zmin, Zmax], into a fixed interval, say [0, 31].

$$Z^* = \frac{Z - Z_{\min}}{Z_{\max} - Z_{\min}} \quad (433)$$

Then Z^* will be further transformed by Eq. (430).

$$Z^{**} = \begin{cases} n, & \text{if } n \le Z^* < n+1 \\ 31 & \text{if } Z^* = 32 \end{cases} \quad (434)$$

Since there are only 32 cutoff values, each ai can be encoded by a 5-bit string. Hence the vector $\vec{a}$ can be encoded by a 15-bit binary string. To encode $\vec{\oplus}$, notice that each $\oplus$ has only two possibilities: ≥ or <. Therefore, a $\vec{\oplus}$ can be encoded by a 3-bit binary string (Table 53). Finally, there are a total of totally 8 logical combinations for three predicates and they can be encoded by 3-bit strings (Table A.2). In sum, a CONDS can be encoded by a 21-bit string (3 for logical combinations, 3 for inequalities, and 15 for the three thresholds). Therefore, each trading strategy can be represented by a 21-bit string.

Table 53: Binary Codes for Inequality Relation

Code	⊕1	⊕2	⊕3
0(000)	≥	≥	≥
1(001)	<	≥	≥
2(010)	≥	<	≥
3(011)	≥	≥	<
4(100)	<	<	≥
5(101)	<	≥	<
6(110)	≥	<	<
7(111)	<	<	<

Table 54: Binary Codes for Logical Combinations.

Logic Code	Logical Combination of Predicates
0(000)	Cond 1 OR (Cond 2 AND Cond 3)
1(001)	Cond 1 AND (Cond 2 OR Cond 3)
2(010)	(Cond 1 OR Cond 2) AND Cond 3
3(011)	(Cond 1 AND Cond 2) OR Cond 3
4(100)	(Cond 1 OR Cond 3) AND Cond 2
5(101)	(Cond 1 AND Cond 3) OR Cond 2
6(110)	Cond 1 OR Cond 2 OR Cond 3
7(111)	Cond 1 AND Cond 2 AND Cond 3

17.14.2 Ordinary Genetic Algorithms

The GA described below is a very basic version of a GA, and is referred to as the ordinary genetic algorithm (OGA). More precisely, it is very similar to the GA employed in Bauer (1994).
The genetic algorithm maintains a population of individuals,

$$P_i = \left\{g_1^i, ..., g_n^i\right\} \text{(435)}$$

for iteration i, where n is population size. Usually, n is treated as fixed during the whole evolution. Clearly, $P_i \subset G$.

- Evaluation step: Each individual g_j^i represents a trading strategy at the ith iteration (population). It can be implemented with the historical data rt−1, rt−2, and rt−3 by means of Eq. (260). A specific example is given in Eq. (261). Each trading strategy g_j^i is evaluated by a fitness function, say Eq. (264).
- Selection step: Then, a new generation of population (iteration i + 1) is formed by randomly selecting individuals from Pi in accordance with a selection scheme, which, in this PhD thesis, is the roulette-wheel selection scheme.

$$M_i = P_s\left(P_i\right) = \left(s_1\left(P_i\right), s_2\left(P_i\right), ..., s_n\left(P_i\right)\right) \text{(436)}$$

where

$$s_k : \left\{\binom{G}{n}\right\} \to G \text{ (437)}$$

$k = 1,2,....,n$ and $\left\{\binom{\mathrm{G}}{\mathrm{n}}\right\}$ is the set of all populations whose population size is n. The set Mi is also called the mating pool.

- Alteration step: Some members of the new population undergo transformations by means of genetic operators to form new solutions.

- Crossover: We use two-point crossover ck, which create new individuals by combining parts from two individuals.

$$O_i = P_c(M_i) = (c_1(M_i), c_2(M_i), ..., C_{n/2}(M_i)) \text{ (438)}$$

where

$$c_k : \left\{ \binom{G}{n} \right\} \to G \times G \text{ (439)}$$

k = 1, 2,. . ., n/2. Oi is known as the set of offspring in the GA.

- Mutation: We use bit-by-bit mutation mk, which creates new individuals by flipping, with a small probability, each bit of each individual of Oi.

$$P_{i+1} = P_m(O_i) = (m_1(M_i), m_2(O_i), ..., m_n(O_i)) \text{ (440)}$$

where

$$m_k : \left\{ \binom{G}{n} \right\} \to G \text{ (441)}$$

k = 1, 2,. . ., n.

- After the evaluation, selection and alteration steps, the new population Pi+1 is generated. Then we proceed with the three steps with Pi+1, and the loop goes over and over again until a termination criterion is met. The control parameters employed to run the OGA are given in Table 55.

Table 55: Control Parameters of OGA.

Number of generations	100
Population size (n)	100
Selection scheme	Roulette-wheel
Fitness function	Accumulated returns
Elitist strategy	Yes
Rank min	0.75
Crossover style	Two-point
Crossover rate	0.6
Mutation rate	0.001

17.15. Notes on Advanced Maths for Quantitative Investment

17.15.1 Stochastic processes

Markov processes

The evolution of a system is determined with the initial condition. Consider a process X_t for which the values $x_1, x_2, ...$ are measured at times $t_1, t_2, ...$. Here one dimensional variable x is

used for notational simplicity, through extension to multidimensional systems is trivial. It is assumed that the joint probability density $f(x_1,t_1;x_2t_2,...)$ exists and defines the system completely.

The conditional density function is defined. $t_1 > t_2 > t_3.....t_k$ in this text.

Using the conditional probability density, one can introduce the general equation

$$f(x_1,t_1) = \int f(x_1,t_1;x_2,t_2 \ \left|x_3t_3\right.)dx_2 = \int f(x_1,t_1;x_2,t_2 \ \left|x_3t_3\right.)f(x_2,t_2 \ \left|x_3t_3\right.)dx_2 \text{ (442)}$$

for the markov process

$$f(x_1,t_1\left|x_2,t_2;x_3t_3\right.) = f(x_1,t_1\left|x_2,t_2\right.) \text{ (443)}$$

Then the substitution of the last equation (443) to the equation (442) leads to Chapman-Kolmogorov equation

$$f(x_1,t_1\left|x_3t_3\right.) = \int \ f(x_1,t_1\left|x_2,t_2\right.)(x_2,t_2\left|x_3,t_3\right.)dx_2 \text{ (444)}$$

This equation can be used as the starting point for deriving the Fokker-Planck equation. First, equation (444) is transformed into the differential equation.

$$\frac{\partial}{\partial t}f(x,t\left|x_0,t_0\right.) = -\frac{\partial}{\partial x}\left[A(x,t)(f(x,t\left|x_0,t_0\right.\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[D(x,t)f(x,t\left|x_0,t_0\right.)\right] + \int\left[R(x\left|z,t\right.)f(z,t\left|x_0,t_0\right.)\right] - R(z\left|x,t\right.)f(x,t\left|x_0,t_0\right.) \ \left.\right]dz$$

(445)

In (445) , the drift coefficient and the diffusion coefficient are equal

$$A(x,t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int (z-x)f(z,t+\Delta t\left|x,t\right.)d_z \text{ (446)}$$

$$D(x,t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int (z-x)^2 f(z,t+\Delta t\left|x,t\right.)d_z \text{ (6)(447)}$$

The integral in the right hand side of the Chapman-Kolmogorov equation (445) si determined with the function

$$R(x,\left|z,t\right.) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int f(x,t+\Delta t\left|z,t\right.) \text{ (448)}$$

It describes the possible discontinuous jumps of the random variable. Neglecting this term in equation (445) yields the Fokker-Planck equation

$$\frac{\partial}{\partial t}f(x,t\left|x_0,t_0\right.) = -\frac{\partial}{\partial x}\left[A(x,t)(f(x,t\left|x_0,t_0\right.\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[D(x,t)f(x,t\left|x_0,t_0\right.)\right] \text{(449)}$$

This equation with $A(x,t) = 0$ and $D(x,t) = const$ is reduced to the diffusion equation that describes the Brownian motion

$$\frac{\partial}{\partial t} f(x,t|x_0,t_0) = \frac{D}{2}\frac{\partial^2}{\partial x^2}\left[f(x,t|x_0,t_0)\right] \text{(450)}$$

17.15.2 Stochastic differential equation

The Brownian motion can be represented in the differential form

$$dy(t) = \mu dt + \sigma dW(t) \text{ (451)}$$

The equation is named the stochastic differential equation. Note that the term $dW(t) = \left[W(t+dt) - W(t)\right]$ has the following properties

$$E\left[dW\right] = 0, E\left[dWdW\right] = dt, E\left[dWdt\right] = 0 \text{ (452)}$$

Let us calculate

$$(dy)^2 = \left[\mu dt + \sigma dW^2\right] = \mu^2 dt^2 + 2\mu dt\sigma dW + \sigma^2 dW^2 \approx \sigma^2 d \text{ (453)}$$

It follows from (453) that while dy is random variable $(dy)^2$ is a deterministic one. This allows us to derive the Ito's Lemma. Consider a function $F(y,t)$ that depends on both deterministic t and stochastic, $y(t)$, variables. Let us expand the differential for $F(y,t)$ into the Taylor series retaining linear terms and bearing in mind equation (453).

