MARKOV-SWITCHING MODELS AND RESULTANT EQUITY IMPLIED VOLATILITY SURFACES: A SOUTH AFRICAN APPLICATION

By Mark Fairbrother

(FRBMAR004)

Dissertation for the purposes of completion of MSc in Mathematics of Finance

Supervisor: Professor Ronald Becker

FIRST SUBMITTED VERSION

LAST EDITED ON:

12/11/2012 09:45 AM

NOTES

The dissertation is submitted together with a CD containing the MS Excel models developed for parameter estimation and option pricing. Contained on the CD is also an electronic version of this document.

COPYRIGHT DECLARATION:

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is my own.
2. Each significant contribution to, and quotation in this dissertation, from the work(s) of other people has been attributed, and has been cited and referenced.
3. This dissertation is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

I know the meaning of Plagiarism and declare that all of the work in the document, save for that which is properly acknowledged, is my own

Signature: ________________________

Date: ________________________

ABSTRACT

Standard Geometric Brownian Motion is the stock model underlying Black-Scholes famous option pricing formula. There are however numerous problems with this stock model as certain features do not follow some empirical stylised facts we see from the observation of actual asset prices. In particular, the constant parameter idea behind Geometric Brownian Motion is flawed. It is argued that information flow dictates stock price movements and information is a function macro-economic regimes shifts. As such, we propose an alternative model, one in which the parameters in the Standard Geometric Brownian Motion change according to an underlying Hidden Markov Process. This new model, termed a Markov-Switching model, is presented in extensive detail. Parameter Estimation methods, Simulation Methods and Option Pricing Theory are explored. Summary algorithms are presented so that this dissertation may be used as a good reference guide for those wishing to apply Markov-Switching Models. The model is tested by fitting the model on South African data and using the discussed option theory to create various implied volatility surfaces. The surfaces produced appear to obey some of the empirical observations and theoretical ideas around expected implied volatility surfaces, indicating that the Markov-Switching model has some value for option pricing.

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

1. Introduction

1.1. Asset Prices and Stock Markets

It does not take academic astuteness to realize that stock price movements have a random nature to them. The stock market's nature is mostly random and unpredictable. The prices of publicly listed assets are determined by buy and sell instructions which arrive at random intervals and in random quantities. One empirical observation is that in the long-run, there is often a trend to the prices. Intuition and data observation may predispose us into observing that the price is more likely to go up in the long-run than go down.

We also observe that stock prices will never go below zero. Financial practitioners need a way to handle this uncertainty. A predictive model is unlikely to be accurate but an ability to capture the inherent volatility and essentially "risk"[1] is important for creation of other financial instruments which derive their value from a volatile asset (known as derivatives[2]). Stochastic models of stock prices allow one to speak of the probability that a stock price will be a certain value, breach a certain boundary or fluctuate within a specific band (amongst many other uses).

To this end, a very common model has been used for stock prices, that of Geometric Brownian Motion (GBM). GBM has only 3 parameters: the "drift" parameter, the volatility parameter and the initial or current observed stock price. It has some unique features which (for the most part) obey what we see empirically in the stock market. The key feature is that the log returns are assumed to be normally distributed with a positive mean. This implies that stock prices are lognormally distributed. Another very important consequence is that log returns are independent in non-overlapping time periods. This implies that knowledge of past returns or stock prices tells us nothing about the logreturn in a future period[3].

1.2. Emergence of Option Pricing Formulae

A groundbreaking paper by Black & Scholes (1973) provided a closed-form call and put option pricing formulae for options when the underlying asset follows GBM[4]. This formula is widely known as the Black-Scholes formula. An important result of this is that it is a purely arbitrage-free pricing model. In other words, under the assumption of arbitrage free markets, it emerged that some of our own opinions of what the most likely stock price movements are[5] have no bearing on the ultimate option price. A consequence of this is an almost totally

[1] Here I use the term "risk" to mean a measure of deviation from the expected in whichever direction.
[2] Derivatives in this context is not in the mathematical sense.
[3] In financial literature, this concerns the Efficient Market Hypothesis discussed Section 1.3.
[4] Discussed more in the Section 2
[5] I.e. The actual drift has no bearing on the price, only the volatility.

objective option pricing formulae. The only parameter that is open to subjectivity is the volatility parameter. In the Black-Scholes formula, this parameter is assumed constant[6].

1.3. Information-driven stock price thinking

Stock market prices, just like prices in most goods markets, are determined by supply and demand. Market participants wishing to sell stock, offer a price at which they are prepared to deal called the offer or ask price. Market participants wishing to buy stock, specify a price at which they are prepared to deal called the bid price. The mean of these two prices is called the mid-price.

I address stocks in this section primarily as shares in an equity market but the ideas regarding information arrival can be extended to other instruments[7]. Stock prices should theoretically (and to a large extent do practically) change due to the arrival of new information. The stock price should represent the current fair or present value of all future cash flows[8]. Stock prices are thus, in a sense, a representation of the *expected* future prospects of the companies linked to the stocks in question. If for example, a company brings out its annual financial statements reporting a net profit that is much lower than originally thought by the average market participant, then the stock price is likely to decline in price. Other examples could include a competitor launching a cutting edge product or the announcement of major staff layoffs.

Stock price graphs show that prices are highly volatile and erratic we can infer that information must be arriving rapidly and of course information does not arrive uniformly to every market participant. This creates erratic buy and sell instructions from market participants and contributes to volatility. Any stochastic stock model should then consider a theory of information and stock prices. The most famous of these is the efficient market hypothesis (or EMH).

The EMH introduced by Fama (1970) has 3 forms which are subsets of each other:

- Weak Form EMH

Loosely speaking, this states that past-period information is not captured in the current stock price. In other words, one cannot make "excess returns"[9] based purely on past-information. The implication of this is that being a chartist or technical analyst[10] has no merit in creating excess returns. In mathematical terms, this means that a stochastic stock model of the next period should not depend on the any functions of the stock price in prior periods.

---

[6] The Black-Scholes formula can be easily adjusted if the volatility parameter is a deterministic function of time.
[7] Bonds, derivatives, portfolios, exchange traded funds etc.
[8] In a equity scenario, this would be dividend payments. For bonds, this could represent future coupons
[9] Excess returns in this context means returns in excess of the returns on a market portfolio.
[10] These are analysts who examine past stock movements and look for patterns and create models to predict future movement based on past data.

In a market that is only weak-form efficient, there is scope for fundamental[11] analysts to generate excess returns.

- Semi-Strong Form EMH

Semi-Strong Form EMH states that all public information is captured in the stock price. The market has perfect knowledge with regards to public information. In a Semi-strong form efficient market, both fundamentalists and chartists are unable to generate excess returns. It is only in private information ("soon-to-be-public" information) that excess returns can be made. By definition, this means that all past-period information is also captured in the stock price so that the semi-strong form is merely a more narrowly defined definition of the weak form EMH.

- Strong Form EMH

The strong form EMH states that all information, public and private, is captured in the stock price.

Some studies (Magnusson & Wydick (2002), Simons & Laryea (2004)) have found that we can likely accept that the Weak Form EMH holds in the long run. This is tested by assessing the presence of serial correlation in stock returns. There do appear to be some moments throughout history where there are short periods where autocorrelation is significant. In most of the African markets (besides South Africa), the markets have shown to have exhibited some weak-form inefficiencies (Mlambo & Biekpe, 2007). As discussed above, the Semi-strong form appears to hold to some degree but there is definitely information asymmetry in the market, meaning that the price cannot always accurately reflect all available information. This is because participants act on new information at different times because the information arrives to them at different times (or in different forms). This paves the way for chartists and technical analysts to glean some profits.

The presence of serial correlation brings with it a whole new concept of thinking, that of irrational and inefficient markets.

1.4. Behavioral finance

Behavioural finance considers the behaviour of irrational market participants. There are certain behavioural tendencies or traits. Readers are referred to Shleifer (2000) for discussions on behavioural tendencies.

The result of an irrational market is that we tend to see over-reactions and under-reactions to the arrival of new information. For example, when an unexpected piece of information arrives

[11] Fundamental analysts are those that look at the company's fundamentals information (such as earnings rations, dividend yields etc.) to determine what they believe is a company's intrinsic value. Fundamentalists then buy/sell a stock that is cheaper/dearer than the intrinsic value in the expectation that in the long-run actual prices return to their intrinsic value.

to market participants uniformly (e.g. when annual financial statements are released or dividends are declared) the market tends to exhibit some short term correlation (Barberis, Shleifer, & Vishny, 1998). There tends to be an under-reaction to single-event news such as earnings announcements but an over-reaction of stock prices to a series of good and bad news (Barberis, Shleifer, & Vishny, 1998).

Another reason why this could occur is due to the presence of chartists in the market that tend to actually create short term correlation. In other words there is market segregation between fundamentalists and chartists and the proportion of each tends to change over time. This is an interesting solution for the presence of short-term autocorrelation and is explored by Vigfusson (1997). Vigfusson employs a Hidden Markov Model which governs the change in the proportion of fundamentalists and chartists in the market. Stock models for each type of investor can be anything but a fundamentalist model must possess the Markov Property[12] and that a chartist model must exhibit a degree of auto-correlation.

Single stock analysis thus becomes increasingly complex. Theoretically, short-term autocorrelation should be less persistent in stock indices and well-diversified portfolios because over and under reactions should largely cancel themselves out so we momentarily turn our attention to stock indices – namely the overall stock index like the JSE ALSI[13] or the TOP40. Finance theory breaks down risk or volatility into two components: firm specific risk and market risk[14]. If we create a notional portfolio of individual stocks (much like an index essentially operates) then the firm-specific risk is of the overall portfolio or index begins to diminish. This is because stocks within the portfolio are not perfectly correlated.

A perfectly diversified portfolio should have negligible firm specific risk and it is only in the market risk where the portfolio returns get their variability. Market risk arises because there are factors that affect all stocks uniformly – Information that arises that is not directly linked to any firm or company in the portfolio/index. This primarily arises from macroeconomic changes. The main example of this is the business cycle[15], where economies go in and out of recession and boom periods. Sometimes these shifts between these economic "regimes" are subtle and sometimes they are extreme (for example the global market crash of 2008). It is the study of how to accommodate these regimes into a stock price model that is of interest in this dissertation.

[12] The Markov Property states the probability of any future value of the process is not dependant on any previous values of the process (i.e. only the current value).
[13] JSE ALSI: Johannesburg Stock Exchange All-Share Index
[14] Also known as unsystematic risk and systematic risk respectively.
[15] Other examples of specific macroeconomic changes include raising of company tax rates, downgrading of government credit ratings, lack of foreign investor confidence and unexpected changes in the repo rate etc.

1.5. Bull Markets, Bear markets and Crashes

A bull-market is a buyer's market. Prices rise at a higher rate than average and volatility tends to be low. Bull-markets are often associated with periods of economic boom periods where the economy is growing rapidly. A typical example is the bull-market from 2003-2007 in South Africa. This cannot last forever and the economy is destined to turn the corner as inflation begins to rise as a by-product of economic growth. This leads into what is the opposite of a bear market – a bull market. This market exhibits slower-than-average growth and excess volatility.

In extreme cases, a bull market may transition into a bear market instantly via a market crash. Prices appear to be driven artificially[16] too low. On the other hand, stock market bubbles occur, for example the tech stock bubble of 1998 where prices are driven artificially high. These market states are loosely correlated to the business cycle,

Accommodation of a discrete number of economic regimes in a stock model can be done by a Market Switching Model (MSM) which is discussed in great detail in Section 3. Here we have that the parameters of stock models switch at random intervals to different values, which are intended to mimic these economic regime changes.

1.6. Aims and Methodology

There are two aims to this dissertation: One is to provide an overall practitioner's toolkit for MSMs. While some of the literature describes the steps separately, few outline the full process from model selection, estimation and option pricing. We also briefly develop a new criterion for estimating the number of states.

The second aim is to test the applicability of MSMs in the South African equity market. We briefly compare the standard Geometric Brownian motion model to MSMs in the SA environment. The resultant model is then used to create option prices and resultant implied volatility surfaces. If the model produces implied volatility surfaces that could reasonably be expected, then it is plausible that the model captures some of the inherent pricing factors that go into option pricing in the South African equity market.

We find that we are able to specify a two-state model for the TOP40 data which is a better fit than a standard Geometric Brownian motion model according to our criterion. This indicates that an MSM model explains stock price movements more accurately than the standard GBM model. Furthermore, we find that under different option pricing theories, the implied volatility surfaces produced are plausible and in line with some general implied volatility observations.

[16] Here we mean that prices are much lower than their fundamental values.

The successful results in this dissertation prompts further research into this model and opens the door for further model enhancements and research into alternative studies such as portfolio selection for example.

In the preceding Sections we lay out some empirical observations and theory as to why a model that allows for regime changes should be considered.

The dissertation begins with a brief background to the Black-Scholes model basic option pricing theory. This importantly lays out the foundation of the basic option pricing theory regarding *complete markets, replicating portfolios and risk neutral valuations.* It is a necessary basic prerequisite theory for the option theory we present under the MSM model.

The Heston model of stochastic volatility is also briefly touched on in the following Section to present an example of what type of implied volatility surfaces other stochastic volatility models can produce. The Heston model is critiqued and we will see in a later section that some of its shortcomings can be corrected via use of the MSM model.

Section 3 defines the model and some of the fundamentals behind MSMs. This section outlines the important definitions and framework upon which the model relies on.

Sections 3-6 address the issue of parameter estimation given empirical data. This Section is crucial to laying down the tools required to successfully fit a model to empirical data. Detailed algorithms and routines are supplied for ease of use.

Section 7 and 8 deal with the theory required to create an option pricing formula based on an MSM. This section expands on some of the core ideas discussed in Section 2 in the context of *incomplete markets*. We will see that this is a key difference between the Black-Scholes model and the MSM model.

Section 9 details the theory behind the creation of option pricing formula necessary to price options given our model. We will find that most of the prices will require some form of Monte-Carlo Simulation. Monte-Carlo simulation specific to the MSM model are addressed in Section 10.

After possessing the necessary tools (core theory, an estimated model and option pricing theory and techniques), Section 11 looks at the South African stock market and asses the suitability of MSM to the TOP40 index market using the themes developed throughout this dissertation.

The dissertation then concludes with a summary of the main results and discusses avenues for further research and improvements.

2. Overview of Core Theory and other models

2.1. Risk Neutral valuation and the Black-Scholes Market

The Black-Scholes option pricing formula revolutionised the pricing of options. Prior to the market crash of 1987, option prices tended to behave according to this formula. The Black-Scholes market assumptions give rise to a market that is complete and the idea of risk-neutral valuation.

We need to briefly go over the theory surrounding the Black-Scholes (BS) market and risk neutral valuation in order to use similar principles for option valuation under our proposed model. This section assumes the basic knowledge of Ito calculus and martingales.

Assumptions of the BS market

There exists a deterministic Bank account $B(t) = e^{rt}$, where r is the positive constant risk-free continuous rate of return. Alternatively it can be represented as:

$$dB_t = rB_t dt \tag{2.1}$$

There exists a stock $S(t)$ whose dynamics are governed by Geometric Brownian motion (GBM) with drift parameter μ and a volatility parameter σ. I.e. $dS_t = \mu S_t dt + \sigma S_t dW_s$ under the real-world measure which we will call measure $\mathbb{P}$.

If the value of W_t is known at time t, then so is S_t.

The market is frictionless: there are no transaction costs and short-selling is permitted

The market is infinitely divisible: any amount of the stock and bond can be purchased or sold in the smallest possible time unit.

