The wider picture concerned capitalism itself. Up until the crash it had been almost a kind of dogma, fostered during the Reagan and Thatcher era, that markets were efficient enough to take care of their own problems, and that as far as possible governments should not intervene in the market process. Thus laissez-faire capitalism was re-introduced to the world. After the fall of the Berlin Wall, this thinking was raised to a higher level and globalisation kicked off as a result. But the experience of markets themselves in the post-war period threw up an intriguing paradox: if markets take care of themselves, why is it that, although between 1950 and the late 1980s we observed asset bubbles and crashes, from the 1990s onwards the frequency of such bubbles increased? After all, shouldn't globalisation and greater interconnectedness have made the markets yet more efficient? But as Figure 1 shows, the number of market 'failures' rose the more globalised we became.

Figure 1 Financial market bubbles

Professor Fama's EMT, written in the 1960s, triggered a heated debate between market practitioners and academics from the start. The core philosophy of EMT is that all information is priced into every asset at each moment of time and therefore it is impossible to create a portfolio that would outperform a randomly chosen portfolio with an equal risk profile. Only by adding additional risk can one beat the market.

But this in itself undermines the theory: if this was possible, one is stating that there are periods of moving in and out of assets where, at inception, not the full upside is priced in, and when this materialises one moves out of the asset. However, at inception this asset would have been mispriced and thus by definition not all information is embedded in the price.

EMT continues by arguing that even if one could perform better than a randomly chosen portfolio, it would be impossible to do this on an individual basis. As a consequence, the theory concludes that markets are not predictable, and this makes them ultimately efficient. Therefore the fact that markets are not predictable, as this is another conclusion, in turn means markets are efficient. If markets are unpredictable, one cannot beat the market, because there is no possible upside in hand-picking assets: all assets are correctly priced at all times.
I surmise that this thesis has damaged the financial industry over time. A failure of belief in that theory occurred in 1998 when the hedge fund LTCM collapsed under the weight of its excessively leveraged positions. The fund’s managers believed that they were smarter than the market, backed no doubt by their undoubtedly impressive academic track record.

Strongly believing in the EMT, LTCM was convinced it could take advantage of the inefficiencies of the market. In fact the market was actually taking advantage of LTCM’s own inefficiencies. The mistake the fund managers made was that they believed in the mathematical modelling approach of financial markets, which is not always accurate – it is particularly inefficient in allowing for outlier events, something that Mr Nassim Taleb has pointed out in *Black Swan* – and is actually a completely different subject.

Notwithstanding the LTCM experience, the trend in approaching markets more scientifically continued and soon many bank dealing rooms included mathematicians and financial engineers. These individuals, the so-called ‘quants’, possessed a similar background and invariably use the same models, mathematical approaches that originated in statistical physics.

A similar approach was made by the credit rating agencies. These firms implicitly assumed that real estate prices could never fall on a national scale, simply because statistical evidence suggested this was so. Economic shocks or changes in economic states were typically ignored, since random behaviour can be used to expect future states. If enough market participants ignore the same states of the world, the market price is automatically incomplete, thus not all information is priced in, and hence markets are not efficient.

The danger is however, and this came to the surface during the crash of 2008-2009, that if every derivative trader is using the same model, everybody is going to take the same price (which is produced by the model) for granted. The issue here is that at the moment the market drops out of the comfort zone or interval of what the model is used to, one is faced with substantial systemic risk. An analogy would be this: as long as all the cars on the highway are driving at 100 mph, there is nothing wrong; but from the moment that one of them has to stop then this causes serious disruption.

As EMT discourages an investor from deviating from the benchmark, it has a strong influence on how the financial industry is obsessed with benchmark assessment, unless one adds new risks. One could argue that the fact that many, if not all, large investors follow the same principle increases the inefficiency of the market. It potentially creates bubbles, since there is too much money chasing the same type of assets and all institutions have virtually the same type of portfolio and are all vulnerable to the same type of risk. If this type of risk emerges, the very scale of these institutions tends to transform a potentially minor risk into a major event due to the same end-reaction of all investors.

A benchmarking approach becomes a self-fulfilling prophecy as investors do not want to jeopardise their career prospects and move outside the benchmark. But just as diversification was highly criticised immediately after the crash, so-called ‘alpha’ was criticised as well. Consider Figure 2 which shows the performance of the hedge fund industry at end-2008.

We observe that almost every hedge fund manager lost money during that year. All the strategies shown (except for dedicated shorts and managed futures) showed a negative performance for 2008. We can argue that both dedicated shorts and managed futures are pure directional plays, like betting in a casino, and anticipate a negative downturn, and so would always perform positive in the environment of 2008. Such strategies do not represent the application of Markowitz and EMT. The immediate circumstantial evidence from Figure 2 suggests that ‘alpha’ is a myth as well. But this issue would also suggest that markets are efficient, and I am not entirely convinced that this is the case.

In any case, a Nobel Prize awaits the person or persons who articulate a model that eclipses Markowitz and EMT.