$$dF(y,t) = \frac{\partial F}{\partial y}dy + \frac{\partial F}{\partial t}dt + \frac{1}{2}\frac{\partial^2 F}{\partial y^2}(dy)^2 = \frac{\partial F}{\partial y}dy + \left[\ \frac{\partial F}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 F}{\partial y^2}\ \right]dt \text{ (454)}$$

The Ito's expression (454) has an additional term in comparison with the differential for a function with deterministic independent variables. Namely the term $\frac{\sigma^2}{2}\frac{\partial^2 F}{\partial y^2}$ has a stochastic nature. If $y(t)$ is the Brownian motion (10), then

$$dF(y,t) = \frac{\partial F}{\partial y}\left[\mu dt + \sigma dW(t)\ \right] + \left[\ \frac{\partial F}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 F}{\partial y^2}\ \right]dt =$$
$$\left[\mu\frac{\partial F}{\partial y} + \frac{\partial F}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 F}{\partial y^2}\ \right]dt + \sigma\frac{\partial F}{\partial y}dW(t) \text{ (455)}$$

Let us consider the function $F = W^2$ as a simple example for employing the ito's lemma. In this case $\mu = 0, \sigma = 1$, and equation (455) is reduced to

$$dF = dt + 2WdW \text{ (456)}$$

Finally, we specify Ito's expression for the geometric Brownian motion $F = \exp\left[y(t)\right]$. Since in this case $\frac{\partial F}{\partial y} = \frac{\partial^2 F}{\partial y^2} = F and \frac{\partial F}{\partial t} = 0$ then

$$dF = \left[\mu + \frac{\sigma^2}{2} \right] F dt + \sigma F dW(t) \text{ (457)}$$

Hence if F is the geometric Brownian motion is relative change $\frac{dF}{F}$ behaves as the arithmetic Brownian motion.

17.15.3 Stochastic integral

The stochastic integral is defined as

$$\int_0^T f(t) dW(t) = ms \lim_{n \to \infty} \sum_{i=1}^{n} f(t_{i-1}) \left[W(t_i) - W(t_{i-1}) \right] \text{ (458)}$$

The notation ms-lim means mean square limit. It means that the difference between the Ito integral in the left hand side of (458) and the sum in the right hand side of (458) has a variance that approaches zero as n increases to infinity. Thus (458) is equivalent to

$$\lim_{n \to \infty} E \left[\int_0^T f(t) dW(t) - \sum_{i=1}^{n} f(t_{i-1}) \left[W(t_i) - W(t_{i-1}) \right] \right]^2 = 0 \text{ (459)}$$

Let us consider the integral

$$I(t_2, t_1) = \int_{t1}^{t2} W(t) dW(t) \text{ (460)}$$

as an example of calculating the Ito's integral. If the function

17.15.4 Martingales

The martingale methodology plays an important role in the modern theory of finance. Martingale is a stochastic process X(t) that satisfies the following condition

$$E\left[X(t+1) \middle| X(t), X(t-1) .... \right] = X(t) \text{ (461)}$$

The equivalent definition is given by

$$E\left[X(t+1) - X(t) \middle| X(t), X(t-1) .... \right] = 0 \text{ (462)}$$

This equation means that the expectation to win at every round of the game being conditioned on the history of the game equals zero. In other words, martingale has no trend. A process with positive trend is named submartingale. A process with negative martingale is supermartingale. The prices with discounted risk premium are martingales.

The important property of Ito's integral is that it is a martingale. The integral

$$E\left[\int_t^{t+\Delta t} \sigma(z) dW(z) \right] = 0 \text{ (463)}$$

Therefore

$$E\left[\int_{t}^{t+\Delta t}\sigma(z)dW(z)\right]=\int_{0}^{t}\sigma(z)dW(z) \text{ (464)}$$

Note that the Brownian motion with drift is not a martingale

$$E\left[y(t+dt)\right]=E\left[\int_{t}^{t+\Delta t}dy+y(t)\right]=y(t+\mu(t) \text{ (465)}$$

The first integral satisfies the martingale condition. This result follows from the Doob Meyer decomposition theorem. Which states that a continuous submartingale X(t) at $0<t<\infty$ with finite expectation $E(X(t))<0$ can be decomposed into a continuous martingale and an increasing deterministic process.

17.15.5 Black-Scholes theory – Riskless portfolio

The basic assumptions of the classical option pricing theory are that the option price F(t) at time t is a continuous function of time and it's underlying asset's price follows the geometric Brownian motion.

$$F=F(S(t),t) \text{ (466)}$$

Let's derive the classical Black-Scholes equation. Since it is assumed that the option price F(t) is described with (466) and price of the underlying asset follows equation

$$dS=\mu Sdt+\sigma SdW \text{ (467)}$$

Ito's expression

$$dF(S,t)=\left[\mu S\frac{\partial F}{\partial S}+\frac{\partial F}{\partial t}+\frac{\sigma^2}{2}S^2\frac{\partial^2 F}{\partial S^2}\right]dt+\sigma S\frac{\partial F}{\partial S}dW(t) \text{ (468)}$$

Furthermore, we build a portfolio P with eliminated random contribution dW. namely we choose -1 (short) option and $\frac{\partial F}{\partial S}$ shares of the underlying asset

$$P=-F+\frac{\partial F}{\partial S}S \text{ (469)}$$

The change of the value of this portfolio within the time interval dt equals

$$dP=-dF+\frac{\partial F}{\partial S}dS \text{ (470)}$$

Since there are no arbitrage opportunities, this changes must equal to the interest earned by the portfolio value invested in the risk-free asset

$$dP=rPdt \text{ (471)}$$

The combination of equations yields the black-scholes equation

$$\frac{\partial F}{\partial t} + rS\frac{\partial F}{\partial S} + \frac{\sigma^2}{2}S^2\frac{\partial^2 F}{\partial S^2} - rF = 0 \quad (472)$$

Note that his equation does not depend on the stock price drift parameter μ, which is the manifestation of the risk-neutral valuation. In other words, investors do not expect a portfolio return exceeding the risk-free interest. The black scholes equation is the partial differential equation with the first order derivative.

17.15.6 Optimization in finance

Optimization is a branch of applied mathematics that derives its importance both from the variety of its applications and from the wide variety of its applications and from the availability of efficient algorithms. Mathematically, it refers to the minimization (or maximization) of a given objective function of several decision variables that satisfy functional constraints.

A typical optimization model addresses the allocation of scarce resources among possible alternative uses in order to maximize an objective function such as total profit.

Optimization

- Linear and nonlinear programming

The standard form of the LP is given below

$$\begin{aligned} &\min_x c^T x \\ &Ax = b \quad (473) \\ &x \geq 0, \end{aligned}$$

where $A \in IR^{mxn}, b \in IR^m, c \in IR^n$ are given and $x \in IR^n$ is the variable vector to be determined.

- Quadratic programming

A more general optimization problem is the quadratic optimization or the quadratic programming (QP) problem where the objective function is now a quadratic function of the variables. The standard form QP is defined as follows:

$$\begin{aligned} &\min_x \frac{1}{2}x^T Q x + c^T x \\ &Ax = b \quad (474) \\ &x \geq 0, \end{aligned}$$

where $A \in IR^{mxn}, b \in IR^m, Q \in IR^n$ are given and $x \in IR^n$ is the variable vector to be determined.

- Conic optimization

Another generalization of LP is obtained when the nonnegativity constraints $x \geq 0,$ are replaced by general conic inclusion constraints. For this purpose, we consider a closed convex cone C in a finite dimensional vector space X and the following conic optimization problem:

$$\begin{aligned} &\min_x c^T x \\ &Ax = b \quad (475) \\ &x \in C. \end{aligned}$$

- Integer programming

Integer programs are optimization problems that require some or all of the variables to take integer values. This restriction on the variables often makes the problems very hard to solve. Therefore we will focus on integer linear problems, which have a linear objective. A pure integer linear program is given by

$$\begin{aligned} &\min_x c^T x \\ &Ax = b \quad (476) \\ &x \geq 0, \text{integer} \end{aligned}$$

where $A \in IR^{mxn}, b \in IR^m, c \in IR^n$ are given and $x \in IN^n$ is the variable vector to be determined.

- Dynamic programming

Dynamic programming refers to a computational method involving recurrence relations. The idea is to divide the problem into stages in order to perform the optimization recursively. It is possible to incorporate stochastic elements into the recursion.

- Optimization with data uncertainty

In all the problem classes discussed so far (except dynamic programming) we made the implicit assumption that the data of the problem, namely the parameters such as Q, A, b, c in QP are known. This isn't normally the case.

Stochastic programming and robust optimization are the techniques used to solve this problem.

1. Stochastic programming

The terms stochastic programming refers to a optimization problem in which some of the problem data are random. The underlying problem might be a linear program, an integer program or a nonlinear program.

For example, a two stage stochastic program with recourse can be written:

$$\begin{aligned} &\max_x a^T x + E\left[\max_{y(w)} c(w)^T y(w)\right] \\ &Ax = b \quad (477) \\ &B(w)x + C(w)y(w) = d(w) \\ &x \geq 0, y(w) \geq 0 \end{aligned}$$

where the first-stage decisions are represented by vector x and the second-stage decisions by vector y(w) , which depend on the realization of random events w.

The objective function then contains a deterministic term and the expectation of the second-stage objective taken over all the realizations of the random event w.

2. Robust optimization

Robust optimization refers to the modelling of optimization problems with data uncertainty to obtain a solution that is guaranteed to be "good" for all possible realizations of the uncertain parameters.