Girsanov's theorem and martingale creation

Theorem: Girsanov's Theorem

Let W be a Brownian Motion with respect to measure $\mathbb{P}$ and filtration $\Im_t$. Let $W^q = (W + \int_0^t q(s)\, ds)$ or equivalently $dW^q = dW + qdt$ and $d\mathbb{Q} = e^{-\frac{1}{2}\int_0^T q(s)^2 ds - \int_0^T q(s) dW(s)} = \partial(T)\, d\mathbb{P}$ such that $\mathbf{E}\left[\int_0^T \partial^4(\tau) q(\tau) d\tau\right] < \infty$.

W^q is a Brownian Motion with respect to $\mathbb{Q}$ and the original filtration $\Im_t$.

Girsanov's theorem essentially allows us to change the drift of the Brownian Motion such that it is still a Brownian motion under a different measure. This is a powerful tool to allow us to change Brownian motions into martingales (under a different measure) by adjusting the drift.

By applying some standard stochastic differential calculus (the quotient rule) we find that:

$$\frac{d\left(\frac{S}{B}\right)}{\frac{S}{B}} = (\mu - r)dt + \sigma dW$$

$$= \sigma d(W + \frac{\mu - r}{\sigma} dt) \quad (2.2)$$

We can now use Girsanov's theorem to adjust the drift such that the above function is a martingale under new measure $\mathbb{Q}$ by letting $dW^q = dW + \frac{\mu - r}{\sigma} dt$

Effectively this means that the discounted stock price is a $\mathbb{Q}$-martingale or that the stock price drifts at the risk-free rate. The stock dynamics under $\mathbb{Q}$ becomes:

$$\frac{dS}{S} = rdt + \sigma dW^q \quad (2.3)$$

The measure $\mathbb{Q}$ is called a risk-nuetral measure.

<u>Creating of a replicating, self-financing portfolio</u>

In theory, if we are able to set up a portfolio that is self-financing (i.e. we can rebalance the portfolio without any further cash injections or costs) and replicating (i.e. that it exactly matches the payoff of the option) then the portfolio and the option have identical payoffs and are "essentially" the same instrument.

If the one was more expensive than the other, we would short the expensive one and buy the cheaper one, creating a zero-cost guaranteed profit or "arbitrage". The concept of arbitrage is discussed in more detail in Section 7. It is reasonable to assume that the market doesn't allow arbitrage (and this is usually a standard assumption when pricing financial instruments). If an arbitrage opportunity existed, it would do so temporarily, as the action of buying and selling to exploit the opportunity will shift the prices such that the opportunity vanishes.

Therefore under a no-arbitrage principle, the portfolio and the option price must have the same price. If we can determine the price of the portfolio, we can determine the price of the option.

Assume a portfolio $V(t)$ exists such that:

$$V(t) = \psi(t)B(t) + \phi(t)S(t) \tag{2.4}$$

such that the pair $(\phi(t)B(t), \psi(t)S(t))$ is previsible. A previsible process is a process that only depends on information available up to the current time and not any future information.

The self-financing condition implies that:

$$dV(t) = \psi(t)dB(t) + \phi(t)dS(t) \tag{2.5}$$

It is also evident that once $\phi(t)$ is determined, $\psi(t)$ is easily found by the proceeds from the sale or purchase of stock so that the portfolio maintains the self financing condition. In other words, we can choose $\phi(t)$ freely but $\psi(t)$ is fixed because we enforce the self-financing condition. We wish to have that $V(T) = X$, where X is the claim (i.e. $\max(S_T - K, 0)$ for a vanilla European call) so that the portfolio is replicating.

Thus, it remains to find a $\phi(t)$ that satisfies the above two conditions. For this we need the *martingale representation theorem.*

Firstly, we state an alternative specification of V_t:

$$V_t = B_t \mathbf{E}_{\mathbb{Q}}[\frac{X}{B_T}|\Im_t] \tag{2.6}$$

It is clear by the tower rule that V_t/B_t is a $\mathbb{Q}$-martingale and additionally that $V_T = X$ so that the above specification for V_t is replicating. We need to reconcile the two specifications.

<u>The Martingale representation theorem</u>

Theorem: Martingale Representation Theorem

Let M_t be a square-integrable martingale with respect to the BM filtration $\Im_t^W$. Then there exists an adapted process $\xi(t,\omega)$ such that we have:

$$M_t = M_0 + \int_0^t \xi(s)dW_s \tag{2.7}$$

Equivalently, we have $dM = \xi dW$.

We apply this theorem to V_t/B_t, a $\mathbb{Q}$-martingale to get:

$$d\left(\frac{V}{B}\right) = \xi dW^q \tag{2.8}$$

And further since we know that $d\left(\frac{S}{B}\right) = \sigma\left(\frac{S}{B}\right) dW^q$ from before, so we can write the above equation as:

$$d\left(\frac{V}{B}\right) = \frac{\xi}{\sigma}\frac{d\left({}^{S}/_{B}\right)}{{}^{S}/_{B}} = \frac{\xi}{\sigma}\left(\frac{dS}{S} - \frac{dB}{B}\right) \tag{2.9}$$

Again by the quotient rule, we have the following expression for $d\left(\frac{V}{B}\right)$:

$$\frac{d\left(\frac{V}{B}\right)}{\frac{V}{B}} = \frac{dV}{V} + \frac{d\left(\frac{1}{B}\right)}{\frac{1}{B}} - \frac{dV}{V}.\frac{dB}{B}$$

$$\Rightarrow d\left(\frac{V}{B}\right) = \frac{V}{B}\left(\frac{dV}{V} - \frac{dB}{B}\right) \tag{2.10}$$

Equating the right hand sides of the equations we get:

$$\frac{V}{B}\left(\frac{dV}{V} - \frac{dB}{B}\right) = \frac{\xi}{\sigma}\left(\frac{dS}{S} - \frac{dB}{B}\right)$$

$$\frac{dV}{B} - \frac{VdB}{B^2} = \frac{\xi BdS}{\sigma S} - \frac{\xi dB}{\sigma B}$$

$$dV = \frac{VdB}{B} - \frac{\xi BdS}{\sigma S} + \frac{\xi dB}{\sigma}$$

$$dV = \left(\frac{V}{B} - \frac{\xi}{\sigma}\right)dB + \frac{\xi B}{\sigma S}dS \tag{2.11}$$

If we set $\phi = \frac{V}{B} - \frac{\xi}{\sigma}$ and $\psi = \frac{\xi B}{\sigma S}$, then we have the self-financing condition. Importantly, it doesn't matter what ξ is (and hence ϕ and ψ), only that it exists which is true via the martingale representation theorem.

We have now effectively reconciled the two versions of $V(t)$ to show that:

$$\phi(t)S(t) + \psi(t)B(t) = B(t) = \mathbf{E}_{\mathbb{Q}}\left[\frac{X}{B(t)}\middle|\mathfrak{I}_t\right] \ \forall t \leq T \tag{2.12}$$

Furthermore, we can then deduce by a no-arbitrage argument that for a call option:

$$O_t = B(t)\mathbf{E}_{\mathbb{Q}}\left[\frac{(S_T - K)^+}{B(t)}\middle|\mathfrak{I}_t\right]$$

$$= \mathbf{E}_{\mathbb{Q}}\left[e^{-(T-t)}(S_0 e^{\left(r-\frac{1}{2}\sigma^2\right)t+\sigma\sqrt{t}W_{(T-t)}}\middle|\mathfrak{I}_t\right] \tag{2.13}$$

We can apply the definition of an expectation to the probability density function to find the answer. We'll omit the details of the integration process, but the result becomes:

$$O_t = S_0 N(d_1) - Ke^{-r(T-t)} N(d_2) \tag{2.14}$$

Where $N(x)$ represents the $\Pr[Z \leq x]$ if Z is a standard normal random variable. Furthermore:

$$d_1 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r + \frac{1}{2}\sigma\right)(T-t)}{\sigma\sqrt{T-t}} \tag{2.15}$$

$$d_2 = d_1 - \sigma\sqrt{T-t} \tag{2.16}$$

There are many forms of the risk neutral argument above, however most of them follow a similar vein. The Girsanov theorem is used to change the discounted stock price to a martingale under a particular measure. Then by imposing the self-financing condition and creating a portfolio of "pricable" financial instruments, replication is shown to be possible via the martingale representation theorem.

A key implication here is the uniqueness of the measure that makes the discounted stock a martingale, for example should the stock depend on two sources of Brownian motion then we may have more than one measure that enables it to be a martingale. This touches on the concept of a *complete* market which is discussed in significant more detail Section 7.

2.2. Implications of the BS Pricing Formula and Implied Volatility

The BS formula depends on some unrealistic assumptions such as the lack of transaction costs, infinite divisibility and unlimited short-selling. Even if some of these are true, instantaneous rebalancing is impractical and can only realistically be done at small intervals.

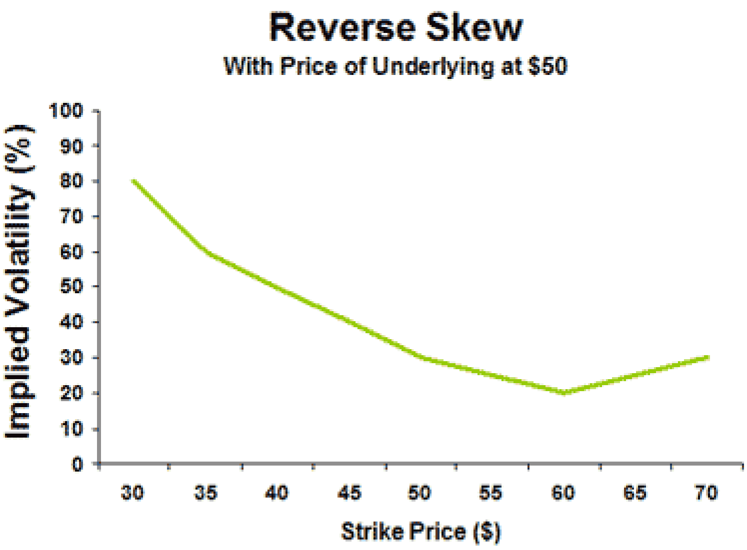
To this end, not all investors in the market can create the same replicating portfolio in order to take account of potential option mispricings and hence there may actually be an arbitrage free "band" of permissible prices within which the option price may fluctuate. The mispricing difference between the theoretical option price and the actual option price may be too small to take advantage of after the limitations on trading are imposed. However, this could be considered a small disadvantage because under these arrangements, we expect on average the BS price to be correct.

Why then do option prices not behave exactly as the BS formula expects them to behave? If option prices behaved like the BS formula, then if we calculated the implied volatility (knowing all other parameters, including the actual option price) then we should find that the volatilities are the same regardless of term or strike price. In fact, this was roughly true up until 1987. However, subsequent to this period, the implied volatility surface[17] deviated from the expected flat nature.

At-the-money options tend to be the most-liquid and will exhibit the lowest prices while those in-the-money or out-the-money have less liquidity and hence could demand a higher price. Higher prices imply, higher volatilities and vice versa so a plot of implied volatility vs. the strike price would indicate a “U-shape” or volatility smile. Varying liquidity can also cause different bid-ask spreads which affect the slope of the curve.

On the other hand, (which is most common in equity options), we can find a “volatility skew” or otherwise known as a “reverse skew”. This happens when there is a “half-smile” such as the one below from a finance website[18]:

Figure 2.1: Typical reverse equity volatility skew



Here we see that implied volatilities are very high for in-the-money calls and out-the-money puts and that the lowest volatility point is not at-the-money. One explanation of this is that investors are worried about market crashes and buy puts for protection, driving up the price.

The other argument is that in-the-money calls are alternative to straight stock purchases since they provide leverage. This reverse skew was first present after the 1987 crash.

Another variant is the “forward skew” which is a mirror image of the reverse skew but this is more present in the commodities market. A further variant is the “volatility frown” but this is more a name for a shape that is not seen often in reality.

On the other hand, perhaps the underlying stock model of Geometric Brownian motion is incorrect. This is the most plausible explanation for the non-flat implied volatility curves and surfaces.

[17] The is a three dimensional implied volatility surface plotted over varying strikes and terms to maturity. The implied volatility is determined using the BS formula.
[18] http://www.theoptionsguide.com/volatility-smile.aspx

A key characteristic of the standard GBM model is that for every fixed length time period going forward, the log return has the same probability of being a specific value. Empirically, this is not what is observed. For example, in the recent financial market crash[19] volatility was high and average returns were mostly negative. In this period, returns appeared more likely to be negative than positive. One could argue that of course the standard fixed parameter GBM model does account for this but the probability of successive negative returns is very low, too low to simply chalk down the event to an extra-ordinary low probability event. Furthermore, short term volatility in this period was also excessively high[20] further giving credence that it was not just an extreme event. Indeed it seems that if the GBM model does hold true, then the parameters themselves may also indeed be stochastic giving us a second order type of model.

A model of asset prices should replicate some stylized facts of asset prices and thereafter deviation from the given smile could be discussed. It is this choice of stock model that is the main theme of this thesis.

2.3. Stylised asset price facts and the Heston Volatility Model

The below facts is a subset of those discussed in Cont (2001):

(1) *Absence of auto-correlations*. Linear autocorrelations are often statistically insignificant except on a very small scale. Effectively, this implies we need a stock model with the Markov property. Geometric Brownian motion is one such model.

(2) *Heavy Tails of unconditional stock returns.* Stock returns tend to exhibit excess kurtosis. Fixed-parameter GBM assumed all log-returns are normal and so the standard model fails to capture this stylised fact.

(3) *Volatility Clustering.* Volatility measures display autocorrelation over several days. The standard GBM model assumes a fixed volatility parameter and so clustering is not possible.

(4) *Intermittency:* Returns display a high degree of variability at any time scale where there tends to be irregular bursts of volatility. Under the GBM model, volatility should be constant over time.

(5) *The leverage effect.* Volatility measures tend to be negatively correlated with the returns of that asset. In the standard GBM model, since volatility is constant, the volatility measure is independent of the asset return. In other words, as the underlying price becomes low, so does the volatility.

[19] Roughly occurring between 2008 and 2009 worldwide.
[20] In other words, it was higher than it had been for an extended period prior to the crash.

It is quite clear that standard GBM model fails to capture facts 2-5. One particular conclusion that can be drawn from above is that volatility appears to be non-constant. Stochastic volatility is a major theme in the literature and the most common model to try and mimic this is a volatility model by Steven Heston (Heston, 1993).

The Heston model retains the appealing functional form of the GBM model but allows the volatility parameter to vary stochastically. In fact the volatility process is a CIR diffusion process[21]. The following specification is used:

$$dS_t = \mu S_t dt + \sqrt{v} S_t dW_t^1 \tag{2.17}$$

The volatility process is defined as:

$$dv_t = \kappa(\theta - v_t)dt + \sigma\sqrt{v_t} dW_t^2 \tag{2.18}$$

Here, κ is known as the mean-reversion "speed", θ the long-term volatility and σ the volatility of volatility. Furthermore, the Heston model allows for correlation between the two Brownian motion components such that $Corr(W_t^1, W_t^2) = \rho$.

The latter flexibility allows one to model the leverage effect by setting ρ as a negative number to allow for full or partial negative correlation. The Heston model of course also brings the solution to (3) since the volatility process by definition exhibits auto-correlation.

As before, using Risk-Neutral pricing, we find that we can let the stock price drift at the risk-free rate and that option prices are the expectation of the discounted payoff under the risk neutral measure.