I would like to thank the Editorial Panel for their involvement, the CISI, the contributing authors and of course you the reader. I hope everyone has as much fun reading this as I had working on it.

Professor Moorad Choudhry FCSI, Editor
A REVIEW OF HIGH-FREQUENCY TRADING & FLASH CRASHES
Chyng Wen Tee, Assistant Professor and Christopher Ting, Associate Professor
LKC School of Business, Singapore Management University
cwt@smu.edu.sg

ABSTRACT
We review recent literature and discussions surrounding the 6 May 2010 Flash Crash, the implications on high-frequency trading (HFT), and flash crashes in general. Different interpretations of the event lead to a split opinion on whether HFT should bear the main responsibility for the crash. The conclusions of two main schools of thought are presented, followed by an in-depth exposition of recent work on mini flash crashes.

INTRODUCTION
In 1971, even before the celebrated Black-Scholes formula was published, Fischer Black was already thinking about automation and computer trading systems (Black, 1971). In a two-part article titled 'Toward a Fully Automated Stock Exchange', he provided an intuitive definition of liquidity (existence of a two-way market, small bid/offer spread, execution volume's impact on market price, etc), and asserted that a liquid market is both continuous and efficient. He went on to make the prescient argument that executions of sizeable order volume will always exert pressure on price, regardless of method or technological advances, and that price continuity is guaranteed only when trading is characterised by volumes of small, individual trades.

6 May 2010 Flash Crash
Today, we live in a world where high-frequency trading (HFT) is increasingly gaining prominence as the financial market continues to move towards higher level of automation by leveraging on technological advancements and pursuing algorithms with growing sophistication. Over the past few years, there have been a number of sizeable flash crashes, with the 6 May 2010 Flash Crash being the most notable and obvious one. On this day, the US stock market experienced one of its most severe intra-day price drops in history, with the Dow Jones Industrial Average (DJIA) index tumbling 900 points within a time span of 20 minutes before rebounding and recovering much of its losses. Several stocks (eg, ACN, CNP, EXC) crashed within the space of a few seconds and traded at a penny before shooting back up to their prevailing market prices.

Huge effort has since been exerted to make sense of all this, with academics, regulators and practitioners in the financial market teaming up to conduct in-depth investigation and analyses of the Flash Crash. Although the crash took place within a single day, to gain a qualitative and quantitative understanding of it is no easy task. The trading activity in the US equity market is split among over 13 public exchanges, more than 30 dark pools and over 200 internalising broker-dealers, making it difficult for all participants of the capital markets to access liquidity in an organised manner. It should be no surprise that the sheer scale of the market and its complex dynamics and interactions render any investigation effort extremely challenging.

Should we blame HFT? Opinions are split between two main conclusions

Despite the difficulties, more than three years after the 2010 Flash Crash, we are now in possession of a respectable volume of literature covering the events leading up to it, along with numerous ‘post-mortem’ analyses to investigate what went wrong in the system to cause such an extreme event. The key question which every researcher tried to answer is obvious: ‘Should we blame HFT for the flash crash?’ We surveyed the literature in this field and found that opinions are split between two main conclusions. One school of thought concluded that HFT is the culprit, while the other concluded that HFT should be credited with improving market conditions, and that human errors are largely to blame for flash crashes. To get a feel for these contradicting views and opinions, let’s compare these two magazine articles in November 2013, one in Bloomberg Business Week and the other in Financial Post, published just one week away from each other. The first article, titled ‘The Nemesis of High-Speed Traders’, vilifies HFT and makes harsh criticisms about it, while the second article, titled ‘In Praise of High-Frequency Traders’, praises the practice and points out that the market as a whole is a better place thanks to the participation of HF traders. As a concise summary: the school of thought critical of HFT thinks that HFT poses a threat to all other market participants because traders react so quickly to price movements, to the extent that they are either consciously using their speed to gain unfair advantage, or impetuously reacting to price movement, which might potentially cause the market to spiral out of control.

The other school of thought uses empirical evidence to show that the presence of HFT leads to improvements in the market microstructure by reducing price volatility and shrinking bid/offer spreads. Furthermore, it argues that even when some of the high-frequency (HF) traders are ‘informed’ traders, their presence in the market actually leads to improved price discovery.

**FLASH CRASH & HFTs**

With that in mind, let’s take a high level review of a few key interpretations of the 2010 Flash Crash. The ‘official’ version is based on the report published by the US Commodity Futures Trading Commission and Securities & Exchange Commission. In this version, the source or trigger of the crash is traced to Waddell & Reed (W&R), an asset management and financial company applying fundamental trading strategies. On 6 May 2010, W&R placed a large sell order of 75,000 June 2010 E-mini contracts as a hedge via Barclays’ automated execution algorithm. The order was placed with an execution rate of 9% of trading volume (over the previous minute). This was a precautionary measure to avoid causing adverse market movement. On any normal business day this would have been fine and gone through smoothly. Unfortunately, this took place on an already jittery day carrying negative market sentiment due to concerns arising from the European sovereign debt crisis. The large sell order was initially absorbed by HF traders and intermediaries. However, when they in turn tried to unload their inventory not much later on, liquidity dried up and volatility shot up, snowballing into the infamous Flash Crash as we know it. The findings in the SEC report are partly based on Kirilenko, Kyle, Samadi and Tuzun (2011), whose main conclusion is that HFT was not the trigger of the Flash Crash. Rather, it was the responses to the large selling pressure on that day which exacerbated and aggravated market volatility. To cut a long story short – the large sell order was the trigger of the whole Flash Crash, its effect cascading and tumbling down the system, dragging the whole US financial market along. For a bird’s-eye view of recent progress and development in this field, we refer the reader to Kirilenko and Lo (2013).