A mathematical model for finding constraint-robust solutions will be described.

First consider an optimization problem of the form:

$$\begin{matrix} \min_x f(x) \\ G(x,p) \in K \end{matrix} \quad (478)$$

The constraint robust optimal solution can be found by solving the following problem

$$\begin{matrix} \min_x f(x) \\ G(x,p) \in K, \forall p \in \forall U \end{matrix} \quad (479)$$

A related concept is objective robustness, which occurs when when uncertain parameters appear in the objective function. This is often referred to as solution robustness in the literature. Such robust solutions must remain close to optimal for all possible realizations of the uncertain parameters. Next, consider an optimization problem of the form:

$$\begin{matrix} \min_x f(x,p) \\ x \in S \end{matrix} \quad (480)$$

Then, an objective-robust solution is obtained by solving:

$$\min_{x \in S} \max_{p \in U} f(x,p) \quad (481)$$

U is the uncertainty set of values of uncertain parameters p.

Mean-Variance optimization

The theory of optimal selection of portfolios was developed by Harry Markowitz in the 1950's.

$$\begin{aligned} E[x] &= x_1\mu_1 + ... + x_n\mu_n = \mu^T x, \\ Var[x] &= \sum_{i,j} \rho_{ij}\sigma_i\sigma_j x_i x_j = x^T Q x \end{aligned} \quad (482)$$

Markowitz's portfolio optimization problem, also called mean-variance optimization, can be formulated in three different but equivalent ways. One formulation is

$$\begin{matrix} \min_x x^T Q x \\ e^T x = 1 \\ \mu^T x \geq R \\ x \geq 0, \end{matrix} \quad (483)$$

Problem shown here is finding a minimum variance portfolio of the securities from 1 to n that yields at least a target value R of expected return.

17.15.7 Black-Litterman

Prior to advancing, it is important to introduce the Black-Litterman formula and provide a brief description of each of its elements. Throughout this article, K is used to represent the number of

views and N is used to express the number of assets in the formula. The formula for the new Combined Return Vector (E[R]) is :

$$E[R] = [(\tau \sum)^{-1} + P' \Omega^{-1} P]^{-1} [(\tau \sum)^{-1} \prod + P' \Omega^{-1} Q] \quad (484)$$

where

- $E[R]$ is the new (posterior) Combined Return Vector (N x 1 column vector);
- τ is a scalar;
- $\sum$ is the covariance matrix of excess returns (N x N matrix);
- P is a matrix that identifies the assets involved in the views (K x N matrix or 1 x N row vector in the special case of 1 view);
- Ω is a diagonal covariance matrix of error terms from the expressed views representing the uncertainty in each view (K x K matrix);
- $\prod$ is the Implied Equilibrium Return Vector (N x 1 column vector); and,
- Q is the View Vector (K x 1 column vector).

17.15.8 Statistical Physics and Finance

Markets as complex systems

Complex systems are composed of many particles, or objects, or elements that may be of the same or different kinds. The elements may interact in a more or less complicated fashion by more or less nonlinear couplings.

One example of standard complex system is Earth's climate, encompassing all components of the atmosphere, biosphere, cryosphere, and oceans and considering the effects of extraterrestrial processes such as solar radiation and tides.

The question of whether a system is complex or simple depends strongly on the level of scientific knowledge. We have currently no system to evaluate the complexity of a system only we can do in an intuitive way and let by intuitive knowledge.

1. Determinism versus chaos

We want to come now to the mathematical treatment of arbitrary complex system.

The deterministic principle has been shaken twice in modern physics.

First, in quantum mechanics tell us that we are not able to make accurate predictions of the trajectory of a particle. However we can argue that the deterministic character is still conserved as a property of the wave functions.

Second, the theory of deterministic chaos has shown that even in classical mechanics predictability cannot be guaranteed without absolutely precise knowledge of the microscopic configuration of the complete global system.

Chaos is not observed in linear systems. Mathematically, the signature of linearity is the superposition principle, which states that the sum of two solutions of the mechanical equations describing the system is again a solution. The theory of linear mechanical systems is fully

understood except for some technical problems. The breakdown of linearity, and therefore the breakdown of the superposition principle, is a necessary condition for the behavior of a nonlinear mechanical system to appear chaotic.

Trajectories of a system through 2N-dimensional phase space are unstable against small perturbations. The stability of an arbitrary trajectory to an infinitesimally small perturbation is studied by the Lyapunov exponents.

2. The probability distribution

Although most many-body systems exceed the technical possibilities of the mathematical calculus of mechanics, we are able to calculate the properties of large systems by applying of methods belonging to statistical physics.

In order to formulate the probability concept more precisely, we use the language of set theory. The elements

3. The Liouville equation

To address the task of whether we can determine the evolution of the probability distribution for a given microscopic system, to this aim we assume that an initial state Γ_0 was realized with a certain probability $\rho(\Gamma_0, t_0)d\Gamma$.

In the course of the microscopic motion, the initial state is shifted into another microscopic state along the trajectory $\Gamma(t) = \{q_i(t), p_i(t)\}$ with the boundary condition $\Gamma(0) = \Gamma_0$. In other words, the probability density ρ is conserved along the trajectory of the complex system.

This circumstance requires

$$\frac{d\rho}{dt} = \frac{\partial p}{\partial t} + \sum_{i=1}^{N}\left[\frac{\partial \rho}{\partial t}\dot{q}_i + \frac{\partial \rho}{\partial p}\dot{p}_i\right] = 0 \text{ (485)}$$

After replacing the velocities $\dot{q}_i$ and forces $\dot{p}_i$ by the equations we arrive at

$$\frac{d\rho}{dt} + \hat{L}\rho = 0 \text{ (486)}$$

where we have introduced the Liouvillian of the system.

$$\hat{L} = \sum_{i=1}^{N}\left[\frac{\partial H}{\partial p_i}\frac{\partial}{\partial q_i} + \frac{\partial H}{\partial q_i}\frac{\partial}{\partial p_i}\right] = 0 \text{ (487)}$$

The relation (486) is called the Liouville equation and it's the most important equation of statistical physics. The Liouvillian plays the same role that the Hamiltonian plays in Newtonian mechanics. The Liouville defines the equation of motion, which is represented by de distribution function ρ.

The meaning of the Liouville equation for the evolution of an economic system and also for many other complex systems lies in the combination of the probabilistic and deterministic features of the evolution process. The Liouville equation conserves our degree of belief.

17.15.9 Probability distributions

Basic definitions

Consider the random variable X. The probability density function P(x) defines the probability to find X between a and b

$$\Pr(a \le X \le b) = \int_a^b P(x)dx \quad (488)$$

The probability density must be nonnegative function and must satisfy the normalization condition

$$\sum_{X\min}^{X\max} P(x)dx = 1 \quad (489)$$

where the interval $[X\max, X\min]$ is the range of all possible values of X.

In fact, the infinite limits $[-\infty,+\infty]$ can always be used since P(x) may be set to zero outside the interval $[X\max, X\min]$. As a rule, the infinite integration limits are further omitted. Another way of describing random variable is to use the cumulative distribution function

$$\Pr(X \le b) = \int_{-\infty}^b P(x)dx \quad (490)$$

Obviously probability satisfies the condition

$$\Pr(X > b) = 1 - \Pr(X \ge b) \quad (491)$$

Two characteristics are used to describe probable values of random variable X: mean (or expectation) and median.

Mean of X is the average of all possible values of X that are weighted with the probability density P(x)

$$m \equiv E(X) = \int xP(x)dx \quad (492)$$

Median of X is the value, M for which

$$\Pr(X > M) = \Pr(X < M) = 0.5 \quad (493)$$

Median is the preferable characteristic of the most probable value for strongly skewed data samples.

The expectation of a random variable calculated using some available information I_t (that may change with time t) is called conditional expectation.

The conditional probability density is denoted by $P(x|I_t)$.

Conditional expectation equals

$$E(X_t|I_t) = \int xP(x|I_t)dx \quad (494)$$

Variance, Var is the standard deviation, σ, are the conventional estimates of the deviations from the mean values of X

$$Var(X) \equiv \sigma^2 = \int (x-m)^2 P(x)dx \text{ (495)}$$

In financial literate, the standard deviation of price is used to characterize the price volatility.
The higher moments of the probability distributions are defined

$$m_n \equiv E(X^n) = \int x^n P(x)dx \text{ (496)}$$

According to this definition, mean is the first moment $(m \equiv m_n)$ and variance can be expressed via the first two moments, $\sigma^2 = m_2 - m^2$. Two other important parameters, skewness S and kurtosis K, are related to third and fourth moments, respectively,

$$\sigma^2 = m_2 - m^2 \text{ (497)}$$

Both parameters are dimensonless. Zero skewness implies that de distribution is symmetrical around its mean value. The positive and negative values of skewness indicate long positive tails and long negative tails, respectively.

$$S = E\left[(x-m)^3\right]/\sigma^3, K = E\left[(x-m)^4\right]/\sigma^4 \text{ (498)}$$

Kurtosis characterizes the distribution peakedness. Kurtosis of the normal distribution equals 3. The excess kurtosis, $K_e = K - 3$, is often used as a measure of deviation from the normal distribution.
In particular, positive excess kurtosis or leptokurtosis indicates more frequent medium and large deviations from the mean value than is typical for the normal distribution.
Leptokurtosis leads to a flatter central part and to so-called fat tails in the distribution. Negative excess kurtosis indicates frequent small deviations from the mean value. In this case the distribution sharpens around its mean value. While the distribution tails decay faster than the tails of the normal distribution.
The joint distribution of two random variables X and Y is the generalization of the cumulative distribution

$$\Pr(X \le b, Y \le c) = \int_{-\infty}^{b}\int_{-\infty}^{c} h(x,y)dxdy \text{ (499)}$$

In this equation, $h(x,y)$ is the joint density that satisfies the normalization condition:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)dxdy = 1 \text{ (500)}$$

The two random variables are independent if their joint density function is simply the product of the univariate density functions: $h(x,y) = f(x)g(y)$. Covariance between 2 variables provides a

measure of their simultaneous change. Consider two variables X and y , that have the means mx and my , respectively. Their covariance equals

$$Cov(x,y) \equiv \sigma_{xy} = E\left[(x-m_x)(y-m_y)\right] = E[xy] - m_x m_y \quad (501)$$

Obviously, covariance reduces to variance if X=Y.