We won't expand on the details much further but it may be helpful to try get an understanding of some of the volatility smiles we may experience under the Heston Model. The parameter values chosen below were chosen to be similar to the empirical data explored in Section 7 for comparison purposes. The following parameters were chosen (time is measured in days so all parameters are daily parameters):

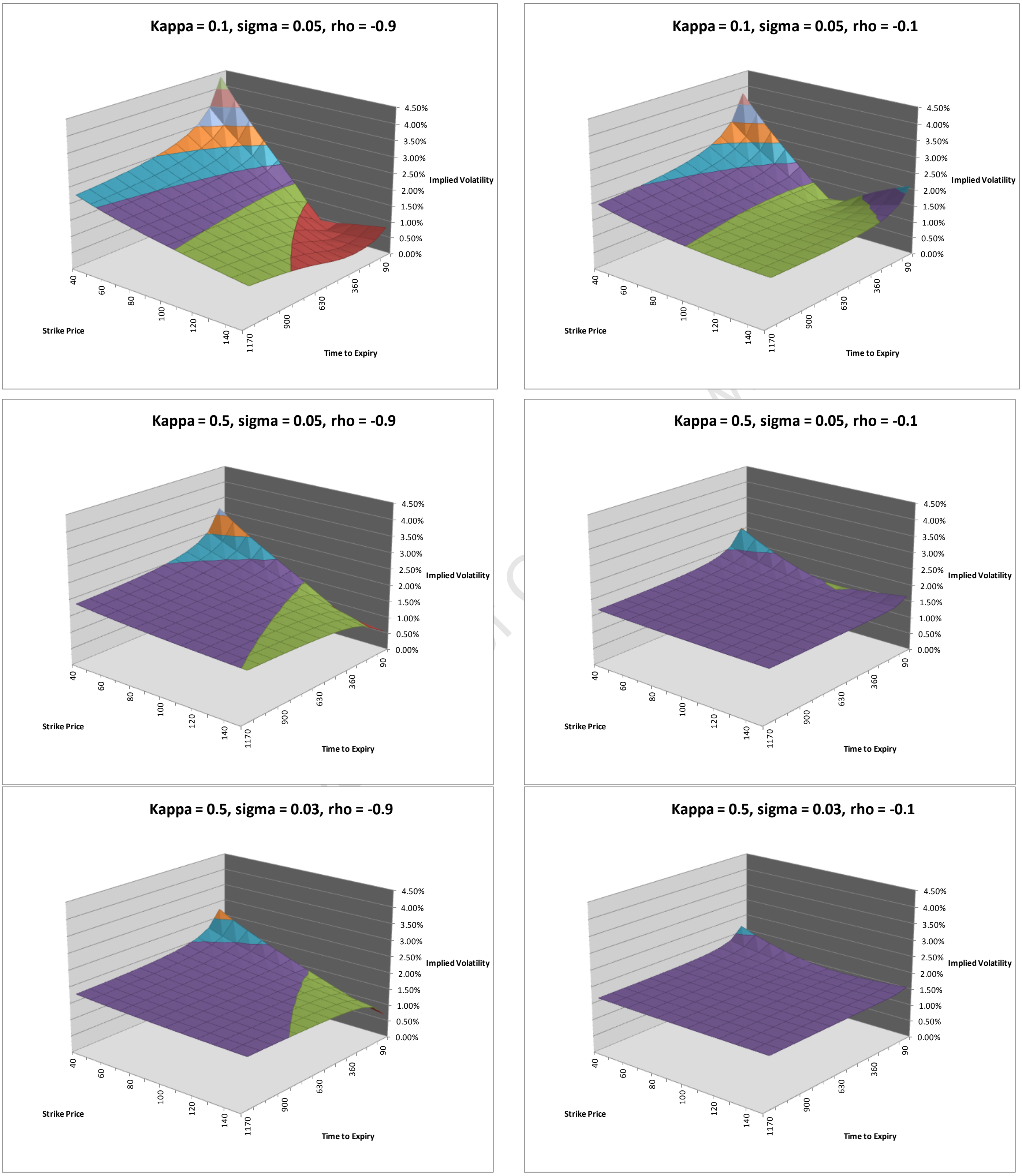
S_0	100
r	0.0280%
θ	0.0243%
v_0	0.0243%

The long-term volatility is chosen to be such that it matches the underlying sample standard deviation. Three further unknowns exists, that of ρ, σ and κ.

It is noted that a low κ and high σ represents a "wilder" volatility process, less restrained by the mean-reversion component and a higher random component to the volatility. The converse is a higher κ and a lower σ which represents a more "tamer" or stable volatility process. The following figures are shown:

[21] The process is mean-reverting with scalable volatility, in that the smaller the value, the smaller the volatility and vice versa

Figure 2.1: Heston Volatility Surfaces



The graphs in the top row represent the “wild” volatility; the final row represents “tamer” volatility and the middle row a scenario somewhere in the middle. It is clear that modifying these parameters can change the shape of the surface dramatically. However there does tend to be some noticeable themes. Importantly, “wilder” volatility bring more extreme curvature, and higher negative correlation tends to exhibit skews more than smiles.

Furthermore, it appears the shorter dated options tend to exhibit more extreme curvature than the long-dated options. Implied volatilities also tend to be lower for long-dated options.

This is perhaps slightly counter-intuitive; long-dated options are less liquid than shorter dated options. Lower liquidity instruments tend to fetch higher prices than lower liquidity instruments (all else equal) and therefore we should expect to find higher implied volatilities. There may also be more uncertainty with long-dated options and this could attract a risk premium (therefore a higher price and a higher volatility).

The Heston model won’t be mentioned much further but suffice to say that it does seem to miss some dynamics for longer-dated implied volatilities. With the preliminaries out the way we now concentrate on the proposed model in this thesis, the Markov-Switching model and discuss some preliminaries on stochastic processes.

3. Processes with Markov Switching

3.1. Stochastic and Markov Processes

A stochastic process is in its simplest form, a sequence of random variables. There may be infinitely many random variables in the sequence. The sequence may further be countable or uncountable. In most cases, the sequence is indexed by time, t, so that the order of the sequence follows nicely with the passage of time and allows us to use the terms "past values", "current values" and "future values" without ambiguity.

Most definitions in this Section come from Guo R. (2006) and have been adjusted slightly to fit our purpose.

Definition: Stochastic Process

A stochastic process is an ordered collection of random variables.

For example, a process can be expressed as $\{Y_t\}_{t\geq 0}$ which is a collection of values Y_t for some or all values of $t \geq 0$.

From here on out, we will assume that all stochastic processes dealt with here are indexed by time. An important subdivision of stochastic processes refers to the "countability" of the random variables in the stochastic process. Should there be countable[22] time periods between successive values of the stochastic process, then the process is known as a discrete time stochastic process[23]. As an example, the time index t be permitted to only take on integer values such that a process could be a collection of the variables Y_0, Y_1, Y_2 ... and so on such that $t \in \mathbb{N}^+$.

Definition: Discrete time stochastic process indexed by time

A discrete time process $\{Y_t\}_{t>0} = \{Y_t\ ;\ t = 0, \Delta t, 2\Delta t, \ldots\}$ is a countable collection of random variables indexed by countable periods of time.

In other cases, it is possible that $t \in \mathbb{R}^+$. Here, successive values are not countable as the process is defined for all values in an arbitrarily small time period. The process can take values at any point in time.

[22] Either finitely countable or infinitely countable
[23] Importantly, the lengths of these time periods are all the same.

Definition: Continuous time stochastic process

A continuous time process $\{Y_t\}_{t>0} = \{Y_t\,; 0 \le t < \infty\}$ *is an uncountable collection of random variables indexed by non-negative real numbers.*

Both types of processes may have a defined length (maturity) or run ad infinitum.

The range of possible values a stochastic process can take is often restricted to a set range. We can define an environment of “states” which are the possible values the process can take. There may be a finite or infinite number of “states” the process can be in.

Many processes have what is known as the Markov property, which loosely means that the only dependency that a stochastic random variable has to other random variables in the process is through the previous random variable. A stochastic process possessing the Markov property is known as a Markov Process.

Definition: Markov Property (General)

A stochastic process has the Markov property if the conditional probability distribution of future states depends only on the current state (Dodge, 2003*).*

Definition: Markov Chain

A Markov Chain is a discrete-time stochastic process with the Markov Property such that

$$\Pr[X_{t+\Delta t} = x \,|X_t, X_{t-\Delta t}, \cdots, X_0] = \Pr[X_{t+\Delta t} = x|X_t] \text{ for all } t \ge 0 \qquad (3.1)$$

The term *Markov Chain* specifically refers to discrete-time processes but the more general term Markov Process can be used to describe both discrete and continuous time. It will be clear from the context whether the statements using this term are specific for continuous time only.

Definition: Markov Process (Continuous time)

A Continuous time Markov Process is a stochastic process with the Markov Property such that:

$$\Pr[X_t = x \,|\Im_s] = \Pr[X_t = x|\sigma(X_s)] \text{ for all } t \ge s \ge 0 \qquad (3.2)$$

$\{\Im_t\}_{t\ge0}$ *is the natural filtration of stochastic process* $\{X_t\}_{t\ge0}$ *and the* $\sigma(*)$ *operator refers to the sigma algebra generated by the argument.*

An alternative representation is:

$$\Pr[X_t = x | \mathfrak{I}_t] = \Pr(X_t = x | X_{t^-}) \tag{3.3}$$

t^- refers *to* $\lim_{dt \downarrow 0}(t - dt)$*.* It is the instantaneous moment before time t.

As an example, we could consider a process that lives in a three state world, so that the process can only take three values. The Markov property is brought in such that the probability of moving to state i (and then the process taking on a specific value) is only dependant on the current state[24]. Processes such as these are very common in the current world. For example, the game of snakes and ladders can be defined in terms of a Markov Chain by letting each square be a state. Given the current square/state or position of the playing piece, there are only a limited number of states that can be reached depending on the dice throw. These "allowable" next states change as we move through the board. We can easily calculate the probabilities of getting to each square conditional on our current state based on a randomised dice. Clearly the probability of landing on a certain square depends solely on our current position and not on any previous positions – the Markov property.

We could expand our snakes and ladders game to allow for access to sets of intervals on the real number line. For example, depending on our current position, we could perhaps then move to any number in a certain finite interval on the real number line. If we were in a different position, we could possibly move to a number in a different finite interval. The process is clearly still Markov but the range of allowable states has changed.

Definition: Discrete and continuous state processes

Let $\{X_t\}_{t \geq 0}$ *be a stochastic process such that* $X_t \in L$ *for all* $t \geq 0$*. If* L *contains countably many elements, then the process is known as a discrete-state process. If* L *contains uncountably many elements, then the process is known as a continuous-state process.*

A simple example of a continuous state process would be a process that takes the value of the previous value of the process plus a random draw from a standard normal distribution. An example of this is our stock price process which has the state space $(0, \infty)$.

Our stock price at time t in the simplest case is defined as $S_t = f(S_{t^-}, Z_t)$ where Z_t is a random draw from *i.i.d* random variables of known distribution. The process is Markov because the following state only depends on the current state and no other previous values. Depending on how the time index is defined, the process may be a continuous time process or a discrete time process[25]. Put in another way, the discrete or continuous nature of the time index is separate from the discrete or continuous nature of the states.

[24] There may be other dependencies on outside random variables but the only dependency on other variables in the process is through the previous value (i.e. "state" that the process was in)

[25] In discrete time we may have $f(S_{t^-}, Z_t) \equiv f(S_{t-\Delta t}, Z_t)$

We now turn to the concept of time-homogeneity, an important idea in this thesis.

Definition: Time homogeneous Markov Chain

If $\{X_t\}_{t\geq 0}$ is a time homogeneous Markov Chain then

$$\begin{cases} \Pr[X_t = i|X_{t-\Delta t} = k] = \Pr[X_s = i|X_{s-\Delta t} = k] & for\ t > 0 \\ X_0 & for\ t = 0 \end{cases} \tag{3.4}$$

for any time s and t, time period Δt and state i and k

A time inhomogeneous Markov Chain is a Markov Chain which does not meet the above definition. The continuous time version of the above is a simple extension of the above and we will omit the definition.

Another way of explaining the feature of time homogeneity is to say that the probability of state movements in a particular time period depend only on the length of the time period and not on the position in time. Time inhomogeneous chains are not considered in great detail in this thesis. The reasons are discussed in Section 12.

A Markov Chain is fully defined by the knowledge of the initial state (or initial state distribution) and knowledge of the “one-step transition probabilities” (proof below).

Firstly define $p_{ij}(t)$ as the probability that the Markov Chain transitions from state i to state j in a period of length t. The time argument is commonly dropped when referring to the “one-step” time period.

Proof: Markov Chain specification

The joint probability distribution is:

$$\Pr[X_0 = x_0, X_{\Delta t} = x_1, X_{2\Delta t} = x_2, \ldots]$$

By using the law of total probability and knowledge of the initial state, we can write it as:

$$= \Pr[X_{\Delta t} = x_1, X_{2\Delta t} = x_2, \ldots |X_0 = x_0] \Pr[X_0 = x_0]$$

$$= \Pr[X_{2\Delta t} = x_2, X_{3\Delta t} = x_3, \ldots |X_0 = x_0, X_{\Delta t} = x_1] \Pr[X_0 = x_0]\, p_{x_0 x_1}(\Delta t)$$

$$= \Pr[X_{2\Delta t} = x_2, X_{3\Delta t} = x_3, \ldots |X_{\Delta t} = x_1] \Pr[X_0 = x_0]\, p_{x_0 x_1}(\Delta t)$$

$$= \quad \vdots$$

$$= \quad \Pr[X_0 = x_0]\, p_{x_1 x_0}(\Delta t) p_{x_2 x_1}(\Delta t), \ldots \tag{3.5}$$

We can capture these transition probabilities in a "transition probability matrix" $\boldsymbol{P}(t) = \{p_{ij}(t)\}_{0 \le i,j \le n}$.

Definition: Transition probability matrix

A transition probability matrix $\boldsymbol{P}(t)$ associated with Markov Chain $\{X_t\}_{t \ge 0}$ is a square matrix with entries $p_{ij}(t)$ and the following properties:

(1) the sum of all the entries in each row is 1 (Stochastic Matrix)
(2) $\boldsymbol{P}(0) = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix
(3) If n is the number of possible states, then the matrix is of dimension $n \times n$
(4) $p_{ij}(u) = \Pr[X_t = j | X_{t-u} = i]$ for any u
(5) The Chapman-Kolmogorov equations holds:

$$\boldsymbol{P}(s)\boldsymbol{P}(t) = \boldsymbol{P}(s+t) \tag{3.6}$$

and

$$\boldsymbol{P}(t)\boldsymbol{P}(s) = \boldsymbol{P}(s+t) \tag{3.7}$$

We also drop the time argument to write $\boldsymbol{P} \equiv \boldsymbol{P}(\Delta t)$ with entries $p_{ij} = \Pr[X_t = j | X_{t-\Delta t} = i]$ for any t.

A continuous-time Markov processes can be seen as a case of a limiting discrete-time process. We can think of continuous time as if the time gap Δt tends to zero. The consider the instantaneous change in $\boldsymbol{P}(t)$ we can look at the differentiability of $\boldsymbol{P}(t)$, denoted $\boldsymbol{P}'(t)$.

Define a matrix $\boldsymbol{Q}$ to be the "rate" matrix or generator matrix representing the instantaneous transition probability change just after time zero. In other words:

$$\begin{aligned} \boldsymbol{Q} &= \lim_{\Delta t \downarrow 0} \frac{\boldsymbol{P}(\Delta t) - \boldsymbol{P}(0)}{\Delta t} \\ &= \lim_{\Delta t \downarrow 0} \frac{\boldsymbol{P}(\Delta t) - \boldsymbol{I}}{\Delta t} \end{aligned} \tag{3.8}$$

An interesting result now presents itself, which addresses the differentiability of $\boldsymbol{P}(t)$.