On the other hand, the opposite school of thought maintains that W&R never drained liquidity away from the market. This points to the fact that the sell orders W&R posted via Barclays’ algorithm never crossed over the bid/offer spread to hit a bid. Instead, they were always posted above the market and executed with prudence, taking the effort to minimise market impact by stopping trade execution on downward market movement. It was the buyers of these future who were reckless when they began to unload their accumulated position. Unlike W&R, they agitated the market with successive bursts of sell orders hitting the bids in the order book, draining liquidity away from an already fragile market and adding further strain to it. Consequently, W&R was not at all the source of the 2010 Flash Crash. Instead, HF traders and intermediaries should bear the majority of the responsibility, since they took liquidity away from the market, causing it to spiral downwards. In short, HF traders caused the Flash Crash.

Despite a further number of contrary opinions published after the release of their original report, the SEC has stood by its report and findings. Its position is that these crashes, especially on single stocks, are generally caused by human error, instead of malicious HFT algorithms wreaking havoc on the market. In its opinion, what is more disturbing is the chronic sloppiness and the lack of due diligence in the human agents maintaining the HFT systems. To support its position, the SEC cited Menkveld and Zhou (2013), one of the latest examples of independent academic research, whose findings and conclusions are consistent with its own. In this paper, the authors used empirical studies to show how a large sell order in the futures market trickled down into the whole financial market. The conclusion of this work is in the same vein as the SEC’s: that the blame can’t be put on a single market agent. The Flash Crash was the product of the interaction and dynamics among market participants. As a result, it’s not a trivial task to formulate a clear-cut recommendation. More work is required to understand the interaction mechanism and the potential pitfalls. This is consistent with the SEC’s finding that W&R’s execution of the large sell order indirectly caused the Flash Crash through the response of other market participants.

In general, the majority of the academic research is sympathetic towards HF traders and finds them benign. For instance, in a highly cited paper, Hendershott, Jones and Menkveld (2011) investigated the empirical relationship between high-speed algorithmic trading and liquidity. They found that for large stocks in particular, algorithmic trading can be credited with narrowing bid/offer spreads and reducing adverse selection. These go to show that algorithmic trading is capable of improving liquidity and enhancing the information content of market quotes. More importantly, algorithmic trading contributes to enhanced price discovery.
without trading by increasing the information content of the quotes. This highlights the importance of algorithmic traders as liquidity suppliers, given their role in lowering the costs of trading and increasing the informativeness of live quotes. In another similar study, Hasbrouck and Saar (2013) investigated the impact of low-latency trading activities on the quality of the market under both normal and stressed market condition. They concentrated on market events in the millisecond environment, and discovered that increased low-latency trading activities lead to an improvement in liquidity and reduced short-term volatility, suggesting that increased low-latency activities do not invariably cause a detrimental effect on fundamental investors and other market participants. Even more interesting is the fact that the same conclusion applies whether the market is under stress or not.

The endeavour to identify the cause and implication of flash crashes aside, preventive and contingency measures have also been proposed and tested for robustness. In a speech by then SEC Chairperson Mary Schapiro, she called for a re-examination of the circuit breaker mechanisms that directly limit price volatility. Nevertheless, circuit breakers should be considered only as a last resort. From a practical perspective, it is a lot better to promote and ensure reliable price discovery with effective monitoring, and only fall back to circuit breakers as a fail-safe mechanism. One popular method proposed to address this issue is the Volume-Synchronised Probability of Informed Trading (VPIN) flow toxicity metric proposed by Easley, Lopez de Prado and O’Hara (2012). This method monitors flow toxicity and has significant forecasting power over toxicity-induced volatility. To show that VPIN is a useful indicator of short-term, toxicity-induced volatility, the authors applied this metric to the period just before the 2010 Flash Crash and demonstrated that their metric is capable of correctly capturing the increasing toxicity of the order flow prior to the May 2010 Flash Crash.