Positive covariance between two variables implies that these variates tend to change simultaneously in the same direction rather than in opposite direction. Conversely is straightforward. Another popular measure of simultaneous change is correlation coefficient.

$$Corr(x,y) = Cov(xy)/(\sigma_x \sigma_y) \quad (502)$$

17.15.10 Copulas

Sklar's theorem

The theorem proposed by Sklar underlies most applications of the copula. Sklar's theorem states that given a joint distribution function H for p variables, and respective marginal distribution functions, there exists a copula C such that the copula binds the margins to give the joint distribution.

For the bivariate case, Sklar's theorem can be stated as follows. For any bivariate distribution function H(x, y), let F(x) = H(x, (−∞,∞)) and G(y) = H((−∞,∞), y) be the univariate marginal probability distribution functions. Then there exists a copula C such that:

$$H(x,y) = C(F(x), G(y)) \quad (503)$$

(where we have identified the distribution C with its cumulative distribution function). Moreover, if marginal distributions, say, F(x) and G(y), are continuous, the copula function C is unique. Otherwise, the copula C is unique on the range of values of the marginal distributions.

Gaussian copula

One example of a copula often used for modelling in finance is the Gaussian Copula, which is constructed from the bivariate normal distribution via Sklar's theorem. For X and Y distributed as standard bivariate normal with correlation ρ the Gaussian copula function is:

$$\phi_{X,Y,p}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[x^2 + y^2 - 2\rho xy \right] \right) \quad (504)$$

is the density function for the bivariate normal variate with Peasson's product moment correlation coefficient ρ , ϕ is the density of the N(0,1) distribution , the marginal density.

Archimedean copulas

One particularly simple form of a copula is

$$H(x,y) = ^{-1} \left(\psi(F(x) + \psi(G(y))\right) \quad (505)$$

where ψ is known as a generator function. Such copulas are known as Archimedean. Any generator function which satisfies the properties below is the basis for a valid copula.

Clayton copula

$$\Psi(x) = x^{\theta} - 1; H(x,y) = (F(x)^{\theta} + G(y)^{\theta} - 1)^{\frac{1}{\theta}} \quad (506)$$
$$\theta \leq 0$$

For θ = 0 in the Clayton copula, the random variables are statistically independent. The generator function approach can be extended to create multivariate copulas, by simply including more additive terms.

17.15.11 Important distributions

Normal distribution

$$\begin{aligned} &f(x)\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}, x \in R, \sigma > 0 \\ &Mean: E[X] = \mu \\ &Variance: \mathrm{var}[X] = \sigma^2 \\ &Moment, generating, function: e^{\mu t + \frac{1}{2}\sigma^2 t^2} \\ &Characteristic, function: e^{i\mu t + \frac{1}{2}\sigma^2 t^2} \end{aligned} \quad (507)$$

Poisson distribution

$$\begin{aligned} &f(x) = e^{\lambda}\frac{\lambda^x}{x!}, x = 0,1,2..., \lambda > 0 \\ &Mean: E[X] = \lambda \\ &Variance: \mathrm{var}[X] = \lambda \\ &Moment, generating, function: e^{\lambda(e^t - 1)} \\ &Characteristic, function: e^{\lambda(e^{it} - 1)} \end{aligned} \quad (508)$$

Lognormal distribution

$$\begin{aligned} &f(x) = \left\{ \begin{matrix} \frac{e^{-(\ln x - \mu)^2/2\sigma^2}}{\sigma x\sqrt{2\pi}}, x > 0 \\ 0, x \leq 0 \end{matrix} \right\}; \sigma > 0 \\ &Mean: E[X] = e^{\mu + \frac{1}{2}\sigma^2} \\ &Variance: \mathrm{var}[X] = e^{2\mu}(e^{2\sigma^2} - e^{\sigma^2}) \\ &Kth, moment: e^{k\mu + \frac{1}{2}\sigma^2 k^2} \end{aligned} \quad (509)$$

Statistical Physics and Finance

1. Introduction
2. Finance and financial mathematics

The mathematical of describing finance starts from well defined hypotheses that consider more or less idealized rules and specific initial boundary conditions. The financial market is a complex

system from a physical point of view. In such a system, the rates of stocks and other asset prices are characterized as relevant degrees of freedom. All other degrees of freedom are irrelevant quantities.

The intention to want to describe the evolution of the share prices conceals in itself behind this division, whereas for instance, the mental states of the traders are interpreted as more or less irrelevant information for financial transactions.

The number of quantitatively available financial quantities such as shares , trading volumes , and funds is approximately of order of magnitude of $10^8 - 10^{10}$. The number is extremely small compared to the size of the irrelevant set containing the degrees of freedom with direct or indirect contact with the underlying structures of the financial market.

The dynamics of the irrelevant microscopic degrees of freedom may be formally eliminated bay application of a suitable projection formalism. It was demonstrated that one obtains probabilistic equations describing such a system of the relevant macroscopic level.

At first, it seems reasonable to describe markets within the framework of a Markov approximation. But this is maybe just not the best approximation.

3. Scales in financial data

In the natural sciences, especially in physics, the problem of reference units is considered basic to all experimental and theoretical work. In finance we have scales used often given in units that are fluctuating in time.

The most important quantity is price, which is indicated in units of a home currency or a key currency. These values show substantial fluctuations.

The causes for these fluctuations are miscellaneous and cannot usually be separated in detail.

Another issue is the choice of the appropriate timescale to use for analyzing financial data. The difficulty consists in the fact that we do not know how we have to handle the discontinuances and possible arrival of information when markets are closed. Although these markets are active 24 hours per day, the social environment and several biological cycles push the market activity to a permanent change of intensity in each financial region of the world.

Another possible timescale describing financial time series is the number of transactions. This scale eliminates the effect of the randomly distributed time intervals elapsing between transactions. Another source of randomness is the volume. Trading time – Time that elapses during market hours is the most common choice.

It should be remarked that financial time series is discontinuous however we can formally extend this series to continuous functions. Behind this artificial choice is concealed the fact that the microscopic processes contributing to the price formation are continuous.

4. Measurements of Price fluctuations

Let us define $X_\alpha(t)$ as the price of a financial asset α at time t. Then, we may ask which is the appropriate variable describing the stochastic behavior of the price fluctuations.

$$R_\alpha(t,\delta t)=\frac{\delta X_\alpha(t)}{X_\alpha(t)}=\frac{\delta X_\alpha(t+\delta t)}{X_\alpha(t)}-1 \quad (510)$$

$$\varepsilon_\alpha(t,\delta t)=Ln\frac{X_\alpha(t+\delta t)}{X_\alpha(t)}=LnX_\alpha(t+\delta t)+LnX_\alpha(t) \quad (511)$$

5. Empirical analysis
6. Probability Distributions

From a physical point of view, a financial market is a complex system whose evolution is given by the A-dimensional vector $X(t)=\{X_1(t),...,X_A(t)\}$ of all simultaneous observes share quotations and other asset prices.

Alternatively we can use the vector of the logarithmic price differences $\varepsilon(t,\delta t)=\{\varepsilon_1(t,\delta t,.....,\varepsilon_A(t,\delta t)\}$ in the place of X(t).

Both vectors are representations of the set of relevant degrees of freedom on the same macroscopic level of the underlying financial system.

The vector function $\varepsilon(t,\delta t)$ is denoted as the trajectory of the financial market in the A-dimensional space of the asset prices. On the macroscopic level each complex system can be described by a probabilistic theory. In this sense the probability distribution function $p\delta t(\varepsilon,t)$ is the central quantity of a physical description of financial markets.

By definition $p\delta t(\varepsilon,t)$ is the probability density for a change of the logarithmic prices by the value ε from the beginning to the end of the time interval $[t,t+\delta t\ ]$.

The quantity $p\delta t(\varepsilon,t)$ is interpreted in the context of Bayesian statistics as the hypothetical probability that a change ε could have taken place at time t independently of the realized event.