Proof: Forward equation

From the Chapman-Kolmogorov equations:

$$\boldsymbol{P}(t+\Delta t) = \boldsymbol{P}(t)\boldsymbol{P}(\Delta t)$$

$$\Leftrightarrow \quad \frac{\boldsymbol{P}(t+\Delta t) - \boldsymbol{P}(t)}{\Delta t} = \frac{\boldsymbol{P}(t)\boldsymbol{P}(\Delta t) - \boldsymbol{P}(t)}{\Delta t}$$

$$\Leftrightarrow \quad \frac{\boldsymbol{P}(t+\Delta t) - \boldsymbol{P}(t)}{\Delta t} = \boldsymbol{P}(t)\left(\frac{\boldsymbol{P}(\Delta t) - \boldsymbol{I}}{\Delta t}\right)$$

$$\Leftrightarrow \quad \lim_{\Delta t \downarrow 0} \frac{\boldsymbol{P}(t+\Delta t) - \boldsymbol{P}(t)}{\Delta t} = \boldsymbol{P}(t)\left(\lim_{\Delta t \downarrow 0} \frac{\boldsymbol{P}(\Delta t) - \boldsymbol{I}}{\Delta t}\right)$$

$$\Leftrightarrow \quad \boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{Q} \tag{3.9}$$

The "backward" equation $\boldsymbol{P}'(t) = \boldsymbol{Q}\boldsymbol{P}(t)$ follows an almost identical proof except starting with $\boldsymbol{P}(t+\ \Delta t) = \boldsymbol{P}(\Delta t)\boldsymbol{P}(t)$.

From the forward equation, we know (together with the condition that matrix $\boldsymbol{P}(t)$ is a stochastic matrix) from basic differential equations that the result is:

$$\boldsymbol{P}(t) = e^{\boldsymbol{Q}t} \tag{3.10}$$

Therefore an analogy to the discrete-time process, the continuous-time Markov Process is fully defined by the initial state and a generator matrix $\{q_{ij}(t)\}_{0 \leq i,j \leq n}$.

Definition: Generator matrix

A generator matrix $\boldsymbol{Q}$ associated with a continuous-time Markov Process is a square matrix with entries q_{ij} and satisfying the following properties:

(1) *The sum of all the entries in each row is 0*
(2) *If n is the number of possible states, then the matrix is of dimension $n \times n$*
(3) $q_{ii} = -\sum_{j \neq i} q_{ij}$
(4) $q_{ii} < 0$
(5) *For an infinitesimally small interval $[t, t+dt]$:*
    a. $\Pr[X_{t+dt} = j \mid X_t = i] = q_{ij}\, dt$
    b. $\Pr[X_{t+dt} = i \mid X_t = i] = (1 - \sum_{j \neq i} q_{ij})\, dt = (1 + q_{ii})dt$
(6) $\boldsymbol{P}(t)$ *is linked to* $\boldsymbol{Q}$ *via the following equations:*
    a. $\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{Q}$ – *the forward equation*
    b. $\boldsymbol{P}'(t) = \boldsymbol{Q}\boldsymbol{P}(t)$ – *the backward equation*

3.2. Hidden Markov Models

Often in reality there are stochastic processes governing the behaviour of other stochastic processes but only a subset of these stochastic processes are observable. If an observable stochastic process $\{Y_t\}_{t\geq 0}$ is governed by some other unobservable stochastic process, then the unobservable stochastic process is often called a latent process (Borsboom, Mellenberg, & Van Heeden, 2003).

One model that makes use of a latent process is a Hidden Markov Model (HMM). Here we have a stochastic process whose value depends on an unobservable Markov Process[26] $\{X_t\}_{t\geq 0}$. Generally and throughout this dissertation, this latent process will refer to a discrete-state process with finitely many states.

Definition: Hidden Markov Model (HMM)[27]

A Hidden Markov Model (HMM) is a triple $(\{Y_t\}_{t\geq 0}, \{X_t\}_{t\geq 0}, \boldsymbol{\eta})$ *or* $(\{Y_t\}_{t\geq 0}, \{X_t\}_{t\geq 0}, X_0)$ *- A system of stochastic processes comprising of one observable Markov Process,* $\{Y_t\}_{t\geq 0}$ *,one latent or unobservable "underlying" discrete-state Markov Process* $\{X_t\}_{t\geq 0}$ *and an initial state distribution* $\boldsymbol{\eta}$ *for* $\{X_t\}_{t\geq 0}$ *or the value of* X_0*.*

Additionally, the conditional values of Y_t *(conditional on the value of* X_t*) are independent of each other.*

In other words, the only connection between consecutive values of Y_t is through the value of X_t. In auto-regressive cases, Y_t (conditional on the value of X_t) will depend on all the prior observed values of Y_t.

There are three problems associated with Hidden Markov Models:

(1) Computing the probability of an output sequence of the observed process.

(2) Given the parameters of the model, find the (hidden) state sequence that is likely to have generated the observed sequence

(3) Given an output sequence, what are the most likely parameters that generated the output sequence?

Problem one can be solved quite efficiently using the *forward algorithm (*the complement of the *backwards algorithm)*, which is closely related to the *Viterbi Alogrithm,* which can be used to solve problem 2.

[26] Either continuous time or discrete time
[27] This is not the general definition of a Hidden Markov Model but we have tailored it for the purposes of this dissertation.

Problem 3 is the one that usually requires the most attention. We need an efficient maximum likelihood estimate (MLE) of the parameters. There is no exact maximum likelihood function that can be maximized to determine unique MLE's due to the unobserved nature of the latent process. One efficient method is the *Baum-Welch algorithm* which is a specific case of the *Expectation-Maximization* algorithm dealt with in Section 4.

3.3. Markov Switching Models

A variant of the general HMM model is a so-called Markov Switching Models (MSMs). Here the observable process is a stock-price process with stochastic parameters that are governed by a discrete-state Markov Process. Therefore, the stochastic parameter will take a finite set of values of the same number as the number of states in the Markov Process. They were first introduced by Hamilton (1989) under the auto-regressive process context.

Heston's model explored in Section 2 is one way to model stochastic volatility. We wish to study an alternative based on some other observations about volatility. As discussed in the introductory paragraph, the market and economy appears cyclical and tend to follow certain regimes. In other words, we have *regimes of volatility* which have been explored by Alexander & Lazar (2005). We are interested in MSMs of exponential family of distributions where we consider both mean and variance to be stochastic.

The stock price process S_t is permitted to take any real number (i.e. a continuous state process), while the underlying process X_t is only permitted to take a countably finite number of values (discrete-state Markov Process). Importantly though, both processes could still either be a discrete-time process or a continuous time process.

3.4. Gaussian Markov Switching Models

MSMs will be of particular interest to us in the context of the Normal (Gaussian) distribution. In our stock models, one common assumption is that stock returns are independent of each other and follow a normal distribution.

One possible solution (of which there are many) is to assume that the stock returns are modelled by a MSM. Here we would assume that each return comes from a normal distribution with mean μ_i and variance σ_i sourced from vectors $\overline{\boldsymbol{\mu}}$ and $\overline{\boldsymbol{\sigma}}$ respectively (both with dimension n. The mean and variance are stochastic and are driven by an underlying latent process. If the underlying latent process has n states then both the mean and variance can take up to a maximum of n values[28].

[28] Depending on the current state of the underlying Markov Chain, not all states may be reached and so some values of mean and variance may not be taken at the next step.

We can get a feeling for the long-run distribution which is a mixture of Gaussian distributions. That is the MSM converges in distribution to a mixture of Gaussian distributions.

Knowledge of the steady state probabilities allows us to calculate the long-run expected returns and standard deviation of returns and should be consistent to what we observe in reality. This thus gives an additional tool when estimating the variables for practical use.

Secondly, knowledge of the steady state variables allows us to specify an initial state vector for the process. In other words, since the states are actually unobservable, we could only conceivably specify the likelihood of the current state. The steady state vector would be the most appropriate given that we know nothing about the previous value of the chain.

3.5. Defining the model

We now introduce the proposed stock model. Firstly we need to define the following terms. Often a subscript or a functional argument is used interchangeably to denote time, usually to allow for allow variable sub and superscripts. It is obvious in the context of the equations in the forthcoming sections which is needed. Furthermore, the dependency on X_t for other variables is often taken as implied and is sometimes dropped from certain expressions.

n	Number of hidden states in the MSM governing the log returns
$X_t, X(t)$	The underlying MSM processes' value at time t. X_t can take up to n values.
$r_t(X_t)$	Risk free rate of return (continuously compounded per annum) at time t when the latent process is in state X_t at time t
$\sigma_t(X_t)$	Standard deviation / volatility parameter of the log returns when the MSM is in state X_t at time t
$\mu_t(X_t)$	drift parameter of the log returns when the MSM is in state X_t at time t
	Discrete Time:
$S_t, S(t)$	The stock price at time t
Δt	The length of time that passes between each step in the process
$\boldsymbol{P}$	A stochastic matrix $\{p_{ij}\}_{1\le i,j\le n}$ where p_{ij} represents the probability that the hidden process X_t will transition from state i to state j in one step of time period length.
	Continuous time:
$\boldsymbol{Q}$	The generator matrix of process X_t represented by $\{q_{ij}\}_{1\le i,j\le n}$

The variables representing the risk free rate, drift and volatility can each take up to n possible values[29]. Often, $r_t(X_t), \sigma_t(X_t)$ and $\mu_t(X_t)$ is abbreviated to μ_i, σ_i and r_i respectively representing the value of each respective parameter when the underlying Markov chain is in state i.

[29] Again, these can only take up to n values because not all states may be reachable from a point in the chain.

Here we have that i represents the state the latent process is in so that there are exactly n values of each parameter. When we refer to the value of the drift, volatility and risk free rate at time t, we will use μ_t, σ_t and r_t.

I introduce some additional useful terms for ease of expression in later applications:

$\boldsymbol{e_i}$	A vector of dimension n which has a value of 1 at entry i and zero everywhere else
$\boldsymbol{r_t}$	A vector of dimension n containing $r_t(X_t)$ for every possible realisation of X_t
$\boldsymbol{\sigma_t}$	A vector of dimension n containing $\sigma_t(X_t)$ for every possible realisation of X_t
$\boldsymbol{\mu_t}$	A vector of dimension n containing $\mu_t(X_t)$ for every possible realisation of X_t

This will allow us to use matrix notation to create some shortcuts. For example, we can now write $\boldsymbol{e_i}^T\boldsymbol{\sigma_t} = \sigma_i$. We can now define the final model.

Definition: Gaussian Markov Switching Stock Price Model[30]

S_t is a stochastic process representing the stock price at time t. The stock price may take any value on the interval $(0, \infty)$.

The process may be a continuous-time or discrete time process.

Let the log return, $R(u,t)$, from time u to time t on a stock be defined as follows:

In continuous time:

$$R(u,t) = \ln\left(\frac{S_t}{S_u}\right) = \int_u^t \left(\mu_s(X_s) - \frac{1}{2}\sigma_s(X_s)^2\right) ds + \int_u^t \sigma_s(X_s)\, dW_s \qquad (3.11)$$

for $0 \le u \le t$ and W_s is Standard Brownian Motion, i.e. $W_t \sim N(0,t)$

In the discrete-time:

$$R(u,t) = \ln\left(\frac{S_t}{S_u}\right) = \sum_{k=0}^{\frac{t-u}{\Delta t}-1} \left(\mu_{u+k\Delta t}(X_{u+k\Delta t}) - \frac{1}{2}\sigma_{u+k\Delta t}(X_{u+k\Delta t})^2\right) + \sum_{k=0}^{\frac{t-u}{\Delta t}-1} \sigma_{u+k\Delta t}(X_{u+k\Delta t})\sqrt{\Delta t}\, Z_k \qquad (3.12)$$

for $0 \le u \le t$, u and t multiples of Δt, $Z_k \sim N(0,1)$ is a set of i.i.d random variables

The stock price at time t is then defined as follows:

$$S_t = S_u e^{R(u,t)} \qquad (3.13)$$

Further properties include:

[30] Simply called a Markov-Switching Model or MSM throughout the dissertation.

(1) S_t and X_t are governed by the same time index. Therefore if S_t is a discrete or continuous-time process, then so is X_t
(2) $R(u,t)$ is a Wiener process and therefore has independent increments given the value of $X(t)$
(3) X_t is a time-homogeneous, unobservable Markov Process with n-integer number of states.
(4) X_t is governed by transition probability matrix $\boldsymbol{P}$ in the discrete-time case and generator matrix $\boldsymbol{Q}$ in the continuous-time case.
(5) $\mu(X_t)$, $\sigma(X_t)$ and $R(X_t)$ (where applicable) can take up to n possible values at each point in time (or each time step for discrete-time).
(6) S_0 is known.
(7) X_0 may be known or unknown. If it is unknown, the initial state distribution $\boldsymbol{\eta}$ is known.

We sometimes simplify matters by merely adjusting the time index such that $\Delta t = 1$. As an example, we can use the idea of a series of daily stock returns. If the time index is measured in days then it is natural that $\Delta t = 1$.

Our main choice of model will be a continuous-time model, although both versions are explored. This is because stock prices trade virtually continuously, and day to day volatility is rather high. When trading times end, the process is merely paused to resume the following day.

In the next section we discuss model estimation in a discrete-time setting. Even though our preferred model may be a continuous one, we can only record discrete measurements when observing a continuous process. More detail is provided in the following section.

4. Parameter Estimation for the MSM

4.1. Introduction: Preliminaries

It is important to note that even though the MSM may be a continuous time model, we can only always observe observations at discrete intervals. It is therefore plausible that we can fit a discrete-time MSM model to the observations. We will see later than transformation from discrete-time to continuous time is fairly straightforward without much loss of information[31] .

Direct estimation of continuous time models are briefly dealt with in Section 5. In this Section, we look to fit a discrete-time model to the discrete observations. We firstly briefly discuss the concept of Maximum Likelihood Estimates and then how we extend this rationale to MSM models. A known existing algorithm, the *Baum-Welch* algorithm (a special case of the expectation-maximisation algorithm) is developed and modified to prevent numerical under and overflow. Finally, the overall estimation routine is summarised for ease of use.

4.2. Introduction: MLE's and the Expectation-Maximization Algorithm

Parameter estimation in statistical models is typically done via maximum likelihood estimators (MLE's). We briefly discuss MLE's and then how to modify the concept when we have an unobservable process. The latter technique is known as the Expectation-Maximization algorithm.

In order to perform an estimation of the parameters we need an observation sequence from the statistical model.

For example, we can assume an observational sequence follows a normal distribution and then the only unknown parameters are the mean and the standard deviation. It comes as no surprise that the longer (or larger) the observational sequence is, the more reliable the estimates become[32].

Establishing MLE's involves choosing parameters that maximizes the joint probability of the observation sequence. This joint probability is termed the likelihood function and may be represented as follows (given n observations of y_i, a model M and parameter set $\boldsymbol{\theta}$).

$$L_Y(\boldsymbol{\theta}|M) = \prod_{i=1}^{n} \Pr[Y = y_i|M,\boldsymbol{\theta}] \tag{4.1}$$

[31] This implies estimation a generator matrix $\boldsymbol{Q}$ from the estimated transition probability matrix $\boldsymbol{P}$
[32] In other words, the standard deviation of the MLE's decrease as n increases

It can be shown that the choice of $\boldsymbol{\theta}$ that maximises the above is equivalent to the choice of $\boldsymbol{\theta}$ that maximises the log-likelihood function. The latter of which is simpler to work with and to differentiate[33].

$$\ln L_Y(\boldsymbol{\theta}|M) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \ln \Pr[Y = y_i|M, \boldsymbol{\theta}] \tag{4.2}$$

For the distributions from the exponential family such as the normal distribution, maximisation becomes further simplified. The log function and the exponential function essentially cancel out, leaving a linear expression to maximize. Maximisation is merely the case of taking partial derivatives and setting them equal to zero. MLE's often take an intuitive form; It can be shown simply (we will omit the details here) that the MLE for μ and σ in a normal distribution is:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.3}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \mu)^2 \tag{4.4}$$

Appling MLE's in the context of MSM's poses some problems. In particular, we have an incomplete data set. We do not know what the current state is and hence we cannot accurately determine the probability of observing a particular observation. In essence, due to the uncertainty over the latent variables, the likelihood function becomes a random variable in itself, dependent on the value of the underlying.