FROM FLASH CRASH TO MINI FLASH CRASH

Ever since the major 2010 event, flash crashes have been receiving a constant stream of publicity due to their incessant occurrence. We list below a handful of recent ones which have received journalistic coverage:

- 23 April 2013: a false (and worrying) tweet caused the Dow Jones to quickly plunge 140 points, or roughly 1%, before bouncing back9
- 25 July 2013: Whirlpool Corp’s stock rocketed 5% higher, only to fall back to the previous level within seconds10
- 30 Aug 2013: an errant trade caused P&G’s stock to plunge by 5% before instantly snapping back within that minute.11

Given the polarised nature of the general consensus, it seems certain that flash crashes will continue to be investigated by academics and practitioners alike. The ‘scarcity’ of major crashes implies that theories formulated on these events are difficult to test and validate. To find a more common platform for testing, we need to broaden our horizon by considering all the ‘mini’ or ‘micro’ flash crashes, ie, crashes which might go past unnoticed unless we actively look for them. Nanex, a trading technology company and data service provider, proposed a rule-of-thumb definition for what qualifies as a mini flash crash:

- Uni-directional tick for at least ten times
- Time window does not exceed 1.5 seconds
- Price change exceeds 0.8%.

Using these criteria, it scanned its database and found a surprisingly large number of these12. This obviously stirs up a reasonable amount of research interest, as it allows us to analyse a large sample of crash data in bulk, as opposed to focusing on sporadic major crashes, which are fortunately relatively rare.

INVESTIGATIONS OF MINI FLASH CRASHES

What causes all these ‘mini’ flash crashes? If we view a major flash crash as the amplified version of its ‘mini’ counterpart, then it’s reasonable to expect that a good understanding of these events, which are in abundance, can arguably lead to a deeper insight of any potential weakness in the current system. Once again, opinions are split. In a recent paper, Johnson & co-authors (2012) suggest that these crashes may be a result of interaction among several trading algorithms, or a positive feedback loop induced by market environment. Their paper highlights the inherent danger in the transition from a mixed human-machine to all-machine ecology, which is beyond the limits of human response times. A fundamental revamp in regulation is necessary to tackle these ultra-fast extreme events across a wide class of systems. They also point out that effective regulation is impossible without a quantitative

---

8 Order flow is said to be toxic from a liquidity provider’s perspective if the people you trade against have a better idea than you about the fundamental ‘true’ price of the asset.

9 http://money.cnn.com/2013/04/24/investing/twitter-flash-crash/

10 http://www.nanex.net/aqck2/4379.html


12 Nanex Research, http://www.nanex.net/FlashCrashEquities/FlashCrashAnalysis_Equities.html
understanding of the dynamics, and a scientific theory for the underlying human-machine ecology.

On the other hand, another recent study conducted by Golub, Keane and Poon (2012) on a comparable set of mini flash crashes identified another possible cause of these events – the Intermarket Sweep Order13 (ISO) exemption to the Order Protection Rule14 in the new market Regulation NMS15. They found evidence pointing to the fact that mini flash crashes are the result of regulation framework and market fragmentation, in particular due to the aggressive use of ISO, with Regulation NMS protecting only the top of the book. It is not difficult to envisage the detrimental effect of ISO on the market microstructure when it is abused and used aggressively. In principle, this exemption is meant for large orders to sweep through the order book, yet in practice it could be exploited by HF traders attempting to outpace the Stock Information Provider16 (SIP). For instance, consider a pernicious HFT algorithm sending an Immediate-Or-Cancel17 (IOC) ISO to a list of exchanges on the same stock. If the SIP is slow, its NBBO could be pointing to a stale quote which no longer exists when the order book is swept. A HF trader using ISO orders will therefore be able to snap up liquidities while the rest of the market participants conforming to the trade-through test will have their orders rerouted and rejected due to the quotes being stale and invalid.

Naturally, the impact of the new regulatory system on the financial market has also become the focus of growing research effort. An interesting study carried out by Chakravarty, Jain, Upson and Wood (2012) on the intermarket sweep order statistic of 120 equities listed on NYSE or NASDAQ found that ISO orders represent 47% of the 587 million trades and 42% of the 167 billion shares traded in the sample under study. These proportions remain consistent for all capitalisation segments. Using the definition of information content defined in Hasbrouck (1995), the authors also investigated if ISO orders are used by informed traders when there is information in the market. If the information share of ISO orders is higher on days with high information, then their interpretation of ISO orders as the preferred order of informed traders is supported. The overall finding is that ISO orders are indeed dominated by informed traders, trading on time sensitive information. This conclusion complements the point of view that aggressive use of ISO might potentially be the cause of the mini crashes we observed.

In addition to all this work, we’re also beginning to see an increasing amount of research going beyond empirical studies and focusing on simulating trading activities and studying the dynamics and interactions among market participants. For instance, in a recent study, Brewer, Cvitanic and Plott (2012) used a simulation approach to study order flow and its interaction with the market microstructure. They simulated random arrival of buyers and sellers to the market submitting bid or ask orders. They then artificially induced a flash crash by the submission of an extremely large order, and analysed the impact on both the liquidity and the stability of the market. They used this as a platform to evaluate the efficiency of different methods typically used to restore liquidity to the market immediately after a crash, including double-auction, trading halt and call auction. While taking a simulation-based approach to study the impact of HFT isn’t without its critics18, it does seem like the industry on the whole is open to using innovative ways to study market interactions and is committed to invest in this direction19.