Using conditional probabilities the joint probability density can be written as

$$\begin{aligned}p\delta t(\varepsilon^{(N)},t_N;....;\varepsilon^{(0)},t_0)&=p\delta t(\varepsilon^{(N)},t_N\left|\varepsilon^{(N-1)},t_{N-1};....;\varepsilon^{(0)},t_0\right.)\\&\times p\delta t(\varepsilon^{(N-1)},t_{N-1}\left|\varepsilon^{(N-2)},t_{N-2};....;\varepsilon^{(0)},t_0\right.)\\&\times p\delta t(\varepsilon^{(1)},t_1\left|\varepsilon^{(0)},t_0\right.)p\delta t(\varepsilon^{(0)},t_0)\end{aligned} \quad (512)$$

The main problem using this formula is that one must obtain all information from the only trajectory observed in reality. To this end we need some assumptions.

7. Ergodicity in financial data

The first important prerequisite is the assumption of the validity of the ergodic hypothesis, that requires that a system starting from an arbitrary initial state can always arrive at each final state after a sufficiently long time.

The most useful statement of that principle is that the better theory of two competing theories that make exactly the same prediction is the simpler one.

8. Stationarity of financial markets

As a second important prerequisite for the description of financial markets, one assumes stationarity. Stationarity means that all probability distribution functions and all correlation functions are invariant under an arbitrary shift in time.

9. Markov approximation

In principle, the joint probability equation defined above contains all necessary information for the description of a financial market on the macroscopic level using price change at relevant quantities, unfortunately this function is too complicated for practical application. But we may use the markov approximation.

$$p\delta t(\varepsilon^{(N)},t_N;...;\varepsilon^{(0)},t_0)=\prod_{J=0}^{N}p\delta t(\varepsilon^{(j)},t_j) \quad (513)$$

If the logarithmic differences ε_α of the asset prices correspond to a sufficiently long time horizon δt, then we can assume that the price fluctuations take place statistically independently. In the case we cannot use Markovian processes memory effects can play an important role that cannot be neglected.

10. Taxonomy of stocks

Correlation and anticorrelation

In financial markets, many stocks are traded simultaneously. A reasonable way to detect similarities in the time evolution starts from the correlation coefficients

$$\upsilon_{\alpha B}=\frac{\left\langle \varepsilon_\alpha(t,\delta t)\varepsilon_B(t,\delta t)\right\rangle_T-\left\langle \varepsilon_\alpha(t,\delta t)\ \right\rangle_T\left\langle\ \varepsilon_B(t,\delta t)\right\rangle_T}{\sqrt{\left\langle \varepsilon^2{}_\alpha(t,\delta t)\ \right\rangle-\left\langle\ \varepsilon_\alpha(t,\delta t)\right\rangle^2{}_T}\sqrt{\left\langle \varepsilon^2{}_B(t,\delta t)\ \right\rangle-\left\langle\ \varepsilon_B(t,\delta t)\right\rangle^2{}_T}} \quad (514)$$

These coefficients are simply the normalized components of the heuristically determined covariance matrix.

We may discuss the correlation coefficients in terms of a geometric representation. To this end, we introduce the normalized price fluctuations

$$\hat{\varepsilon}_\alpha(t,\delta t)=\frac{\varepsilon_\alpha(t,\delta t)-\left\langle\ \varepsilon_\alpha(t,\delta t)\ \right\rangle_T}{\sqrt{\left\langle\ \varepsilon^2{}_\alpha(t,\delta t)\ \right\rangle_T-\left\langle\ \varepsilon_\alpha(t,\delta t)\ \right\rangle^2{}_T}} \quad (515)$$

So that the correlation coefficient can be written as second moments of these reduced quantities

$$\upsilon_{\alpha B}=\left\langle\hat{\varepsilon}_\alpha(t,\delta t)\hat{\varepsilon}_B(t,\delta t)\right\rangle_T \quad (516)$$

Let us now combine the S records of the normalized price fluctuations $\hat{\varepsilon}_\alpha(t,\delta t)$ into the S-dimensional vector

18. Abbreviations and Acronyms

ADF augmented Dickey-Fuller
a.e. almost everywhere
AIC Akaike information criterion
ANN Artificial Neural Networks
APT asset pricing theory
AR auto regressive
ARCH autoregressive conditional heteroschedastic
ARDL auto regressive distributed lag
ARIMA auto regressive integrated moving average
ARMA auto regressive moving average
a.s. almost surely
ASE American Stock Exchange
BET bond equivalent yield
BGM Brace-Gatarek-Musiela model
BIC Bayesian information criterion
CAPM capital asset pricing model
C(CAPM) conditional capital asset pricing model
CLT central limit theorem
CML capital market line
CrVaR credit risk value-at-risk
CvaR conditional value-at-risk
DAX Geman stock index
d.f (cumulative) distribution functions
DF Dickey-Fuller
DGP data generation process
DJIA Dow Jones Industrial Average
ECM error correction model
ECN electronic communication network
EM expectation maximization
ES expected shortfall
ESR expected shortfall risk
EVT extreme value theory
EWMA exponentially weighted moving average
GARCH generalized autoregressive conditional heteroschedastic
GET general equilibrium theory
GEV generalized extreme value

GMM generalized method of moments
GDP gross domestic product
HFD high frequency data
HJM Heath, Jarrow, Morton model
IC information criteria
IGRACH integrated GARCH
IID independent and identically distributed
IIN independent identically normal
IN independent normal
IR information ratio
lag operator
LIBOR London Interbank Offered Rate
LLN law of large numbers
LP linear program, linear programming
MA moving average
MDA maximum domain of attraction
MBS mortgage-backed securities
MIP mixed integer programming
ML maximum likelihood
MLE maximum likelihood estimator
MPT modern portfolio theory
MSCI Morgan Stanley Composite Index
M-V analysis mean-variance analysis
NASDAQ National Association of Securities Dealers Automated Quotation System
NAV net asset value
NYSE New York Stock Exchange
ODE ordinary differential equation
OLS ordinary least squares
OTC over-the-counter
P/B price-to-book ratio
PCA principal component analysis
PDE partial differential equation
pdf probability density function
QP quadratic program, quadratic programming
RDF resource description framework
RMT random matrix theory
ROI return on investment
SDE stochastic differential equation
S&P 500 Standard & Poor’s 500 Index

SML security market line

ss self similar

sssi self similar with stationary increments

UL unexpected loss

VaR value-at-risk

VAR vector auto regressive

VC theory Vapnik-Chervonenkis theory

19. Programming code

The programming code (C++, Mathematica, Matlab) on most of the experiments in this PhD thesis is available under request . Send me an e-mail miquel.noguer@hotmail.com. If you use the code for research, make sure you cite the code, not just for acknowledgment but also for replication of results.

20. References

[1]Samuelson, Paul A. 1965. “Proof that Properly Anticipated Prices Fluctuate Randomly.” Industrial Management Review, vol. 6, no. 1 (Spring):41–50.

[2] Fama, Eugene F. 1965. “The Behavior of Stock Market Prices.” Journal of Business, vol. 38, no. 1 (January):34–105.

[3] Andrew W. Lo. 1994The Adaptive Markets Hypothesis: Market Effciency from an Evolutionary Perspective

[4] Sharpe, W. F. (1994). "The Sharpe Ratio". Journal of Portfolio Management 21 (1): 49–58.

[5] Khandani, Amir, and Lo, Andrew. 2007. “What Happened to the Quants in August 2007?” Preprint. Available at: web.mit.edu/alo/www/Papers/ august07.pdf.

[6] Schiller, Robert. 2007. “Historic Turning Points in Real Estate.” Cowles Foundation Discussion Paper No. 1610. Available at: cowles.econ.yale.edu.

[7] Edwards, W., (1968), “Conservatism in human information processing,” In: Kleinmutz, B. (Ed.), Formal Representation of Human Judgment. John Wiley and Sons, New York, pp. 17-52.

[8] Tversky, A. and D. Kahneman (1974), “Judgment under uncertainty: heuristics and biases,” Science 185, 1124-1131.

[9] Shefrin, H., and M. Statman (1985), “The disposition to sell winners too early and ride losers too long: Theory and evidence,” Journal of Finance 40, 777–791.

[10] Frazzini (2006), “The Disposition Effect and Underreaction to News.” Journal of Finance, 61.

[11] Silber, W.L. (1994), “Technical trading: when it works and when it doesn't,” Journal of Derivatives, vol 1, no. 3, 39-44.

[12] De Long, J.B., A. Shleifer, L. H. Summers, and Waldmann, R.J. (1990), “Positive feedback investment strategies and destabilizing rational speculation,” The Journal of Finance, 45, 2, 379-395.

[13] Bikhchandani, S., D. Hirshleifer, and I. Welch (1992), "A theory of fads, fashion, custom, and cultural change as informational cascades," Journal of Political Economy, 100, 5, 992-1026.

[14] Wason, P.C. (1960), "On the failure to eliminate hypotheses in a conceptual task," The Quarterly Journal of Experimental Psychology, 12, 129-140.

[15] Garleanu, N. LH Pedersen and (2007), "Liquidity and Risk Management," The American Economic Review, 97, 193-197.

[16] Klaassen, Franc. 2002. "Improving GARCH Volatility Forecasts with Regime-Switching GARCH." Empirical Economics 27(2): 363–394.

[17] Nielsen, Steen, and Jan Overgaard Olesen. 2000. "Regime-Switching Stock Returns and Mean Reversion." Working Paper 11–2000. Department of Economics, Copenhagen Business School.

[18] Van Norden, Simon, and Huntley Schaller. 1997. "Regime Switching in Stock Market Returns." Applied Financial Economics 7: 177–191.