There are a couple of ways to deal with this of which the most well known is the *expectation-maximisation algorithm* (EM algorithm for short) (Dempster, Laird, & Rubin, 1977). The EM algorithm is a generalised method for determining MLE's in the case where there are hidden or latent discrete variables. The observational sequence may be continuous or discrete.

In essence we have the overall data set $Y = (O, X)$, where O is observed and X is unobserved so that Y remains a random variable. Due to the randomness of the likelihood function, the EM algorithm attempts to apply a recursive technique by applying expected values to the log-likelihood function.

E Step (Expectation).

We examine the expected value of the log-likelihood function over possible realisations of X given a parameter set $\boldsymbol{\theta_t}$, the model M, and the observational sequence O.

$$f(\theta|O, M, \theta_t) = \mathbf{E_X}[\ln L_y(\theta_t|M, O)] \tag{4.5}$$

33 Maximising will be done by simply finding the zeroes of the derivative.

M Step (Maximization).

We then determine the choice of θ that maximises the above quantity and denote it as $\boldsymbol{\theta}_{t+1}$.

We then iterate between the E and M step until convergence of $\boldsymbol{\theta}$.

Summary

The logic is as follows:

1) First, provide a guess at the parameter set $\boldsymbol{\theta}$.
2) Given the complete model (θ, M) and the observed data O, find the best estimate[34] of the value of X.
3) Use the estimates for X from (2), the partial model M and observed data O to compute a new estimate of $\boldsymbol{\theta}$.
4) Repeat from step (2) until convergence of θ to get our estimate $\widehat{\boldsymbol{\theta}}$.

As Rabiner (1989) points out, the $\hat{\theta}$ that arises from the above is not necessarily the global maximum of the likelihood function. However, Rabiner (1989) again shows that it is guaranteed to find the local maximum. To avoid this problem, one can try a large array of initial guesses and arrive at a resultant set of $\hat{\theta}$'s and then determine the parameter set that maximises the likelihood function.

The EM algorithm is very general and broad and we will need to apply it to our specific set of circumstances. Firstly, step 2 is not straightforward. In the HMM algorithm, it is akin to problem two discussed in Section 3.2. Even computation of the new $\boldsymbol{\theta}$ is not easy due to the persistent problem of numerical over/underflow[35] and an undefined maximisation routine.

The *Baum-Welch* Algorithm first introduced by (Baum, Petrie, Soules, & Weiss, 1970) is a specific case of the EM algorithm for time homogeneous HMM's and provides us with a distinct routine and explicit optimisation formula for our next iteration of the EM.

[34] Not necessarily the most likely, since we are actually computing the expected value or average value of X which is not necessarily the most likely value. i.e. mean is not necessarily the mode.
[35] This arises due to successive multiplication of probabilities causes probability based expression based on the entire observation set to become very small such that computational software generates either an underflow error (usually resulting in a division by zero error) or an overflow error (the number becomes excessively large).

4.3. The Baum-Welch Algorithm: The Preliminaries

As mentioned above, the Baum-Welch (BW) algorithm is a specific case of the EM algorithm applied to the case of time-homogenous HMM's. It is specifically for discrete state Markov models. The continuous case is dealt with via a Kalmen filter but this is not explored in this thesis. The algorithm can be applied generally to all types of observational distributions and it is only in the final step that we specifically apply it to the MSM case.

The algorithm makes use of two other algorithms, the *forward algorithm* and the *backward algorithm* both of which present alternate ways to specify the full probability of the observation sequence. These two algorithms require some explanation before we can implement BW algorithm. What we will also see is that both algorithms are sensitive to numerical under and overflow and a scaling technique will be developed to combat this.

Ultimately, we will note that that the final estimates of the HMM model will be functions of the other variables developed via the forward and backwards algorithms.

Before we proceed, we need to introduce some notation. In this section, we will assume that $\Delta t = 1$ so that each observation is exactly one time period apart.

T	The total number of observations
$O_1 O_2 \ldots O_t$	The event entailing the occurrence of all observations up to time t. $O = O_1 O_2 \ldots O_T$
$\{p_{ij}\}_{1 \le i,j \le n}$	As before, the transition probabilities for the underlying Markov Process X_t which is collectively described in the stochastic matrix $\boldsymbol{P}$
$w_1 w_2 \ldots w_t$	The event entailing the occurrence of all a specific sequence of the underlying chain up to time t. $\boldsymbol{W} = w_1 w_2 \ldots w_t$. w_t can take up to n values.
π_i	The "steady-state" probability representing the long-run probability of being in state i
$\{b_i(y)\}_{1 \le i \le n}$	The probability of observing y whilst the underlying chain is in state i. B is the matrix representing all possible values of this expression.[36] $\boldsymbol{B}$ is of dimension $n \times y$
M	A full description of the model. i.e $M = \{\boldsymbol{P}, \boldsymbol{B}, \boldsymbol{\Pi}\}$

[36] In the case when y arises from continuous random variable, y can take infinitely many values. In the Gaussian case for example, $\mathbf{B}$ is of dimension $n \times \dim(\mathbb{R})$. This is not computationally tractable but fortunately the BW algorithm avoids this complication as we will see. The $b_i(y)$ are Gaussian probability density functions. In the general case, matrix $\boldsymbol{B}$ is known as an emission probability matrix.

4.4. The forward algorithm and forward probabilities

4.3.1. Introduction

As per Mitra (2010), we need a way to compute $\Pr[O|M]$. To do so we can specify this quantity as follows:

$$\begin{aligned}
\Pr[O|M] &= \sum\nolimits_{All\ W} \Pr[O, W|M] \\
&= \sum\nolimits_{All\ W} \Pr[O|M, W] \,.\, \Pr[W|M] \\
&= \sum\nolimits_{All\ W} b_{q_1}(O_1).\, b_{w_2}(O_2) \ldots b_{w_T}(O_T).\, \Pr[Q|M] \\
&= \sum\nolimits_{All\ W} b_{w_1}(O_1).\, b_{w_2}(O_2) \ldots b_{w_T}(O_T).\, \pi_{w1} p_{w_1 w_2} p_{w_2 w_3} \ldots p_{w_{T-1} w_T} \qquad (4.6)
\end{aligned}$$

There are n^T possible values of W. This is computationally very heavy as the number of observations increases. Furthermore, the structure of the above is such that the successive multiplication of probabilities makes this expression susceptible to numerical underflow. To combat the former problem, the forward algorithm is introduced which requires $n^2 T$ computations.

To combat this we introduce a new variable, the forward variable $\alpha_t(i)$ as follows:

Definition: The forward variable

$$\alpha_t(i) = \Pr[O_1 O_2 \ldots O_t, w_t = i|M] \qquad (4.7)$$

In other words, it is the probability of observing the partial sequence up to time t and being in state i at time t given the model. Using the above we can use the law of total probability to establish $\Pr[O|M]$ in terms of the forward variable as follows:

$$\Pr[O|M] = \sum_{i=1}^{n} \alpha_T(i) \qquad (4.8)$$

To determine $\alpha_T(i)$ we can develop an iterative relation since the forward variable can be expressed in terms of the prior forward variable. We can then determine $\alpha_T(i)$ by recursion. We have the following iterative relationship.

$$\alpha_{t+1}(j) = \Pr[O_1 O_2 \dots O_{t+1}, w_{t+1} = j | M]$$

$$= \Pr[O_1 O_2 \dots O_t, w_{t+1} = j | M] . \Pr[O_{t+1} | w_{t+1} = j, M]$$

$$= \left(\sum_{i=1}^{n} \Pr[O_1 O_2 \dots O_t, q_t = i, w_{t+1} = j | M] \right) . b_j(O_{t+1})$$

$$= \left(\sum_{i=1}^{n} \Pr[O_1 O_2 \dots O_t, w_t = i] . \Pr[w_t = i, w_{t+1} = j | M] \right) . b_j(O_{t+1})$$

$$= \left(\sum_{i=1}^{n} \alpha_t(i) p_{ij} \right) b_j(O_{t+1}) \tag{4.9}$$

This can be understood in probabilistic terms. The first term in the bracket represents the probability of observing the sequence $O_1 O_2 \dots O_t$ and then transitioning to state j - the summation considers all possible originating states at time t[37]. The second term then represents the probability that observation O_{t+1} occurring given that we are in state j at that time.

In all recursive systems, we will need to know either the end value or the beginning value in order to start the recursion. Note that $\alpha_1(i) = \Pr[O_1, w_1 = i | M]$. If we know the initial state, then this is merely the probability of observing the first observation. However, if this is unknown we can make a very good approximation to this probability by employing steady state probabilities for the Markov Chain part[38]:

Algorithm: Forward Variables

$$\alpha_1(i) = \pi_i b_i(O_1) \tag{4.10}$$

We then have the following recursive scheme:

1. *Initialisation for* $1 \leq i \leq n$

$$\alpha_1(i) = \pi_i b_i(O_1) \tag{4.11}$$

2. *Iterative step for* $1 \leq t \leq T-1, 1 \leq j \leq n$

$$\alpha_{t+1}(j) = \sum_{i=1}^{n} \alpha_t(i) p_{ij}\, b_j(O_{t+1}) \tag{4.12}$$

[37] Including state j itself!
[38] This is especially true if the chain has been running for a very long time.

When the process terminates, we have a $\alpha_T(i)$ for each state i allowing us to calculate $\Pr[O|M]$ as described in equation (4.8).

As before, the issue is that there are many multiplications of values less than 1, especially as n and t becomes large.

To overcome this, Rabiner (1989) proposed a scaling technique which protects against numerical underflow which we will explore in the next section.

4.3.2. Scaled forward variables and probabilities

Scaling is done by multiplying by a scaling constant at each time step. The scaling value, c_t is independent of state and as such ensure that scaled forward variables and still in the same proportion to each other as were the unscaled forward variables. However, obviously, the scaled values lose their meaning as probabilities.

We need some new notation: $\hat{\alpha}_t(i)$ is the scaled forward variable; $\ddot{\alpha}_t(i)$ is an intermediate quantity used to determine the scaling constant at time t. The following scheme is proposed:

Algorithm: Scaled forward variables

1. *Initialisation for* $1 \leq i \leq n$

$$\ddot{\alpha}_1(i) = \alpha_1(i) = \pi_i b_i(O_1) \tag{4.13}$$

$$c_1 = \frac{1}{\sum_{i=1}^{n} \ddot{\alpha}_1(i)} \tag{4.14}$$

$$\hat{\alpha}_1(i) = c_1 \ddot{\alpha}_1(i) \tag{4.15}$$

2. *Iterative step for* $1 \leq t \leq T-1, 1 \leq j \leq n$

$$\ddot{\alpha}_t(j) = \sum_{i=1}^{n} \hat{\alpha}_{t-1}(i) p_{ij}\, b_j(O_t) \tag{4.16}$$

$$c_t = \frac{1}{\sum_{i=1}^{n} \ddot{\alpha}_t(i)} \tag{4.17}$$

$$\hat{\alpha}_t(i) = c_t \ddot{\alpha}_t(i) \tag{4.18}$$

Because of the scaling $\hat{\alpha}_t(i)$, we can see that the sum of these values (at each time step) over each possible state will always sum to 1. This protects against numerical underflow.

4.3.3. Likelihood quantities in terms of scaled variables

What if we wish to compute $\Pr[O|M]$? First note that we can show that $\hat{\alpha}_t(i) = \left(\prod_{\tau=1}^{t} c_\tau\right) \alpha_t(i)$ by induction: Clearly the result holds for $t = 1$ by the routine above. Assume it holds for some integer s and try show it holds for the next integer:

$$
\begin{aligned}
\hat{\alpha}_{s+1}(i) &= c_{s+1}\ddot{\alpha}_{s+1}(i) \\
&= c_{s+1}\sum_{j=1}^{n} \hat{\alpha}_s(i)p_{ji}\, b_i(O_{s+1}) \\
&= c_{s+1}\sum_{j=1}^{n} \left(\prod_{\tau=1}^{s} c_\tau(i)\right) \alpha_s(i)p_{ji}\, b_i(O_{s+1}) \\
&= c_{s+1}\left(\prod_{\tau=1}^{s} c_\tau(i)\right)\sum_{j=1}^{n} \alpha_s(i)p_{ji}\, b_i(O_{s+1}) \\
&= \left(\prod_{\tau=1}^{s+1} c_\tau(i)\right) \alpha_{s+1}(i)
\end{aligned}
\tag{4.19}
$$

This result is important to applying the next little trick:

$$
\begin{aligned}
1 &= \sum_{i=1}^{n} \hat{\alpha}_T(i) = \sum_{i=1}^{n}\left(\prod_{t=1}^{T} c_t\right)\hat{\alpha}_t(i) = \left(\prod_{t=1}^{T} c_t\right)\sum_{i=1}^{n}\hat{\alpha}_t(i) \\
&= \left(\prod_{t=1}^{T} c_t\right)\Pr[O|M]
\end{aligned}
\tag{4.20}
$$

Therefore $\Pr[O|M] = {}^{1}\!/\!\left(\prod_{t=1}^{T} c_t\right)$. Denote the denominator as C_T which we will use in later calculations.

We will see how we can adjust the BW algorithm to accommodate the scaled values. We now turn to another algorithm, the *backward algorithm* which can be considered a sort of complement of the forward algorithm. It is another way to calculate $\Pr[O|M]$.

4.4. The backward algorithm and backward probabilities

4.4.1. Introduction

We define a new variable, $\beta_t(i)$, called the backwards variable, or backwards probability, as follows:

Definition: Backwards variable

$$\beta_t(i) = \Pr[O_{t+1}O_{t+2}\dots O_T|w_t = i, M] \tag{4.21}$$

Note that $\Pr[O|M]$ can then be expressed in terms of the backwards variable:

$$\begin{aligned}
\Pr[O|M] &= \Pr[O_1O_2\dots O_T|M] \\
&= \sum_{j=1}^{n} \Pr[O_1O_2\dots O_T, w_1 = j|M] \\
&= \sum_{j=1}^{n} \Pr[O_1O_2\dots O_T|\, w_1 = j, M]\Pr[w_1 = j|M] \\
&= \sum_{j=1}^{n} \Pr[O_2O_3\dots O_T|\, w_1 = j, M]\,\Pr[O_1|w_1 = j, M]\,\pi_j \\
&= \sum_{j=1}^{n} \beta_1(j)\, b_j(O_1)\pi_j
\end{aligned} \tag{4.22}$$

We can also express the backwards variable in terms of the next backwards variable as follows:

$$\begin{aligned}
\beta_t(i) &= \Pr[O_{t+1}O_{t+2}\dots O_T|w_t = i, M] \\
&= \sum_{j=1}^{n} \Pr[O_{t+1}O_{t+2}\dots O_T, w_{t+1} = j|w_t = i, M] \\
&= \sum_{j=1}^{n} \Pr[O_{t+1}O_{t+2}\dots O_T|\, w_{t+1} = j, w_t = i, M]\Pr[w_{t+1} = j|w_t = i, M]
\end{aligned}$$

$$= \sum_{j=1}^{n} \Pr[O_{t+2} \dots O_T | \, w_{t+1} = j, M] \Pr[O_{t+1} | w_{t+1} = j, M] p_{ij}$$

$$= \sum_{j=1}^{n} \beta_{t+1}(j) \, b_j \, (O_{t+1}) p_{ij} \tag{4.23}$$

This allows us to use a recursive scheme to calculate $\Pr[O|M]$ but we need an initial step. Note that (as the name implies) we need a terminal value to initialize in comparison to a beginning value to start the recursion. We need a value for $\beta_T(i)$ independent of the recursion formula. By the original definition, this is not strictly valid but note that $\beta_{T-1}(i)$ is:

$$\beta_{T-1}(i) = \Pr[O_T | w_{T-1} = i, M]$$

$$= \sum_{j=1}^{n} \Pr[O_T, w_T = j | w_{T-1} = i, M]$$

$$= \sum_{j=1}^{n} \Pr[O_T | w_T = j, w_{T-1} = i, M]. \Pr[w_T = j | w_{T-1} = i, M]$$

$$= \sum_{j=1}^{n} \Pr[O_T | w_T = j, M]. p_{ij}$$

$$= \sum_{j=1}^{n} b_j \, (O_T) \, . \, p_{ij} \tag{4.24}$$

Then by using the definition of the backwards variable in equation (4.21) above, it is clear that in order for the definition to work, $\beta_T(i) = 1$. This is then our initial value to start the recursion.