Broadening our attention to include mini or micro flash crashes allows us to see the bigger picture

CONCLUSIONS

We have now entered the era where high-speed machines and algorithms are operating in a time scale that is well beyond the limit of human response time. The debates and research in this space are expected to go on for time to come. Is HFT a force for good or evil? As far as we can gather, the jury is still out. Nonetheless, whichever school of thought you are in and whatever your theory is, we believe we all agree that more work needs to be done to gain a better understanding on the ultrafast dynamics. Only then can we expect to form an impartial and objective opinion on this fundamental question, and use that as a basis to formulate an effective and prudent regulatory framework.

13 Intermarket Sweep Order: Sweep through several exchanges and execute as many buy or sell orders as possible. The responsibility falls on the sender to ensure that the trade-through test is satisfied.
14 A rule to ensure price priority and protects quotes at the top of the book.
15 A set of rules formulated with the aim to improve fairness in transaction and effective quotes dissemination.
16 A processor maintained by NASDAQ which constructs a National Best Bid and Offer (NBBO) from all US equity exchanges.
17 An exchange receiving an Immediate-Or-Cancel ISO order will either execute immediately or cancel it without having to check on potential trade-through of protected quote
REFERENCES


ABSTRACT
Retail investment advisers rely increasingly on outsourced or in-house model portfolios which, in the absence of industry-standard datasets, tend to be exemplified by data from within the last 20-30 years. Unless data are selected from periods (eg, pre-1975) which include bear markets in all major financial assets, expectations may be over-optimistic. The use of average return and variance data obscures differences among different assets and ignores the cyclical nature of all asset markets, making automatic rebalancing of portfolios irrational, problematical and almost certainly sub-optimal.

LITERATURE REVIEW
This paper briefly examines the principles underlying modern financial theory (MFT) and analyses logical and practical flaws in its current application. The genesis of MFT can be found in pre-war analytical work (eg, Williams 1938; von Neumann & Morgenstern 1944) developed by Harry Markowitz (Markowitz 1952, 1959), via modern portfolio theory (MPT), into quantitative finance. CAPM, the capital asset pricing model (Treynor 1961, 1962; Sharpe 1964; Lintner 1965), and the efficient market hypothesis (Fama 1965, 1970; Samuelson 1965) were joined finally by the Black-Scholes option pricing formula (Black & Scholes 1973) to complete the key elements comprising MFT. While currently still entrenched in the teaching of finance and investment, all parts of MFT have been under attack since at least the 1980s, though contradictory evidence had been around since the early 1960s, principally Benoit Mandelbrot’s discovery of power law distributions in a range of financial market data (eg, Mandelbrot 1963, 1964 etc.). The aftermath of the LTCM fiasco and the bursting of the dot.com bubble caused a massive reassessment and criticism of the underlying assumptions and practical application of MFT (eg, Shiller 1992, 2005; Shleifer 2000; Lo & MacKinlay 2002; Sornette 2004; Mandelbrot 2004, 2008; Taleb 2007a, 2007b; Fox 2010, etc.)
INTRODUCTION

In retail markets, the ubiquitous model portfolio approach to asset allocation, with systematic rebalancing, seems unimpeachable. However, this paper challenges that consensus, and argues that practitioners should be aware of the dangers of thinking in averages, and be open to the reality of the cyclical nature of asset prices and the resulting implications for investment theory.

CURRENT COMMON PRACTICE

The use of model portfolios with rebalancing is predicated on a number of assumptions and beliefs:

- Diversification is ‘a good thing’, derived from modern portfolio theory (MPT).
- Asset allocation is the primary determinant of long-term returns.
- Some pro-forma and empirical evidence indicate diversified portfolios can, on average, generate superior returns with lower volatility over the long term compared to any one component asset class.
- “Nobody can predict the market,” so automatic periodic rebalancing avoids subjective judgment and ensures the ‘policy’ asset allocation is maintained fairly constantly.

Providers of model portfolio ‘solutions’ cite historical performance data showing generally enviable outcomes from a selected dataset. It is typically implied that this general methodology is robust and ‘scientific’. A range of portfolio asset allocations is constructed across a range of ‘risk profiles’ which are matched to each client’s score on some species of risk-profile spectrum – typically from a questionnaire. This process is believed to meet the regulator’s ‘suitability’ criteria.

Model portfolio asset allocations vary significantly among providers across a range of ‘risk profiles’, ‘risk tolerances’ or other metric. This is caused by subjective data-selection, evidenced by the fact that the range of datasets, asset classes, as well the asset allocations for each ‘risk level’ is remarkably varied among providers. Datasets can be selected from periods which either exclude severe or extreme market events, or use long data points which tend to generate a normal distribution. Effectively each provider’s ‘solution’ is different, while a genuinely scientific approach would not be expected to generate such a heterogeneous mix.

For an adviser, whether or not outsourcing this function, selecting a suitable asset allocation is a subjective decision to pick from a subjectively created range of model portfolios. It is not a scientific or objective approach to managing clients’ assets, but a doubly subjective choice.