[19] Kaufmann, Sylvia, and Martin Scheicher. 1996. "Markov-Regime Switching in Economic Variables: Part I. Modelling, Estimating and Testing. Part II. A Selective Survey." Institute for Advanced Studies Economics Series no. 38.

[20] Chai, Soo, and Joon Lim. 2007. "Economic Turning Point Forecasting Using Neural Network with Weighted Fuzzy Membership Functions." Lecture Notes in Computer Science, Springer.

[21] Schiller, Robert. 2007. "Historic Turning Points in Real Estate." Cowles Foundation Discussion Paper No. 1610. Available at: cowles.econ.yale.edu.

[22] Alexander, Carol. 2001. Market Models: A Guide to Financial Data Analysis. West Sussex: John Wiley & Sons Ltd.

[23] Fama, Eugene, and Kenneth French. 1992. "The Cross-Section of Expected Stock Returns." Journal of Finance XLVII(2): 427–465.

[24] Grinold, Richard, and Ronald Kahn. 1999. Active Portfolio Management. New York: McGraw-Hill.

[25] Sargent T. and Simms C. "Business cycle modelling without pretending to have much a priori economic theory" working paper 55, Federal reserve bank of Minneapolis 1977

[26] Geweke J. "The dynamic factor análisis of economic time series".in Latent variables in socio-economic models.North-Holland 1977.

[27] Uhlenbeck, George, and Leonard Ornstein. 1930. "On the Theory of Brownian Motion." Physical Review 36: 823–841.

[28] Aït-Sahalia Y.,Wang Y., Yared F. (2001) "Do Option Markets Correctly Price the Probabilities of Movement of the Underlying Asset?" Journal of Econometrics, 101.

[29] Carr P., Madan D. (1998) "Toward a Theory of Volatility Trading" Research Papers, Courant Institute of Mathematical Sciences, New York University.

[30] Almgren and Chriss, "Optimal Execution of Portfolio Transactions."

[31] Robert Almgren and Julian Lorenz, "Adaptive Arrival Price," in Brian R. Bruce (ed.), Algorithmic Trading III: Precision, Control, Execution (London: Euromoney Institutional Investor, 2007), pp. 59–66.

[32] J. P. Bouchaud and M. Potters, Theory of Financial Risks: From Statistical Physics to Risk Management, Cambridge University Press, 2003.

[33] D. Challet, A. Chessa, A. Marsili, and Y. C. Chang, ‘‘From Minority Games to Real Markets,’’ Quantitative Finance 1, 168–176 (2001).

[34] W. B. Arthur, ‘‘Inductive Reasoning and Bounded Rationality,’’ American Economic Review 84, 406–411 (1994).

[35] A. Beja and M. B. Goldman, ‘‘On the Dynamic Behavior of Prices in Disequilibrium,’’ Journal of Finance 35, 235–248 (1980).

[36] M Noguer Alonso(2008) Agent-based modelling of financial markets:The new kind of science market model. Master Thesis. UNED.

[37] Aumann, Robert J. (1987), "game theory,", The New Palgrave: A Dictionary of Economics, 2, pp. 460–82 .

[38] von Neumann, John; Morgenstern, Oskar (1944), Theory of games and economic behavior, Princeton University Press

[39] P. Bak, How Nature Works: The Science of Self-Organized Criticality. Copernicus, Springer, New York, 1996.

[40] for a review, see: D. Challet, M. Marsili, Y.C. Zhang, Minority Games, Oxford University Press (2005)

[41] S. Mike and J. Farmer. An empirical behavioral model of liquidity and volatility, Journal of Economic Dynamics andControl, 32:200 (2008).

[42] W. De Bondt, R. Thaler, Does the market overreact ? Journal of Finance 40, 793-805 (1985).

[43] M. Wyart, J.-P. Bouchaud, Self-referential behaviour, overreaction and conventions in financial markets, JEBO 63, 1(2007).

[44] R. Lyons, The microstructure approach to Foreign Exchange rates, MIT Press, Cambridge MA 2001

[45] J. Hasbrouck, Empirical Market Microstructure, Oxford University Press 2007

[46] Bouchaud. Jean-Philippe.The endogenous dynamics of markets: price impact and feedback loops. arXiv:1009.2928v1 [q-fin.ST] 15 Sep 2010

[47] Meucci, A., 2005, Risk and Asset Allocation (Springer).

[48] Rachev, S. T., 2003, Handbook of Heavy Tailed Distributions in Finance (Elsevier/North-Holland).

[49] Hamilton, J. D., 1994, Time Series Analysis (Princeton University Press).

[50] Beran, J., 1994, Statistics for Long-Memory Processes (Chapman and Hall).

[51] Baillie, R.T., 1996, Long memory processes and fractional integration in econometrics, Journal of Econometrics 73, 5—59.

[52] Cont, R., 2005, Volatility clustering in financial markets: Empirical facts and agent-based models, in A. Kirman, and G. Teyssiere, ed.: Long Memory in Economics. Springer.

[53] Schoutens, W., 2003, Levy Processes in Finance (Wiley.).

[54] , Rama Cont and P. Tankov, 2008, Financial Modelling with Jump Processes (Chapman

and Hall/CRC) 2nd edn.

[55] Merton, R. C., 1976, Option pricing when underlying stocks are discontinuous, Journal of Financial Economics 3, 125—144.

[56] Monroe, I., 1978, Processes that can be imbedded in Brownian motion, The Annals of Probability 6, 42—56.

[57] Mandelbrot, B., 1963, The variation of certain speculative prices, Journal of Business 36, 394—419.

[58] Samorodnitsky, G., and M. S. Taqqu, 1994, Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance (Chapman and Hall).

[59] Kou, S. G., 2002, A jump-diffusion model for option pricing, Management Science 48, 1086—1101.

[60] Barndorff-Nielsen, O.E., 1998, Processes of normal inverse Gaussian type, Finance and Stochastics 2, 41—68.

[61] Madan, D., and F. Milne, 1991, Option pricing with VG martingale components, Mathematical Finance 1, 39—56.

[62] Carr, P., H. Geman, D. H. Madan, and M. Yor, 2003, Stochastic volatility for Levy processes, Mathematical Finance 13, 345—382.

[63] Vasicek, O., 1977, An equilibrium characterisation of the term structure, Journal of Financial Economics 5, 177—188.

[64] Cox, J. C., J. E. Ingersoll, and S. A. Ross, 1985, A theory of the term structure of interest rates, Econometrica 53, 385—407.

[65] Comte, F., and E. Renault, 1996, Long memory continuous time models, Journal of Econometrics 73, 101—149.

[66] Heston, S. L., 1993, Closed-form solution for options with stochastic volatility with applications to bond and currency options, The Review of Financial Studies 6, 327—343.

[67] Hamilton, J. D., 1994, Time Series Analysis (Princeton University Press).

[68] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society. Series B (Methodological) 39 (1): 1–38. JSTOR 2984875. MR0501537.

[69] B. B. Mandelbrot, Fractals and Scaling in Finance, Springer-Verlag, 1997.

[70] H. O. Peitgen, H. Jurgens, and D. Saupe, Chaos and Fractals: New Frontiers in Science, Springer-Verlag, 1992.

[71] J. Y. Campbell, A. W. Lo, and A. C. MacKinlay, The Econometrics of Financial Markets, Princeton University Press, 1997.

[72] C. J. G. Evertsz and B. B. Mandelbrot, Multifractal Measures,

[73] B. B. Mandelbrot: ‘‘Limit Lognormal Multifractal Measures,’’ Physica A163, 306–315 (1990).

[74] B.B. Mandelbrot, L. Calvet, & A. Fisher, Cowles Foundation Discussion Papers: 1164 (1997).

[75] T. Lux, The multi-fractal model of Asset Returns: its estimation via GMM and its use for volatility forecasting, University of Kiel,Working Paper, (2003).

[76] Calvet, Laurent; Adlai Fisher (2001). "Forecasting multifractal volatility". Journal of Econometrics 105: 27–58. doi:10.1016/S0304-4076(01)00069-0.

[77] Roweis, S.; Ghahramani, Z. (1999). "A Unifying Review of Linear Gaussian Models". Neural Computation 11 (2): 305–345. doi:10.1162/089976699300016674. PMID 9950734.

[78] Bernardo J. and A.F.M. Simth (1994). Bayesian theory. Wiley.

[79] Barndorff-Nielsen, O.E., and N. Shephard, 2001, Non-Gaussian Ornstein- Uhlenbeck-based models and some of their uses in financial economics (with discussion), J. Roy. Statist. Soc. Ser. B 63, 167—241.

[80] Van der Werf, K. W., 2007, Covariance of the Ornstein-Uhlenbeck process, Personal Communication.

[81] Sorensen, Eric H., Ronald Hua, and Edward Qian. 2005. “Contextual Fundamentals, Models, and Active Management.” Journal of Portfolio Management, vol. 31, no. 1 (Fall):23–26.

[82] Johansen, S. 1991. “Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models.” Econometrica, vol. 59, no. 6 (November): 1551–81.

[83] Litterman, Robert B. 1986. “Forecasting With Bayesian Vector Autoregressions— Five Years of Experience.” Journal of Business & Economic Statistics, vol. 4:25–38.

[84] B. LeBaron, ‘‘Chaos and Nonlinear Forecastability in Economics and Finance,’’ Philosophical Transactions of the Royal Society of London 348A, 397–404 (1994).