Algorithm: Backwards variable

1. *Initialisation for* $1 \leq i \leq n$

$$\beta_T(i) = 1 \tag{4.25}$$

2. *Iterative step for* $1 \leq t \leq T-1, 1 \leq i \leq n$

$$\beta_t(i) = \sum_{j=1}^{n} \beta_{t+1}(j) \, b_j \, (O_{t+1}) p_{ij} \tag{4.26}$$

4.4.2. Scaled backwards variables and probabilities

In a similar fashion to the forward scaled variables, we begin with defining the scaled backwards variable $\hat{\beta}_t(i)$. We again define an intermediate variable $\ddot{\beta}_t(i)$ which will aid calculation. We will use the same scale factors as per the forward variables. We then use the following recursive scheme:

Algorithm: Scaled Backwards variable

1. *Initialisation for* $1 \leq i \leq n$

$$\ddot{\beta}_T(i) = \beta_T(i) = 1 \tag{4.27}$$

$$\hat{\beta}_T(i) = c_T \ddot{\beta}_T(i) \tag{4.28}$$

2. *Iterative step for* $1 \leq t \leq T-1, 1 \leq i \leq n$

$$\ddot{\beta}_t(i) = \sum_{j=1}^{N} p_{ij} b_j(O_{t+1}) \hat{\beta}_{t+1}(j) \tag{4.29}$$

$$\hat{\beta}_t(i) = c_t \ddot{\beta}_t(i) \tag{4.30}$$

In a similar induction way as per the forward scaled variables, we can show that $\hat{\beta}_t(i) = (\prod_{s=t}^{T} c_s)\beta_t(i)$. Similar to the case with the forward variable, we denote the first product as D_t. We can then make the observation that $C_t.D_{t+1} = C_T$

4.5. The Baum-Welch Algorithm: Functions of Forward and Backward Probabilities

The tools of the forward and backwards probabilities (and their scaled versions) allow us to define two additional variables that will help us finalise our BW algorithm.

It will be useful to be able to calculate the probability of being in state i at time t given the observational sequence. This will ring true if we are for example attempting to determine the likelihood of being in a certain market regime given what we have observed in recent times.

Fortunately, this can be expressed in terms of forward and backward probabilities. Define $\gamma_i(t)$ to be this probability. I.e.

Definition: $\boldsymbol{\gamma_i(t)}$ ***- The probability of being in state*** i ***given the observation sequence***

$$\gamma_i(t) = \Pr[w_t = i | O, M] \quad (3.14)$$

We can then represent this quantity as follows:

$$\begin{aligned} \gamma_i(t) &= \Pr[w_t = i | O, M] \\ &= \frac{\Pr[w_t = i, O | M]}{\Pr[O | M]} \\ &= \frac{\Pr[w_t = i, O_1 O_2 \dots O_t, q_{t+1} = j | M]. \Pr[O_{t+1} O_{t+2} \dots O_T | w_{t+1} = j, M]}{\Pr[O | M]} \\ &= \frac{\alpha_t(i) \beta_t(i)}{\Pr[O | M]} \end{aligned} \quad (4.31)$$

Another useful quantity is the concept of the probability of a specific transition occurring at a certain time given the entire observational sequence. We can define this as $\xi_t(i, j)$.

Definition: $\boldsymbol{\xi_t(i, j)}$***: probability of transitioning to state j from state*** i ***at time*** t ***given the entire observation sequence.***

$$\xi_t(i, j) = \Pr[w_t = i, w_{t+1} = j | O, M] \quad (3.15)$$

This can again be represented in terms of forwards and backwards probabilities.

$$\begin{aligned} \xi_t(i, j) &= \Pr[w_t = i, w_{t+1} = j | O, M] \\ &= \frac{\Pr[w_t = i, w_{t+1} = j, O | M]}{\Pr[O | M]} \\ &= \frac{\Pr[w_t = i, w_{t+1} = j, O_1 O_2 \dots O_t | M]. \Pr[O_{t+1} O_{t+2} \dots O_T | w_t = i, w_{t+1} = j, M]}{\Pr[O | M]} \\ &= \frac{\alpha_t(i) p_{ij}. \Pr[O_{t+1} | w_{t+1} = j, M]. \Pr[O_{t+2} \dots O_T | w_{t+1} = j, M]}{\Pr[O | M]} \end{aligned}$$

$$= \frac{\alpha_t(i) p_{ij} \, b_j(O_{t+1}) \beta_{t+1}(j)}{\Pr[O|M]} \tag{4.32}$$

Note the link between $\gamma_i(t)$ and $\xi_t(i,j)$: $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$. That is by summing over all the destination states we simply get the probability of being in state i at time t given the observational sequence.

Now note some other functions and their meanings which follow intuitively from this:

$\sum_{t=1}^{T-1} \gamma_t(i)$	Expected number of times a transition from state i occurs throughout the sequence.
$\sum_{t=1}^{T} \gamma_t(i)$	Expected number of times a state i is visited throughout the sequence.
$\sum_{t=1}^{T-1} \xi_t(i,j)$	Expected number of times a transition from state i to j occurs throughout the sequence.
$\gamma_1(i)$	The probability of initially being in state i given the observation sequence

4.6. The Baum-Welch Algorithm: Main steps

The functions in the above section allow us to simply get an understanding of the next iteration of estimates for the parameters. We essentially need to estimate three sets of parameters $\{p_{ij}\}, \{\pi_i\}$ and $\{b_j(O_t)\}$. In the MSM case, we have:

$$b_j(y) \equiv b_j(y; \varrho_j, \sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(y-\varrho_j)^2}{2\sigma_j^2}\right) \tag{4.33}$$

It thus suffices to simply estimate $\{\varrho_j\}$ and $\{\sigma_j\}$.

We note that in the MSM case: $\varrho_j = \mu_j - \frac{1}{2}\sigma_j^2$

Omitting the details of the exact calculation of these estimates, we can apply a broad brush intuitive guess as to what they should be. Recall that having found the expected likelihood given the data, we now seek to find the MLE of the parameters denoted by a bar.

Intuitively, $\bar{p}_{ij}$ should be the expected number of transitions from state i to state j divided by the expected number of transitions from state i.

From the previous table above we see that we can express this explicitly and importantly in terms of the scaled forwards and backwards variables:

$$\begin{aligned}
\bar{p}_{ij} &= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\
&= \frac{\sum_{t=1}^{T-1} \alpha_t(i) p_{ij}\, b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(i)} \\
&= \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i)/C_t . p_{ij}\, b_j(O_{t+1}) \hat{\beta}_{t+1}(j)/D_{t+1}}{\sum_{t=1}^{T-1} \hat{\alpha}_t(i)/C_t . \hat{\beta}_t(i)/D_t} \\
&= \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i) p_{ij}\, b_j(O_{t+1}) \hat{\beta}_{t+1}(j)/C_T}{\sum_{t=1}^{T-1} \hat{\alpha}_t(i)/C_t . \hat{\beta}_t(i)/(D_{t+1} c_t)} \\
&= \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i) p_{ij}\, b_j(O_{t+1}) \hat{\beta}_{t+1}(j)/C_T}{\sum_{t=1}^{T-1} \hat{\alpha}_t(i) . \hat{\beta}_t(i)/C_T c_t} \\
&= \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i) p_{ij}\, b_j(O_{t+1}) \hat{\beta}_{t+1}(j)}{\sum_{t=1}^{T-1} \hat{\alpha}_t(i) . \hat{\beta}_t(i)/c_t}
\end{aligned} \tag{4.34}$$

Now performing this calculation over all possible states allows us to get the new estimate of the transition probability matrix. This in turn enables the calculation of the new steady state estimates π_i which will be used in the next round of iterations.

We now turn to $\bar{\varrho}_i$: As per equation (4.3), we can get an intuitive idea of what the estimate should be. For example, if we could separate out the observations where the HMM is in state i we would simply average out these observations to get the MLE. However, due to the hidden nature of the chain we simply weight each observation by the probability that the chain is in state i at that time (given the observation sequence) and sum over all possible times.

$$\begin{aligned}
\bar{\varrho}_i &= \frac{\sum_{t=1}^{T} \gamma_t(i) O_t}{\sum_{t=1}^{T} \gamma_t(i)} \\
&= \frac{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i) O_t}{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i)} \\
&\frac{\sum_{t=1}^{T} \hat{\alpha}_t(i)/C_t . \hat{\beta}_t(i)/D_t . O_t}{\sum_{t=1}^{T} \hat{\alpha}_t(i)/C_t . \hat{\beta}_t(i)/D_t} \\
&\frac{\sum_{t=1}^{T} \hat{\alpha}_t(i)/C_t . \hat{\beta}_t(i)/(D_{t+1} c_t) . O_t}{\sum_{t=1}^{T} \hat{\alpha}_t(i)/C_t . \hat{\beta}_t(i)/(D_{t+1} c_t)} \\
&\frac{\sum_{t=1}^{T} \hat{\alpha}_t(i) . \hat{\beta}_t(i)/c_t . O_t}{\sum_{t=1}^{T} \hat{\alpha}_t(i) . \hat{\beta}_t(i)/c_t}
\end{aligned} \tag{4.35}$$

For $\bar{\sigma}$ we will apply a similar logic by weighting the squared deviations from ϱ_i by the probability that the HMM was in state i at that time. Refer to equation (4.4) for the ordinary MLE.

$$\bar{\sigma}_i = \sqrt{\frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(O_t - \bar{\varrho})^2}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)}}$$

$$= \sqrt{\frac{\sum_{t=1}^{T} \gamma_t(i)(O_t - \bar{\varrho})^2}{\sum_{t=1}^{T} \gamma_t(i)}}$$

$$= \sqrt{\frac{\sum_{t=1}^{T} \hat{\alpha}_t(i)/C_t.\hat{\beta}_t(i)/D_t.(O_t - \bar{\varrho})^2}{\sum_{t=1}^{T} \hat{\alpha}_t(i)/C_t.\widehat{\beta}_t(i)/D_t}}$$

$$= \sqrt{\frac{\sum_{t=1}^{T} \hat{\alpha}_t(i).\hat{\beta}_t(i)/c_t.(O_t - \bar{\varrho})^2}{\sum_{t=1}^{T} \hat{\alpha}_t(i).\hat{\beta}_t(i)/c_t.\widehat{\beta}_t(i)}} \quad (4.36)$$

Equations (4.35) and (4.36) are the expressions given by Harte (2006) and developed further in terms of the scaled forwards and backwards variables.

We calculate the above quantities from equation (4.34), (4.35), π_i and (4.36) and use it to calculate the scaled forward and backwards probabilities in the next iteration.

The question would next be when we should stop the iteration? We'd like to stop the process when the difference between $\Pr[O|M]$ at the current iteration and the previous iteration is negligible[39]. Rabiner (1989) shows that the BW algorithm always produces estimates that guarantee that $\Pr[O|M]$ is an increasing function for each parameter iteration[40]. We omit the theoretical details for this proof. However, this implies that we will only reach a maximum $\Pr[O|M]$ after infinitely many iterations and so we need a stopping mechanism or criteria.

The algorithm requires an initial estimated to get it started. In the following section we discuss one possible method that could be used to give the first guess.

4.7. Choice of Initial Parameter set

To begin the algorithm, we require an initial "guess" at the parameter estimates. The choice is not arbitrary since different initial choices could very well lead to different optimal estimates (Liu, Davis, Lovell, & Kootsookos, 2004).

The optimal set of parameters will be achieved by performing a run of the algorithm based on each possible initial guess; recording the optimal parameters and the value of $\Pr[O|M]$; and then choose the parameter set corresponding to the highest value of $\Pr[O|M]$. Practically, running the algorithm over every possible initial guess is unfeasible and practically we need to constrain ourselves to a small number of good initial guesses.

---

[39] For programming purposes, we would need to define a negligible number in absolute terms. This is difficult since the probability decreases the more observations and states there are. Another option is to keep iterating until the difference in the parameter estimates between iterations becomes negligible. Here negligibility is more easily defined, for a probability, the criteria may be 0.001

[40] Here "increasing" refers to the operator "≥" as opposed to "strictly increasing represented by ">"

Since we have the observational data sequence we can calculate some descriptive statistics including the mean, standard deviation, range and interquartile range.

Therefore one "neutral" set of initial values is as follows:

$$p_{ij} = {}^{1}/_{n} \tag{4.37}$$

$$\varrho_i = {}^{\sum_{i=1}^{n} O_i}/_{n} \tag{4.38}$$

$$\sigma_i = {}^{\sum_{i=1}^{n} (O_i - \varrho_i)^2}/_{n} \tag{4.39}$$

Starting parameter sets could then use this as a base value and then vary from that. There is no guide on exactly how to do this but we could follow some intuition. We may add and subtract different increments to p_{ij} and add and subtract multiples of half the standard deviation calculated in the descriptive statistics to ϱ_i.

The above paragraph discusses the case when we have little understanding of what the parameters might be but this is not usually the case. When it comes to market data, the states are usually representing different sentiment or regimes of volatility. Despite the hidden nature of the chain, the market may well know if it is a prolonged bear phase such as the financial crisis of 2008-2009. Dissecting the time series of returns into regimes may be easily done by observing a graph of it, especially if the variances of the observations under each state are low[41].

However, stock market data is notoriously volatile and if state persistence is short then it becomes unclear from an apparent data anomaly whether a state change occurred or the anomaly is just due to a period of excess volatility.

Nevertheless, there exists scope in some visual techniques to acquire at least a rough approximation to what we could expect the parameters to be – the hope being that the algorithm does the rest and takes them to the right values.

4.8. Summary of complete routine

This section briefly provides a summary of the complete routine for reference purposes. It also provides a guide as to the order of computation. The k'th iteration will be denoted by a superscript (k). We start with two initialization steps:

[41] In the extreme case, a zero variance for each state creates a jump process where observation values merely alternate between their means at each state change.