WHERE DID MODEL PORTFOLIOS COME FROM?

The concept of model portfolios, as well as rebalancing, is derived directly from Harry Markowitz’s modern portfolio theory. Markowitz was aiming to find a mathematical way of selecting the best shares to mix together into a stock market portfolio. To do this he adopted a reductionist approach, to keep the maths relatively simple, based on two metrics: the mean return and the variance from that mean of each stock for selection. To those two factors he added the covariance, or correlation, between every pair of stocks because, if the component stocks correlated with each other perfectly, it would be no better than selecting just a single stock.

Markowitz’s logic was, in a portfolio you want to hold the stocks which have historically given the best returns because it is reasonable to expect them to do so in the future. Similarly, you don’t want stocks that go up and down in perfect synchronicity because that would appear to negate the diversification effect, and it is better to hold stocks whose prices have historically tended to vary from the average as little as possible, because there may be a smaller chance they will suffer large falls relative to the portfolio or market. For all this to work mathematically he assumed a linear relationship between return and ‘risk’; returns from financial markets formed a Gaussian (bell curve) distribution and settled on the standard deviation (volatility) – the square root of the variance – as a ‘catch-all’ for risk.

The problem of how much of each stock to put into the portfolio was addressed by mean-variance optimisation (MVO). Other things being equal, you would want to weight the portfolio to the stock(s) which had the best returns, subject to their not being too volatile and not correlating too much with any other component stocks. MVO is a piece of iterative mathematics where, knowing the mean return, the variance (volatility), and the correlations between each pair of component stocks, you add a bit more of this or a bit less of that, until you find the mix of shares with the lowest correlations but highest weighted average return, for each level of volatility shared by the components. This mean-variance optimisation produces an ‘efficient portfolio’. By repeating the process for the entire range of volatilities in the data you get a set of portfolios, each with different asset weightings and, by joining the dots together graphically, with return up the y-axis and volatility along the x-axis, you plot the well-known ‘efficient frontier’. The data used for these purposes are not standard, but selected subjectively by each portfolio manager.

1 For quantitative finance purposes the key metrics are mean return (μ), volatility/risk (σ), and the correlations between each pair of assets. Remarkably, there is no agreement on what dataset (or data intervals) truly reflects the inherent characteristic values of (μ) and (σ) for any asset; correlations are notoriously unstable. Each manager is therefore free to select data which demonstrate the point he wants to make. This is data-mining.
CAN MPT BE EXTENDED TO MULTI-ASSET PORTFOLIOS?

Model portfolios are derived from MVO principles, with the added theoretical assumption that, if MPT works for stocks, it will work for a mix of different asset classes also, though Markowitz himself did not make such a claim. This leap of faith is derived from the Law of One Price. It is a fundamental principle of classical economics and states that identical goods should share an identical price. Sellers will tend to gravitate to the highest price prevailing and buyers to the lowest; in an efficient market these will almost instantly converge to the ‘market price’. This may work for apples and automobiles, but whether it holds true for the behaviour of different financial assets may be doubtful. Given very different long-term real returns and volatilities attaching to public equities, property, bonds, cash and commodities, etc, it is challenging to accept that sharing a common volatility means you can expect to get the same return from every asset in the mix.

For example, a bond is a contract enforceable in law – an undertaking by the issuer to deliver regular interest payments and return of nominal capital at redemption. If you own property you possess legal title and you can expect both growth in real value over time and a continuing stream of future inflation-linked cash flows from rents. Equities, by contrast, entitle you to absolutely nothing with certainty. You may hope for dividends and capital growth, but you cannot sue the company if it fails to provide either. Given these fundamental differences, it is hard to reason that a shared level of subjectively selected volatility means all asset types share identical levels of risk and return.

The methodology of model portfolios relies on averages – but ignores the cyclical nature of markets in the real world

Even if that approach did apply perfectly to all asset classes, the assumptions needed for the whole of modern financial theory (MFT) to work are highly problematical. MFT comprises theories, all of which rely fundamentally on the assumptions that financial markets are ‘frictionless’ and costless with no chance of market impact, agents act rationally with relevant information instantaneously universally available, all price changes are random and independent (leaving bubbles and crashes unexplained), and the price changes or returns from all financial markets form a normal distribution. These assumptions have been increasingly criticised and largely discredited.

The use of volatility as ‘risk’ is central to MFT, and feeds directly into the mathematical (as opposed to common sense) theory of diversification. MFT describes the character of any investment with just two metrics – its volatility and expected return. Combining the tenets of MFT with the Law of One Price leads logically to the notion that any two investments with equal volatility will have the same future expected return and vice versa. Thus, in the assumed efficient markets of MFT, future price behaviour of all securities is entirely described by their expected return (μ) and volatility (σ) and any two securities with the same foreseeable volatility should have the same expected return (Derman 2002). Since only the values of μ and σ are required to describe a security, it must follow that higher expected returns can be captured only by accepting higher volatility/risk and the corollary, accepting higher risk, brings the expectation of higher returns. This is a central tenet of the efficient market hypothesis: that the only way to ‘beat the market’ is to take on more risk than the market average.