[85] W. A. Brock, D. Hsieh, and B. LeBaron, Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence, MIT Press, 1991.

[86] T. Lux, ‘‘The Socio-economic Dynamics of Speculative Markets: Interacting Agents, Chaos, and the Fat Tails of Return Distributions,’’ Journal of Economic Behavior and Organization 33,143–165 (1998).

[87] R. C. Hilborn, Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers, Oxford University Press, 2000.

[88] P. Berge, Y. Pomenau, and C. Vidal, Order Within Chaos: Towards a Deterministic Approach to Turbulence, Wiley, 1986.

[89] R. C. Hilborn, Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers, Oxford University Press, 2000.

[90] R. C. Hilborn, Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers, Oxford University Press, 2000.

[91] David Ruelle and Floris Takens (1971). "On the nature of turbulence". Communications of Mathematical Physics 20: 167–192. doi:10.1007/BF01646553.

[92] Elliott, Ralph Nelson (1994). Prechter, Robert R., Jr.. ed. R.N. Elliott's Masterworks. Gainesville, GA: New Classics Library. pp. 70, 217, 194, 196. ISBN 978-0932750761.

[93] John Murphy: Technical Analysis of Futures Markets. New York, N.Y.: New York Inst. of Finance, 1986. [ISBN 0-13-898008-X]

[94] Asness, C. S. (2008): The past and future of quantitative asset management, CFA Institute Conference Proceedings Quarterly, December 2008.

[95] Sefton, J. A., and A. Scowcroft (2005): Understanding momentum, Financial Analysts Journal 61 (2), 64—82.

[96] Fama, E.F., and K. R. French (1995): Size and book-to-market factors in earnings and returns, Journal of Finance 50, 131—155.

[97] Petkova, R. (2006): Do the Fama-French factors proxy for innovations in predictive variables?, Journal of Finance 51, 581—612.

[98] Sorensen, E. H. (2009): Active equity management for the future, Journal of Portfolio Management 36 (1), 60—68.

[99] Sefton, J. A., and A. Scowcroft (2005): Understanding momentum, Financial Analysts Journal 61 (2), 64—82.

[100] Jegadeesh, N., and S. Titman (1993): Returns to buying winners and selling losers: Implications for stock market efficiency, Journal of Finance 48(1), 65— 91.

[101] Fama, E.F., and K. R. French (1992): The cross-section of expected stock returns, Journal of Finance 47, 427—465.

[102] Andrews, D. W. K. (1993): Tests for parameter instability and structural change with unknown change point, Econometrica 61, 821—856.

Andrews, D. W. K. (2003): Tests for parameter instability and structural change with unknown change point: A corrigendum, Econometrica 71, 395—397.

[103] Viceira, L. M. (1997): Testing for structural change in the predictability of asset returns, Working paper, Harvard Business School.

[104] Campbell, J. Y., A. W. Lo and A. C. MacKinlay (1997): The econometrics of financial markets. Princeton: Princeton University press.

[105] Patton, A. J. (2009): Are “market neutral” hedge funds really market neutral?, Review of Financial Studies 22(7), 2495—2530.

[106] Jagannathan, R., and Z. Wang (1996): The conditional CAPM and the crosssection of expected returns, Journal of Finance 51(1), 3—53.

[107] Ang, A., and J. Chen (2007): CAPM over the long run: 1926-2001, Journal of Empirical Finance 14(1), 1—40

[108] Lewellen, J., and S. Nagel (2007): The conditional CAPM does not explain asset-pricing anomalies, Journal of Financial Economics 82(2), 289—314.

[109] Petkova and Zhang (2005): Is value riskier than growth?, Journal of Financial Economics 78, 187—202.

[110] Chua, D. B., M. Kritzman and S. Page (2009): The myth of diversification, Journal of Portfolio Management 36 (1), 26—35.

[111] Koenker, R. (2005): Quantile regression. Cambridge: Cambridge University Press.

[112] Leinweber, David J (2009).Nerds on wall street. Wiley

[113] Vapnik, Vladimir N. 1995. The Nature of Statistical Learning Theory. New York: Springer.

[114] Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, Prentice Hall, ISBN 0-13-273350-1

[115] Bishop, C.M. (1995) Neural Networks for Pattern Recognition, Oxford: Oxford University Press. ISBN 0-19-853849-9 (hardback) or ISBN 0-19-853864-2 (paperback)

[116] Kohonen, T. and Honkela, T. (2007). "Kohonen network". Scholarpedia. http://www.scholarpedia.org/article/Kohonen_network.

[117] ^ Elman et al., Jeffrey (1996). Rethinking Innateness: A Connectionist Perspective on Development. Cambridge, MA: MIT Press. ISBN 026255030X.

[118] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proceedings of the National Academy of Sciences of the USA, vol. 79 no. 8 pp.

[119] Ackley, D. H.; Hinton, G. E.; Sejnowski, T. J. (1985). "A Learning Algorithm for Boltzmann Machines". Cognitive Science 9: 147–169. http://learning.cs.toronto.edu/~hinton/absps/cogscibm.pdf.

[120] Wulfram Gerstner (2001). "Spiking Neurons". in Wolfgang Maass and Christopher M. Bishop. Pulsed Neural Networks. MIT Press. ISBN 0262632217. http://books.google.com/books?id=jEug7sJXP2MC&pg=PA3&dq=%22Pulsed+Neural+Networks%22+rate-code+neuroscience&ei=FEo0ScetL4zukgSyldy8Ag.

[121] Fahlman, S, Lebiere, C (1991). The Cascade-Correlation Learning Architecture, created for National Science Foundation, Contract Number EET-8716324, and Defense Advanced Research Projects Agency (DOD), ARPA Order No. 4976 under Contract F33615-87-C-1499. electronic version

[122] Corinna Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, 20, 1995. http://www.springerlink.com/content/k238jx04hm87j80g/

[123] David Meyer, Friedrich Leisch, and Kurt Hornik. The support vector machine under test. Neurocomputing 55(1-2): 169-186, 2003 http:/ / dx. doi. org/ 10. 1016/ S0925-2312(03)00431-4

[124] Corinna Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, 20, 1995. http:/ / www. springerlink. com/ content/ k238jx04hm87j80g/

[125] ACM Website, Press release of March 17th 2009. http:/ / www. acm. org/ press-room/ news-releases/ awards-08-groupa

[126] M. Aizerman, E. Braverman, and L. Rozonoer (1964). "Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control 25: 821–837.

[127] Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola and Vladimir Vapnik (1997). "Support Vector Regression Machines". Advances in Neural Information Processing Systems 9, NIPS 1996, 155-161, MIT Press.

[128] Suykens J.A.K., Vandewalle J., Least squares support vector machine classifiers, Neural Processing Letters, vol. 9, no. 3, Jun. 1999, pp. 293-300.

[129] Quinlan, J. Ross. 1979. “Discovering Rules by Induction from Large Collections of Examples.” In Expert Systems in the Micro-Electronic Age. Edited by Donald Michie. Edinburgh: Edinburgh University Press.

[130] Holland J.(1975), Adaptation in Natural and Artificial Systems

[131] Kjellström, G. (December 1991). "On the Efficiency of Gaussian Adaptation". Journal of Optimization Theory and Applications 71 (3): 589–597. doi:10.1007/BF00941405.

[132] Holland, J. H. (1975). Adaptation in natural and artificial systems. Ann Arbor, MI: University of Michigan Press.

[133] Bauer, R. J. (1994). Genetic algorithms and investment strategies. Wiley.

[134] Palmer, R. G., Arthur, W. B., Holland, J. H., LeBaron, B., & Taylor, P. (1994). Artificial economic life: A simple model of a stockmarket. Physica D, 75, 264–274.

[135] Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). The econometrics of financial markets. Princeton University Press.

[136] Drunat, J., Dufr´enot, G., & Mathieu, L. (1998). Modelling burst phenomena: Bilinear and autogressiveexponential models. In: C. Dunis & B. Zhou (Eds), Nonlinear Modelling of High Frequency Financial Time Series (pp. 201–221). Wiley.

[137] Granger, D.W. J., & Anderson, A. P. (1978). An introduction to bilinear time series models. Gottingen and Zurich: Vandenhoech & Ruprecht.

[138] Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH modelling on finance: A review of the theory and empirical evidence. Journal of Econometrics, 52, 5–59.

[139] Tong, H. (1990). Non-linear time series: A dynamical system approach. New York: Oxford University Press.

[140] Brock,W. A., Hsieh, D.,&LeBaron, B. (1991). Nonlinear dynamics, chaos and instability. Cambridge, MA: MIT Press.

[141] Box, G. E. P., & Jenkings, G. M. (1976). Time series analysis: Forecasting and control. San Fransisco: Holden-Day.

[142] Granger, D.W. J., & Anderson, A. P. (1978). An introduction to bilinear time series models. Gottingen and Zurich: Vandenhoech & Ruprecht.

[143] Subba-Rao, T. (1981). On the theory of bilinear time series models. Journal of the Royal StatisticalSociety, Series B, 43, 244–255.

[144] Subba-Rao, T., & Gabr, M. M. (1980). An introduction to bispectral analysis and bilinear time series models. In: Lecture Notes in Statistics (Vol. 24). New York: Springer-Verlag.

[145] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica, 50, 987–1008.

[146] Bollerslev, T. P. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31, 307–327.