Algorithm: The Complete MSM Estimation Routine

0) *We start with the initial guesses:* $p_{ij}^{(0)}$, $\varrho_i^{(0)}$, $\sigma_i^{(0)}$ *for* $1 \leq i,j \leq n$

1) *Use* $\left\{p_{ij}^{(k)}\right\}_{1 \leq i,j \leq n}$ *to determine* $\left\{\pi_i^{(k)}\right\}_{1 \leq i \leq n}$ *using standard Markov Chain techniques*[42]*.*

We then run the following algorithm until $\Pr[O|M]^{(k)} - \Pr[O|M]^{(k-1)} < \varepsilon$*, for* $k = 0,1,2 \ldots$ *where* ε *is some negligible number*[43]*.*

1) *We then employ the scaled forward algorithm outlined in equations (4.13)-(4.18). In this routine we store the values for* $\hat{\alpha}_t^{(k)}(i)$ *and for every* t *and* i *and* $c_t^{(k)}$ *for every* t*.*

   a) *Initialisation for* $1 \leq i \leq n$

$$\ddot{\alpha}_1^{(k)}(i) = \alpha_1^{(k)}(i) = \pi_i^{(k)} b_i^{(k)}(O_1)$$

$$c_1^{(k)} = \frac{1}{\sum_{i=1}^{n} \ddot{\alpha}_1^{(k)}(i)}$$

$$\hat{\alpha}_1^{(k)}(i) = c_1^{(k)} \ddot{\alpha}_1^{(k)}(i)$$

   b) *Iterative step for* $1 \leq t \leq T-1, 1 \leq j \leq n$*. Here we have that* $b_j^{(k)}(O_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\left(O_t - \varrho_j^{(k)}\right)^2}{2\sigma_j^{(k)}}\right)$*:*

$$\ddot{\alpha}_t^{(k)}(j) = \sum_{i=1}^{n} \hat{\alpha}_{t-1}^{(k)}(i) p_{ij}^{(k)} b_j^{(k)}(O_t)$$

$$c_t^{(k)} = \frac{1}{\sum_{i=1}^{n} \ddot{\alpha}_t^{(k)}(i)}$$

$$\hat{\alpha}_t^{(k)}(i) = c_t^{(k)} \ddot{\alpha}_t^{(k)}(i)$$

2) *Calculate* $\Pr[O|M]^{(k)} = \frac{1}{\prod_{t=1}^{T} c_t^{(k)}}$ *and store until the next iteration (the next* k*)*

We would have liked to run the backwards algorithm in the iterative step but we need to start with the value of c_T *which is unavailable until the entire forward algorithm has run; so we need to run these individually in their own "loop".*

[42] This is addressed in Section 3.1
[43] For the first run of the algorithm we ignore the initial criteria. That is, we essentially assume that there is no restriction on $\Pr[O|M]^{(k)} - \Pr[O|M]^{(k-1)} < \varepsilon$ for the first run since $\Pr[O|M]^{(0)}$ does not make sense.

3) *We now turn to equations (4.27)-(4.30) and run the scaled backwards algorithm storing* $\beta_t^{(k)}(i)$ *for each* t *and* i*:*

   a) *Initialisation for* $1 \leq i \leq n$

$$\ddot{\beta}_T^{(k)}(i) = \beta_T^{(k)}(i) = 1$$

$$\hat{\beta}_T^{(k)}(i) = c_T^{(k)} \ddot{\beta}_T^{(k)}(i)$$

   b) *Iterative step for*$1 \leq t \leq T-1, 1 \leq i \leq n$*. Note that here we need* i *to run from* $T-1$ *down to* 1*. As before we have* $b_j^{(k)}(O_{t+1}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\left(O_{t+1}-\varrho_j^{(k)}\right)^2}{2\sigma_j^{(k)}}\right)$:

$$\ddot{\beta}_t^{(k)}(i) = \sum_{j=1}^{N} p_{ij}^{(k)} b_j^{(k)}(O_{t+1}) \hat{\beta}_{t+1}^{(k)}(j)$$

$$\hat{\beta}_t^{(k)}(i) = c_t^{(k)} \ddot{\beta}_t^{(k)}(i)$$

4) *We now begin to calculate the next set of estimates from equations. Firstly from equation (4.34) for* $1 \leq i, j \leq n$

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t^{(k)}(i) p_{ij}^{(k)} b_j^{(k)}(O_{t+1}) \hat{\beta}_{t+1}^{(k)}(j)}{\sum_{t=1}^{T-1} \hat{\alpha}_t^{(k)}(i).\hat{\beta}_t^{(k)}(i)/c_t^{(k)}}$$

5) *We then calculate* $\pi_i^{(k+1)}$ *for every* i *using* $\left\{p_{ij}^{(k+1)}\right\}_{1 \leq i,j \leq n}$.

6) *Then we use equations (4.35) and (4.36) to find the remaining parameters as follows:*

$$\varrho_i^{(k+1)} = \frac{\sum_{t=1}^{T} \hat{\alpha}_t^{(k)}(i).\hat{\beta}_t^{(k)}(i)/c_t^{(k)}.O_t}{\sum_{t=1}^{T} \hat{\alpha}_t^{(k)}(i).\hat{\beta}_t^{(k)}(i)/c_t^{(k)}}$$

$$\sigma_i^{(k+1)} = \sqrt{\frac{\sum_{t=1}^{T} \hat{\alpha}_t^{(k)}(i).\hat{\beta}_t^{(k)}(i)/c_t^{(k)}.(O_t - \varrho_i^{(k+1)})^2}{\sum_{t=1}^{T} \hat{\alpha}_t^{(k)}(i).\hat{\beta}_t^{(k)}(i)/c_t^{(k)}.\widehat{\beta_t^{(k)}}(i)}}$$

7) *Let* $k = k+1$ *and return to step 1.*

Note again that $\varrho_i = \mu_i - \frac{1}{2}\sigma_i^2$ so that $\mu_i = \varrho_i + \frac{1}{2}\sigma_i^2$.

As a final add-on section, we briefly discuss the estimation of the risk-free rate which has not been considered in the above methodology.

4.9. The issue of the risk free rate

So far, we have skirted over the issue of a Markov-modulated risk free rate. This becomes important when we consider option pricing under the MSM.

We could simplify the issue by simply assuming that the risk free rate is constant regardless of the state. The difficulty comes when we wish to have the risk-free rate also governed by an underlying hidden process.

There are two issues here:

One is that although stock returns and the risk free rate are loosely correlated. In the past, the equity and bond markets have been shown to be negatively correlated (Wainscott, 1990). It is likely that the two will not be governed by the same chain or further than state jumps will not occur simultaneously. However, this is arguable, especially given our idea of states representing stages of the economy.

Secondly, even if we assume that state transitions and probability transition matrices are consistent, how should we go about estimating the respective rates in each state?

Markov-regulated interest rate models are beyond the scope of this thesis but have been explored elsewhere. See for example Smith (2002).

A simple method for example would be to simply average out all past-period risk free rates (sourced possibly from a government T-bill rate of appropriate term) and then make suitable estimates of the rates in each state such that the long-run rate is that of the average.

When we study the SA market, we will make a simple approximation for illustration purposes.

Since this is not the theme of this thesis, we will not mention more on the estimation of the risk free rate.

This Section has outlined an estimation procedure for a discrete-time MSM but it can be applied to continuous time MSM's since we can only record observations in discrete time (for example the closing prices of the daily stock prices. One aspect not discussed is the "dimension" of the model – i.e. how many states there are. This is dealt with in Section 6.

Before we deal with this aspect, we have a brief diversion to discuss another estimation procedure. This involves uses moment-matching to estimate continuous-time MSM's more directly.

5. Moment Based Methods for Determining Drift and Volatility

5.1. Introduction

In this section we examine an alternative method of estimation. We apply regression techniques to determine the drift and volatility under each of the hidden states. This section largely summarises the work of Elliot, Krishnamurthy, & Sass (2008) and is aimed at estimating a continuous-time directly. It is not used in the remainder of the dissertation and has been provided merely as an alternative estimation technique for completeness.

When we explored the Baum-Welch algorithm, we "discretised" the stock price information since the stock price is essentially a continuous-time process[44]. This is an appropriate method if we choose the time interval to be sufficiently small. "Sufficiently small" means that the probability of a two state jumps during the time interval is negligible[45]. In the cases we explore, we are ideally looking for long term market shifts (bull or bear markets) but a similar theory can be applied to match much shorter term investor sentiment (See Section 1 for more discussion). It thus becomes necessary to explore the scenario where state changes occur frequently. As Elliot, Krishnamurthy, & Sass (2008) points out, there is often a minimal interval we can apply due to the nature of the process[46] and so it becomes necessary to explore some techniques in continuous-time parameter estimation. In this case, we no longer turn to estimation of the probability transition matrix[47], but rather estimation of the generator matrix directly. One method of finding the generator matrix is described in Israel, Rosenthal, & Wei (2001).

Although estimating the generator matrix directly is particularly appealing, continuous time estimation of the generator matrix (using so called "Markov Chain Monte Carlo" methods) for "noisy" chains is not sufficiently stable or fast when compared to the discretisation of the time interval (Hahn & Sass, 2009)[48]. The authors go on to try and marry the two methods (discretisation vs direct generator estimation) to propose a compromise solution.

Even if the discretisation method is chosen (and then the BW algorithm will apply) then we are still stuck with the problem of the initial estimates for the drift and volatility. In this section we explore a moment-matching method in order to estimate these parameters.

For other techniques in estimating these parameters, please consult Hahn, Fruhwirth-Schnatter, & Sass (2007).

[44] Stock prices are available virtually every few seconds.
[45] If only one state jump occurs then it will be as if the state jump occurred at the end of the time period. If two state jumps occur then the first state jump will not be recognised at all and hence there is a loss of information. Negligible is again not a clearly defined term but we could keep reducing the interval until the BW parameter estimates converge.
[46] We often only have daily opening, closing stock or mid prices.
[47] This loses most of its meaning when analysing continuous time Markov processes.
[48] These methods use a Bayesian approach.

5.2. Drift and Volatility Calculations

We can write $R(u,t)$ from Section 3 in terms of Ito integrals as follows :

$$R(u,t) = \int_u^t \varrho_z(X_z)\, dz + \int_u^t \sigma_z(X_z)\ dW_z \tag{5.1}$$

Here, W_z is the Brownian motion and $\varrho_z = \mu_z - \frac{1}{2}\sigma_z^2$. Let $R_{\Delta t} \equiv R(0,\Delta t) \equiv R(u, u+\Delta t)$ for all $u > 0$. We can then consider a moment matching technique whereby we calculate many generalised moments, $\mathbf{E}[R_{l\Delta t}^m]$ where $l, m \in \mathbb{N}^+$. In other words $R_{l\Delta t}^m$ is the m'th raw moment of $l\Delta m$-period log returns.

Elliot, Krishnamurthy, & Sass (2008) use the notation $\hat{R}_{m,l}$ to describe this generalised moment based on the observational data which is the notation I will also adopt.

We can easily calculate a set of observed moments as follows:

$$\hat{R}_{m,l} = \frac{1}{[N/L]} \sum_{i=0}^{[N/L]} \left(R_{(i+l)\Delta t} - R_{i\Delta t}\right)^m \tag{5.2}$$

Here the "$[*]$" operator denotes the integer part of the argument. L is the number of lags we choose. We can then choose a parameter set that minimises the squared difference between the theoretical moments[49] and the observed moments as follows:

$$\sum_{m=1}^{M} \sum_{l=1}^{L} (\hat{R}_{m,l} - \mathbf{E}[R_{l\Delta t}^m])^2 \tag{5.3}$$

M represents the highest raw moment considered. We now need a definition of $\mathbf{E}[R_{l\Delta t}^m]$ for all m and l. Elliot, Krishnamurthy, & Sass (2008) provides a comprehensive proof for an explicit routine/formula for determining the generalised moments. We will omit the details of the proof here and just repeat the result for a two state homogenous HMM chain. Readers are referred to the original text for more information.

Recall from Section 3.1 that the sum of entries in the row of the rate matrix $\boldsymbol{Q}$ must sum to zero. In other words, for a general time homogenous two-state MSM we have:

$$\boldsymbol{Q} = \begin{pmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{pmatrix}$$

Then define a negative sum of the diagonal entries as $\omega = q_1 + q_2$. Then define some functions we will need to simplify the process:

$$\overline{\varrho^k} = \pi_1 \varrho_1^k + \pi_2 \varrho_2^k \tag{5.4}$$

[49] Which are functions of the parameter sets for the drift and volatility

$$\overline{\sigma^l} = \pi_1\sigma_1^l + \pi_2\sigma_2^l \tag{5.5}$$

$$\overline{\varrho^k\sigma^l} = \pi_1\varrho_1^k\sigma_1^l + \pi_2\varrho_2^k\sigma_2^l \tag{5.6}$$

When $k = 1$ or $l = 1$, we will omit the superscript. We can then express the first four raw moments in terms of ω, ϱ^k, σ^l and $\varrho^k\sigma^l$ as follows:

$$\mathbf{E}[R_t] = \bar{\varrho}t \tag{5.7}$$

$$\mathbf{E}[R_t^2] = \overline{\sigma^2}t + \overline{\varrho^2}t^2 + 2\left(\overline{\varrho^2} - \bar{\varrho}^2\right)\left(e^{-\omega t} - 1 + \omega t - \frac{(\omega t)^2}{2}\right)/\omega^2 \tag{5.8}$$

$$\begin{aligned}\mathbf{E}[R_t^3] &= 3\overline{\varrho\sigma^2}t^2 - \overline{\varrho^3}t^3 + 6\overline{(\varrho\sigma^2} - \bar{\varrho}\overline{\sigma^2})\left(e^{-\omega t} - 1 + \omega t - \frac{(\omega t)^2}{2}\right)/\omega^2 \\ &+ 12\left(\bar{\varrho}\overline{\varrho^2} - \bar{\mu}^3\right)\left(1 - \omega t + \frac{(\omega t)^2}{2} - \frac{(\omega t)^3}{6} - e^{-\omega t}\right)/\omega^3 \\ &+ 6\left(\overline{\varrho^3} - 2\bar{\varrho}\overline{\varrho^2} + \bar{\varrho}^3\right)\left(-2 + \omega t + (2 + \omega t)e^{-\omega t} - \frac{(\omega t)^3}{6}\right)/\omega^3\end{aligned} \tag{5.9}$$

$$\begin{aligned}\mathbf{E}[R_t^4] &= 3\overline{\sigma^4}t^2 + 6\overline{\varrho^2\sigma^2}t^3 + \overline{\varrho^4}t^4 + 6\left(\overline{\sigma^4} - \overline{\sigma^2}^2\right)\left(e^{-\omega t} - 1 + \omega t - \frac{(-\omega t)^2}{2}\right)/\omega^2 \\ &+ 6\left(4\overline{\varrho^2}\,\overline{\sigma^2} + 8\bar{\varrho}\overline{\varrho\sigma^2} - 12\bar{\varrho}^2\overline{\sigma^2}\right)\left(1 - \omega t + \frac{(\omega t)^2}{2} - \frac{(\omega t)^3}{6} - e^{-\omega t}\right)/\omega^3 \\ &+ 6\left(6\overline{\varrho^2\sigma^2} - 4\overline{\varrho^2}\,\overline{\sigma^2} - 8\bar{\varrho}\overline{\varrho\sigma^2} - 6\bar{\varrho}^2\overline{\sigma^2}\right)\left(-2 + \omega t + (2 + \omega t)e^{-\omega t} - \frac{(\omega t)^3}{6}\right)/\omega^3 \\ &+ 72\left(\overline{\bar{\varrho}^2\overline{\varrho^2}} - \bar{\varrho}^4\right)\left(e^{-\omega t} - 1 + \omega t - \frac{(\omega t)^2}{2} + \frac{(\omega t)^3}{6} - \frac{(\omega t)^4}{4!}\right)/\omega^4 \\ &+ 72\left(\varrho\overline{\varrho^3} - 2\overline{\varrho^2}\bar{\varrho}^2 + \bar{\varrho}^4\right)\left(3 - 2\omega t + \frac{(\omega t)^2}{2} - (3 + \omega t)e^{-\omega t} - \frac{(-\omega t)^4}{4!}\right)/\omega^4 \\ &+ 24\left(\overline{\varrho^4 + 3\bar{\varrho}^2\overline{\varrho^2}} - 3\bar{\varrho}\overline{\varrho^3} - \bar{\varrho}^4\right)\left(-3 + \omega t + \left(3 + 2\omega t + \frac{(\omega t)^2}{2}\right)e^{-\omega t} \right. \\ &\left. - \frac{(\omega t)^4}{4}\right)/\omega^4\end{aligned} \tag{5.10}$$

One can imagine how the formulae for higher moments look for those HMMs with more than two states!