This theory crystallises as high volatility = high expected return, and low volatility = low expected return. If volatility truly reflects all risk, that relationship between risk and reward might be rational. Volatility, even though expedient for the original construction of MPT, does not in fact exhibit a linear relationship to returns when tested empirically; the exact opposite is true. The evidence is voluminous and compelling (eg, Fama & French 1992, 2004; Haugen 2002, 2010; Ang et al 2006; Clarke et al 2006; Montier 2009). In exact contradiction to theory, low volatility stocks (typically ‘value’ stocks measured on a range of metrics and reflecting relative cheapness in the bottom, say, 10% or 20% of the stock universe) in aggregate tend to generate higher future growth rates with less volatility than high volatility growth stocks (top 10% or 20% of stock universe) which tend to significantly lower future growth rates. It is therefore not necessary to take on higher risk in order to obtain higher expected returns; stock selection among low volatility value stocks can be expected to provide higher returns with less volatility over time. Nevertheless, model portfolios will weight progressively to equities as a class (without further refinement) as risk appetite increases, and to bonds as risk aversion increases.

However, the methodology of model portfolios relies on averages – (μ) and (σ), plus correlations which are unstable and unreliable – but ignores the cyclical nature of markets in the real world. Given the present macro-economic background, it might be unwise to assume that interest rates will never rise. If they did, then falls in both equity and bond markets would be unsurprising: what then for risk-averse clients who have been heavily weighted to bonds? To uncritically rely on the theory that historical long-term averages will persist into the future may bring nasty shocks for clients.
AUTOMATIC REBALANCING – GOOD ONLY FOR CLUELESS MANAGERS

It is assumed that automatic rebalancing of diversified portfolios avoids market timing risk. Ignoring the fact that each manager likely has a different asset allocation, the rationale is that the policy (initial) portfolio is optimal and, if some component(s) rise or fall faster than the others by value, then an appropriate portion of that holding is sold or bought in order to return the asset allocation to policy levels by value. This process generates conflicts of interest through dealing charges and commissions; one major platform reports that over 75% of all deals are now placed by managers of model portfolios. Logically, the process is almost certain to be sub-optimal and makes no sense unless an investment manager is literally clueless about what else to do.

The future return from any investment is determined by a single factor – the entry point into the asset price cycle

Prices of all real asset classes and markets are cyclical: all go up and down over time, either in real terms or relative to ‘fair value’, and some markets display mean reversion. The future return from any investment is determined by a single factor – the entry point into the asset price cycle. If you buy at a relatively low point in the cycle you can reasonably expect to do quite well in the future; if you buy near the top, you will probably have to wait a very long time for a real return, and may face hefty losses in the meantime. If he has no idea of the current point in any asset price cycle then rebalancing is probably the only thing a manager can do, but it is almost certain to result in a sub-optimal outcome.

Consider a diversified portfolio equally allocated to bonds, equities and property – i.e. 33½ % in each – to ‘match’ a given level of desired risk or ‘risk profile’. Initially you have no view as to where any of these markets is in its price cycle and so have a policy of rebalancing every, say, three or six months. Over an investing life of perhaps 20 years or more, assume you will experience at least one full cycle in each asset class, but that the cycles are not necessarily synchronous or of equal length. As each asset is in the rising part of its cycle (relative to at least one of the other assets), you will be selling and reducing your holdings, although by value you will always rebalance to 33½ % of the portfolio, whatever its total value is. The strategy automatically decreases your average selling price, and reduces your ability to stay with the trend and run profits, while holdings become ever smaller over time. You may have very little left at the top of the cycle, and be forced by policy to start buying at top prices immediately after that market peak.

Conversely, in a bear market you will force yourself to start buying too soon while prices are still falling, this time raising your average buying price so you may have to wait, after the bottom is reached, for a significant market rise before your average entry price is reached. Except you might never get there, because you would start selling as soon as the asset price turns up. Follow this strategy long enough, and you may count it a success to finish up with your original nominal capital. It is entirely feasible, particularly with dealing costs added, to systematically lose money over time.

The process is sub-optimal, in aggregate selling too soon in a rising market and buying too soon in a falling market: you will always be on the wrong side of the trend. The only circumstance in which you will do well is if a market reverses sharply between one rebalancing and the next, when there is a unique possibility you may have bought immediately after a market has bottomed or sold just after the top. However, if the new trends continue, you will be back on the treadmill of selling and buying too soon, quite possibly selling below your average buying price, and buying above your average selling price. To meet TCF

2 requirements, these possibilities should be explained clearly to investors.