[147] Tong, H. (1983). Threshold models in nonlinear time series analysis. In: Lecture Notes in Statistics(Vol. 21). Heidelberg: Springer-Verlag.

[148] Refenes, A.-P. (1995). Testing strategies and metrics. In: A.-P. Refenes (Ed.), Neural Networks in theCapital Markets (pp. 67–76). New York: Wiley.

[149] Jobson, J. D., & Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. Journal of Finance, 36(4), 889–908.

[150] Arnold, S. F. (1990). Mathematical statistics. New Jersey: Prentice-Hall.

[151] Sharpe, W. F. (1966). Mutual fund performance. Journal of Business, 39(1), 119–138.

[152] Chen, S.-H., & Tan, C.-W. (1999). Estimating the complexity function of financial time series: An estimation based on predictive stochastic complexity. Journal of Management and Economics, 3.

[153] Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economic Letters, 6, 255–259.

[154] Chen, S.-H., & Lu, C.-F. (1999). Would evolutionary computation help for designs of artificial neural nets in financial applications? In: Proceedings of 1999 Congress on Evolutionary Computation.IEEE Press.

[155] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74, 427–431.

[156] Moody, J., & Wu, L. (1997). What is the “true price”? – state space models for high frequency FX data. Proceedings of the Conference on Computational Intelligence for Financial Engineering. IEEE Press.

[157] Barnett, W. A., Gallant, A. R., Hinich, M. J., Jungeilges, J. A., Kaplan, D. T., & Jensen, M. J. (1997). A single-blind controlled competition among tests for nonlinearity and chaos. Paper presented at the 1997 Far Eastern Meeting of the Econometric Society (FEMES’97), Hong Kong, July 24–26, 1997 (Session 4A).

[158] Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH modelling on finance: A review of the theory and empirical evidence. Journal of Econometrics, 52, 5–59.

[159] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Publishing Company, 1989.

[160] T. Hellstr€om, K. Holmstr€om, Predicting the stock market, Technical Report IMa-TOM-1997-07, Department of Mathematics and Physics, Malardalen University, Sweden, 1997.

[161] D. Noever, S. Baskaran, Genetic algorithms trading on the S&P500, The Magazine of Artificial Intelligence in Finance 1 (3) (1994) 41–50.

[162] S. Mahfoud, G. Mani, Financial forecasting using genetic algorithms, Applied Artificial Intelligence 10 (6) (1996) 543–565.

[163] A. Muhammad, G.A. King, Foreign exchange market forecasting using evolutionary fuzzy networks, in: Proceedings of the IEEE /IAFE 1997 Computational Intelligence for Financial Engineering, 1997, pp. 213–219.

[164] F. Kai, X, Wenhua, Training neural network with genetic algorithms for forecasting the stock price index, in: Proceedings of the 1997 IEEE International Conference on Intelligent Processing Systems, Beijing, China, October 1997, vol. 1, pp. 401–403.

[165] A.P.N. Refenes, A.N. Burgess, Y. Bentz, Neural networks in financial engineering: a study in methodology, IEEE Transactions on Neural Networks 8 (6) (1997) 1222–1267.

[166] T. Bollerslev, Generalized autoregressive conditional heteroskedaticicty, Journal of Econometrics 52(1986) 307–327.

[167] J.Y. Campbell, A.W. Lo, A.C. MacKinlay, The Econometrics of Financial Markets, Princeton University Press, Princeton, New Jersey, 1997.

[168] F. Castiglione, Forecasting price increments using an artificial neural network, Advances in Complex Systems 4 (1) (2001) 45–56.

[169] P. Campolucci, A. Uncini, F. Piazza, B.D. Rao, On-line learning algorithms for locally recurrent neural networks, IEEE Transactions on Neural Networks 10 (2) (1999) 253–271.

[170] C.L. Giles, S. Lawrence, A. Chung Tsoi, Rule inference for financial prediction using recurrent neural networks, in: Proceedings of IEEE/IAFE Conference on Computational Intelligence for Financial Engineering (CIFE), 1997, pp. 253–259.

[171] B. LeBaron, A.S. Weigend, A boostrap evaluation of the effect of data splitting on financial time series, IEEE Transactions on Neural Networks 9 (1997) 213–220.

[172] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991) 79–87.

[173] A.S. Weigend, M. Mangeas, A.N. Srivastava, Nonlinear gated experts for time series: discovering regimes and avoiding overfitting, International Journal of Neural Systems 6 (1995) 373–399.

[174] A.S. Weigend, Time series analysis and prediction using gated experts with application to energy demand forecast, Applied Artificial Intelligence 10 (1996) 583–624.

[175] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman and Hall, London, 1989.

[176] S. Shi, A.S. Weigend, Taking time seriously: Hidden Markov experts applied to financial engineering, in: Proceedings of the 1997 IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, 1997, pp. 244–252.

[177] L.R. Rabiner, B.H. Juang, An introduction to Hidden Markov models, IEEE ASSP Magazine 3 (1) (1986) 4–16.

[178] J.H. Holland, Adaptation, in: R. Rosen, F.M. Snell (Eds.), Progress in Theoretical Biology, vol. 4, Academic Press, New York, 1976, pp. 263–293.

[179] T. Kovacs, Evolving optimal populations with XCS classifier systems, MSc. Dissertation, University of Birmingham, UK, 1996.

[180] P.L. Lanzi, A study of the generalization capabilities of XCS, in: Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97), Morgan Kaufmann, San Francisco, CA, 1997, pp. 418–425.

[181] S.W. Wilson, Generalization in the XCS classifier System, in: Proceedings of the Third annual Genetic Programming Conference, Morgan Kaufmann, San Francisco, CA, 1998, pp. 665–674.

[182] S.W. Wilson, State of XCS classifier system research, Second International Workshop on Learning Classifier Systems during GECCO99, 1999.

[183] S.W. Wilson, Get Real! XCS with Continuous-Valued Inputs, in: Festschrift in Honor of John H. Holland, L.Booker, S. Forrest, M. Mitchell, and R. Riolo (Eds.), Center of Study of Complex Systems, May 15–18, The University of Michigan, ANN Arbor, MI, 1999, pp. 111– 121.

[184] P.L. Lanzi, Adding memory to XCS, in: Proceedings of the IEEE Conference on Evolutionary Computation (ICEC98), 1998, pp. 609–614.

[185] P.L. Lanzi, A. Perrucci, Extending the representation of classifier conditions part II: from messy coding to S-expressions, in: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '99), vol. 1, 1999, pp. 345–353.

[186] R.E. Dorsey, R.S. Sexton, The use of parsimonious neural networks for forecasting financial time series, Journal of Computational Intelligence in Finance 6 (1) (1998) 24–31.

[187] P.J.B. Hancock, Pruning neural nets by genetic algorithm, in: International Conference on Artificial Neural Networks, Elsevier, 1992, pp. 991–994.

[188] C.L. Giles, S. Lawrence, A. Chung Tsoi, Rule inference for financial prediction using recurrent neural networks, in: Proceedings of IEEE/IAFE Conference on Computational Intelligence for Financial Engineering (CIFE), 1997, pp. 253–259.

[189] A.S. Weigend, H.G. Zimmermann, Exploiting local relations as soft constraints to improve forecasting, Journal of Computational Intelligence in Finance 6 (1998) 14–23.

[190] S.E. Fahlmann, C. Lebiere, The cascade-correlation learning architecture, Technical Report CMU-CS-90-100, Carnegie Mellon University, 1990.

[191] C.L. Giles, S. Lawrence, A. Chung Tsoi, Rule inference for financial prediction using recurrent neural networks, in: Proceedings of IEEE/IAFE Conference on Computational Intelligence for Financial Engineering (CIFE), 1997, pp. 253–259.

[192] A.S. Weigend, B.A. Huberman, D.E. Rumelhart, Predicting sunspots and exchange rates with connectionist networks, in: Proceedings of the 1990 NATO Workshop on Nonlinear Modelling and Forecasting, Addison Wesley, Santa Fe, NM, USA, 1991.

[193] Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate data analysis. New York: Prentice-Hall; 1995.

[194] Elman JL. Finding structure in time. Cognitive Science 1990;14:179–211.

[195] Kasabov N, Watts M. Spatial-temporal adaptation in evolving fuzzy neural networks for on-line adaptive phonemerecognition. Technical Report TR99/03, Department of Information Science, University of Otago, 1999.

[196] F. Takens (1981). "Detecting strange attractors in turbulence". in D. A. Rand and L.-S. Young. Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol. 898. Springer-Verlag. pp. 366–381.

[197] F. Smetsand R. Wouters(2004), Comparing Shocks and Frictions in US and Euro Area Business Cycles: A Bayesian DSGE Approach, European Central Bank, Working Paper Series, No. 391. Also in Journal of Applied Econometrics, Vol. 20, No. 1, 2005, pp. 161-183.

[198] F. Smetsand R. Wouters(2007), Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach, European Central Bank, Working Paper Series, No. 722. Also in American Economic Review, Vol. 97, No. 3, 2007, pp. 586-606.

[199] Erb,Claude, Harvey Campbell (2006) The tactical and strategic value of commodity futures. SSRN.

[200] Andrews, D. W. K. (1993): Tests for parameter instability and structural change with unknown change point, Econometrica 61, 821—856.

[201] Andrews, D. W. K. (2003): Tests for parameter instability and structural changewith unknown change point: A corrigendum, Econometrica 71, 395—397.