Although the very suggestion that automatic rebalancing is irrational will be seen as heretical by some practitioners, the empirical evidence is clear. These strategies first started to become used around the late 1980s and early 1990s, when computing power and MVO software became accessible. Using data from the preceding ten or so years, the best-performing asset had been Japanese equities so, in accord with MFT, optimisation engines weighted heavily to these assets. In a quantitative world of efficient markets, that asset allocation was rational and could be expected to generate the highest future returns. Had markets continued as previously, optimisation would have allocated more to Japan but, as it happens, from 1990 the asset started more than a decade of steep decline.

Sticking to the theory, rebalancing say every six months, produced an interesting effect. At each rebalancing point, the dataset was rolled forward six months to incorporate recent market returns. This being the case, the mathematics found Japan to have slightly lower mean return values after six months, and so resulted in some Japanese holdings being sold and re-allocated elsewhere. At the next rebalancing that happened again, and the next, and so on. By the millennium, Markowitz’s portfolio theory and optimisation methodology had achieved the remarkable feat of being long at the top of the Japanese market and selling steadily all the way down, until finally such optimised portfolios were left with effectively no holdings in Japan at the bottom.

2 TCF – ‘treating customers fairly’ is a high-level regulatory principle in the UK.
Optimisation sounds more impressive than systematic rebalancing of standard or model portfolios, but they share essentially similar flaws. Both rely, as the principal determinant of asset allocation, on the mean return. But the mean is a lagging value, always behind the curve of what the market is doing now. An optimisation approach must logically result in always playing ‘catch-up’, always failing to buy more when cheap, chasing the price up, and then selling on the downward leg. In terms of the average buying price and selling price over a market cycle, this is obviously inefficient.

A model portfolio approach may avoid the most egregious consequences of optimisation, but at a price: everything depends on the initial policy asset allocation being ‘right’ in the future, and automatic rebalancing destroys the very possibility of benefiting from cyclical market trends, either up or down. It compels you to sell too soon in a rising market and to buy too soon in a falling market.

**WHAT CAN BE DONE?**

Before acquiescing to purveyors of model portfolios who insist market timing is impossible, and that a diversified ‘risk-targeted’ model portfolio is a better bet, practitioners must pause to consider. Timing the tops and bottoms of markets may be practically impossible, but that is not the same as denying the possibility of a rational assessment of where we are in asset price cycles. There are fundamental valuation methods available for some asset markets and common sense for others. It may not be sensible to buy an asset class on the basis of average expected return; in the upper parts of its cycle an asset will almost certainly generate below-average future returns. There can be rational reasons for not buying, even if that means not following the consensus.

**CONCLUSION**

Professional training and exams continue to embrace orthodox modern financial theory, now more than half a century old and still predominant. The tyranny of averages means students and practitioners can easily overlook the cyclical nature of financial markets, missing rational buying opportunities and sustaining the Panglossian myth that every day is a good day to buy, even though for extended periods buying will almost certainly result in below-average long-term returns. Society and investors might benefit if examination curricula and syllabuses required the teaching and examining of the weaknesses in MFT, asset price behaviour, and the economic history of financial markets.

**REFERENCES:**


ABOUT THE EDITOR

Professor Moorad Choudhry FCSI is IPO Treasurer at the Royal Bank of Scotland.

He is Visiting Professor at the Department of Mathematical Sciences, Brunel University; Visiting Professor, Department of Economics, London Metropolitan University and Visiting Teaching Fellow, Department of Management, Birkbeck, University of London.

Professor Choudhry is Managing Editor of the International Journal of Monetary Economics and Finance and on the Editorial Board of Qualitative Research in Financial Markets. He is also a member of the CISI Editorial Panel, which chooses key content for the Securities & Investment Review. He is author of The Principles of Banking (Wiley 2012).

mooradchoudhry@gmail.com

ACADEMIC JOURNAL PANEL MEMBERS

Prof Moorad Choudhry FCSI
Prof Carol Alexander
Dr Edward Bace MCSI
Dr Paul Cox
Scott Dobbie FCSI(Hon)
Peter Land, Chartered FCSI
Gino Landuyt
Prof Donald Lawrence MCSI
Sian Lloyd, Chartered FCSI
Gregor Logan MCSI
David Moskovic
Prof Jim Steeley
Nigel Sydenham, Chartered FCSI

Royal Bank of Scotland
University of Sussex
CISI, Middlesex University
University of Birmingham
Deutsche Bank
Brewin Dolphin
Luxembourg Financial Group
University College London
CISI
Family Investment, Nutmeg
Royal Bank of Scotland
Aston Business School
BPP Professional Education

SUBMISSION GUIDELINES

CISI members are invited to submit to the Institute for consideration papers on any aspect of wealth management, capital markets and banking.

Articles must be:
• Original work and previously unpublished
• Between 1,500 and 3,500 words in length and accompanied by an Abstract of 80-150 words.

All papers submitted will be refereed by the journal editorial panel or its recommended reviewers.

For further details about the Review of Financial Markets and how to submit articles, see cisi.org/academic

HAVE YOUR SAY

If you would you like to comment on any of the articles in this issue, contact CISI Communications Editor Richard Mitchell: email richard.mitchell@cisi.org or call +44 20 7645 0749