One point of consideration is the choice of L and M in equation (5.3). The greater the value, the larger the number of computations needed. Further, as the amount of computations increase, higher moments and as such it can become applicable to make use of a weighting scheme to attach importance to the certain moments. For example, equation (5.3) can also be written in matrix form as follows:

$$(\boldsymbol{\eta}-\widehat{\boldsymbol{\eta}})^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{\eta}-\widehat{\boldsymbol{\eta}}) \tag{5.11}$$

where $\boldsymbol{\eta}$ is a vector of form $(\mathbf{E}[R_{\Delta t}], \dots \mathbf{E}[R_{L\Delta t}] \dots, \mathbf{E}[R^M_{\Delta t}], \dots, \mathbf{E}[R^M_{L\Delta t}])$ and $\hat{\eta}$ represents the empirical version of this matrix with contents represented by equation (5.2). $\boldsymbol{\Lambda}$ is a square weighting matrix of choice. Hence the matrix $\boldsymbol{\eta}$ has ML components. Elliot, Krishnamurthy, & Sass (2008) propose the following potential three weighting schemes (see also Anderson & Sorenson (1996) and Hansen (1982)).

$$\begin{aligned} &(1) \quad \Lambda_{ii} = 1/\eta_{ii}^2 \\ &(1) \quad \Lambda_{ij} = 0 \text{ for } i,j = 1 \dots n, i \neq j \end{aligned} \tag{5.12}$$

$$\begin{aligned} &(2) \quad \Lambda_{ii} = 1/\gamma_{ii} \\ &(2) \quad \Lambda_{ij} = 0 \text{ for } i,j = 1 \dots n, i \neq j \end{aligned} \tag{5.13}$$

$$(3) \quad \boldsymbol{\Lambda} = \{\gamma_{ij}\}^{-1}_{1 \le i,j \le n} \tag{5.14}$$

where γ_{ij} refers to the covariance between the moments. Defining an expression for the covariance between these terms can get complex but the following definition is suitable:

$$\gamma_{ij} = \gamma_{i,j} = \begin{cases} \mathbf{Cov}[R_{i\Delta t} R_{j\Delta t}] & 1 \le i,j \le L \\ \mathbf{Cov}[R^2_{(i-L)\Delta t} R^2_{(j-L)\Delta t}] & L < i,j \le 2L \\ \mathbf{Cov}[R^3_{(i-2L)\Delta t} R^3_{(j-2L)\Delta t}] & 2L < i,j \le 3L \\ \vdots & \\ \mathbf{Cov}[R^M_{(i-(M-1)L))\Delta t} R^M_{(j-(M-1)L)\Delta t}] & (M-1)L < i,j \le ML \end{cases} \tag{5.15}$$

The definition of γ_{ij} for i and j which are not in the same "L interval" length follows logically. For example: if $L < i \le 2L$ and $2L < j \le 3L$ then $\gamma_{ij} = \mathbf{Cov}[R^2_{(i-L)\Delta t} R^3_{(j-2L)\Delta t}]$.

Weighting scheme (1) corresponds to minimising the relative error given the sample mean (Elliot, Krishnamurthy, & Sass, 2008). Weighting scheme 3 is an obvious choice but since the dimension of the matrix is so large in comparison to the other schemes, it can lead to large scale estimation errors. For this reasons Ho, Perraudin, & Sorenson (1996) proposed that just the diagonal elements are used (which is weighting scheme (2)).

Despite the above method, we are still assuming that we know the number of hidden states when reality we do not. In the next section we explore some rationale and reasoning behind determining the correct number of states to be assumed.

6. Examining different regimes

6.1. Introduction

In the previous two sections we discussed how we can estimate the drift, volatility and transition probability of the underlying MSM. Implicit in the assumption was that the number of states in the chain is known but in reality this is rarely the case. Indeed, knowledge of the number of states violates the idea of the underlying process being hidden.

In the context of stock models, regimes changes are expected and practitioners often have a good feeling of whether a regime shift has occurred: This is often because regime shifts are often linked to other observable processes. For example, the collapse of many major banks in the US in late 2008 spurred on an extended period of excess volatility which in the context of our models can be described as a regime shift.

Ultimately, we understand that the market's view on the fortunes of the constituent companies that make up the stock market are what ultimately drives the market in the long term. Thus the factors that affect the fortunes of these companies give us some insight into whether regime switches have occurred or at the very least, the times when a regime shift is more likely than another time. Factors include general economic variables, budget speeches, monetary policy review meetings and foreign exchange policies amongst many others.

There may therefore then be some basis for empirically examining past return data to look for regime switches. Derman (1999) uses such an approach when examining implied volatilities of option prices on the S&P 500. This approach will certainly be acceptable if we do not believe in time varying transition probabilities and that the number of regimes present in the past data will persist into the future[50].

We can also turn to more rigorous methods to determine the "dimension" of the underlying HMM. The main problem is that the test statistics have no standard distribution (Li & Hao, 2005) so testing hypothesis regarding the number of states becomes difficult.

In the simple case, we could imagine that we can just include n (the number of states) as an unknown parameter in the BW algorithm. This presents one immediate problem: the number of other parameters required to be estimated are directly dependent on the number of states. The dimension of parameter vector $\boldsymbol{\theta}$ is non-constant. The second problem is that the likelihood under the BW algorithm is a non-decreasing function in n (Figueiredo, Leitao, & Jain, 1999). This essentially means we cannot use the BW algorithm to determine n and need to seek other methods in order to establish the correct state.

The standard likelihood ratio test cannot be used to determine the number states (Titterington, Smith, & Markov, 1985). The standard likelihood ratio is: $L^* = -2[\ln L(\boldsymbol{\theta}|M_{H_1}) - L(\ln L(\boldsymbol{\theta}|M_{H_0})$. Here the subscripts under the respective M's, H_1 and H_0 denotes the alternative and null hypothesis respectively. Under certain regularity conditions, this

[50] In other words, the dynamics of the underlying HMM will not vary over time.

distribution has approximately a χ^2 distribution with degrees of freedom equal to the dimension of $\boldsymbol{\theta}$. Guidici, Ryden, & Vandekerkhove (2000) highlight these regularity conditions in the appendix in technical detail. Loosely speaking, the regularity conditions are satisfied if the underlying Markov Chain is irreducible[51] and aperiodic[52]. The more technical regularity conditions break down when the number of states is included in the parameter vector $\boldsymbol{\theta}$. This results in the information matrix becoming singular and the asymptotic distribution of the likelihood ratio statistic is not χ^2 (Rios, 2008) and thus renders the standard likelihood ratio test useless in this context. In essence models with different numbers of states are just too different. Guidici, Ryden, & Vandekerkhove (2000) do show that the likelihood ratio is approximately χ^2 but results from this test will likely need to be backed up against with other tests.

A simple alternative is to simply overparameterise the model by choosing a state number known to be too large. Franq & Roussignol (1997) used this method and proposed that the "excess" states would then simply yield the same parameters as the other states and one would be able to thus effectively determine the number states required.

A very common method is the use of an Information Criterion such as the Akaike Information Criteria (AIC) or the Bayes Information Criteria (BIC). Various papers discuss the appropriateness of these methods (these are discussed in the forthcoming sections). The idea is to determine the information criteria value under the assumption of many different states and then choose the model which yields the lowest value. These statistics are often a function of the likelihood function but incorporate a term that penalizes the statistics for high values of n.

The above approaches are all based on the EM algorithm (or in our case the BW algorithm). Stochastic approaches also exist: Most notably Markov Chain Monte Carlo methods (MCMC).

The next section develops these ideas a further and gives an understanding of the practical application needed in order to determine the number of states.

6.2. Information Criterion

6.2.1. Introduction

The process of using an information criterion can be computationally expensive. Although the value is usually a direct function of the likelihood function, the full BW algorithm needs to be run each time for each different state. The problem is magnified because of the local maximum problem[53] which may lead to long run times.

[51] It is possible to get to any state from any other state (all states can communicate with each other).
[52] That is, returns to states can occur at irregular times (the minimum period that elapses for state return times is 1 time step).
[53] Recall that the BW algorithm finds a local maximum only and that the algorithm may need to be run many times using different initial guesses before the global maximum is found through trial and error.

It could be argued though, that the number of hidden states a stock market environment has will stay constant over time, even if a valid argument for time dependent transition probabilities, means and variances exist. I thus don't regard computational intensity as a major criterion when selecting the method for determining the number of states.

6.2.2. Akaike Information Criterion (AIC)

Definition: AIC Criterion

$$AIC(M) = 2k - 2L(\boldsymbol{\theta}|M) \tag{6.1}$$

where k is the number of parameters in the model so that models with many parameters are penalized.

The AIC is a broad brush approach to testing for model selection.

We can then choose a model with say n^* states of the MSM which is the model with the minimum AIC.

6.2.3. Bayesian Inference Criterion (BIC) & Minimum Description Length (MDL)

The BIC corresponds to the MDL (Figueiredo, Leitao, & Jain, 1999) and is given by the following equation.

Definition: BIC Criterion

$$MDL(M) = BIC(M) = L(\boldsymbol{\theta}|M) - \frac{N_n(M)}{2}\ln T \tag{6.2}$$

where $N_n(M)$ is the number of parameters needed to specify model M with n hidden states.

Also recall that T is the number of observations in our model.

Again we choose the number states equal to n^*, the value of n corresponding to the lowest value of MDL. MacKay (2002) shows in her paper that the MDL/BIC method has some mathematical validity for model comparisons when determining the number of states. The MDL was first introduced by Rissanen (1978) under the motto: "Choose the model that gives the shortest description of data" (Chambaz, Garivier, & Gassiat, 2005).

There is one more criterion often mentioned in the literature, the Minimum Message Length (MML) criterion. However, Oliver, Baxter, & Wallace (1996) state that the MML criterion is "superficially" similar to the MDL criterion and so the MML criterion will not be discussed further.

6.2.4. Mixture Minimum Description Length (MMDL)

One shortcoming of the MDL criterion is that it effectively weights each type of parameter evenly (by $\frac{logT}{2}$). In our model, the parameters associated with an arbitrary state j, μ_j and σ_j, are only calculated from observations were the HMM was in state j and not the entire observation sequence. Only the transition probabilities and stationary probabilities are calculated on the entire data set. There is thus an argument against equal-weighting.

The MMDL criterion was developed as a variation of the MDL to apply the theory behind MDL to mixture HMMs in particular. However, building on the principles of MMDL, we can create our own so-called "cost function" to use in our model.

We start by decomposing N_n into its constituents by observing that $N_n = N_n(\boldsymbol{P}) + N_n(\boldsymbol{\pi}) + N_n(\mathbf{I})$ where $N_n(*)$ in this context is the number of parameters needed to estimate the argument * under the assumption of n states of the latent process.

As discussed above, $N_n(\boldsymbol{P})$ and $N_n(\boldsymbol{\pi})$ are based on the entire sample and should accordingly receive the standard weight of $\frac{\ln T}{2}$. Recall also that matrix $\boldsymbol{B}$ is the emission probability matrix. For the i'th row of $\boldsymbol{B}$ we only require estimates of 2 sets of parameters, μ_i and σ_i and these parameters are only estimated from a subset of the observational sequence – those observations where the latent process was in state i. Since π_i can be thought of as the average time spent in state i in the long run, $T\pi_i$ represents the expected number of observations that were generated when the latent process was in state i. A weight of $\ln(T\pi_i)$ is thus appropriate (Bicego, Murino, & Figeuiredo, 2003). Since each row only has two parameters, we require $2n$ estimates to fully specify $\boldsymbol{B}$.

Recall that $\boldsymbol{P}$ is our transition probability matrix and so it is a stochastic matrix[54]. Therefore, one only needs to estimate $n-1$ parameters in each row. We thus require only $n(n-1)$ parameter estimates to fully specify $\boldsymbol{P}$.

Using similar logic, it is easy to see that we require $n-1$ estimates to fully specify $\boldsymbol{\pi}$ (the stationary distribution of the underlying latent process). We can now manipulate the original MDL criterion from equation (6.2) and apply our unequal weighted scheme:

[54] This means that the rows of the matrix sum to unity.

$$
\begin{aligned}
MMDL(M) &= L(\boldsymbol{\theta}|M) - \frac{N_n(\boldsymbol{P}) + N_n(\boldsymbol{\pi})}{2}\ln T - \frac{N_n(\boldsymbol{B})}{2}\sum_{i=1}^{n}\ln(T\pi_i) \\
&= L(\boldsymbol{\theta}|M) - \frac{n(n-1)+(n-1)}{2}\ln T - \frac{2n}{2}\sum_{i=1}^{n}\ln(T\pi_i) \\
&= L(\boldsymbol{\theta}|M) - \frac{n^2-1}{2}\ln T - n\sum_{i=1}^{n}\ln(T\pi_i) \\
&= L(\boldsymbol{\theta}|M) - \frac{n^2}{2}\ln T - n\sum_{i=1}^{n}\ln(T\pi_i) - \frac{1}{2}\ln T \qquad (6.3)
\end{aligned}
$$

Since we will be comparing models based only on differences between n we can ignore terms independent of n: In other words, when comparing different values of the MMDL, the term $\frac{1}{2}\ln T$ will remain constant. We may thus drop it to arrive at an appropriate expression for our model:

Definition: MMDL (Customised for MSM)

$$MMDL(M) = L(\boldsymbol{\theta}|M) - \frac{n^2}{2}\ln T - n\sum_{i=1}^{n}\ln(T\pi_i) \qquad (6.4)$$

6.3. Stochastic Methods: MCMC

MCMC methods involve assigning parameters a prior distribution and then applying a Bayesian approach to obtaining the posterior distribution (and posterior estimate) of the number of states. This approach has the advantage of estimating both the MSM's parameters and the number of states n simultaneously (Richardson & Green, 1997).

The methods involve simulating via Monte Carlo methods to achieve a posterior distribution for the number of regimes (Neal, 1991). A common method is to use the Dirichlect distribution as the prior (used in Neal (1991) and Otranto & Gallo (2001)).

Bayesian methods have one very nice appealing feature namely that we are able to favour a specific regime through our choice of the prior. This can be quite meaningful in this context again when we should have an idea of the number of expected regimes just by observing empirical data.

However as many authors point out, (Figueiredo, Leitao, & Jain (1999), Neal (1991) and Otranto & Gallo (2001)) these methods are very computationally expensive. Even though I stated that computing time is not a high cost in the context we use the models, the novelty of using information criteria together with the BW algorithm is appealing. Furthermore, use of MMDL has a solid theoretical underpinning. MCMC methods are thus not explored in any further depth.

6.4. "Sequential Pruning"

Sequential Pruning is discussed briefly by Bicego, Murino, & Figeuiredo (2003). The premise provides us with an algorithm once we have chosen a selection criterion.

Algorithm: Sequential Pruning

(0) *Determine* n_{max} *and* n_{min} *which are the minimum and maximum number of states allowed or that are feasible. In our context,* $n_{min} = 2$*.* n_{max} *will be discussed when we look specifically at the data.*

Now start a loop in n*, intialised by* $n = n_{max}$ *until* $n = n_{min}$*.*

(1) *Run the BW-algorithm*[55] *and store* $\boldsymbol{\Pi}, \boldsymbol{P}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ *and* $L[\boldsymbol{\theta}|M^{(n)}]$*. We call the optimal set of parameters under* n *states model* $\widehat{M}^{(n)}$*.*

(2) *Compute the information criterion value for* n *states, represented by* $C^{(n)}$*.*

(3) *Identify the state with lowest steady state probability*[56]*. Remove all entries corresponding to that state from the parameter vector. For the now truncated matrix* $\boldsymbol{P}$*, uniformly scale up the remaining elements in each row so that it once again becomes a stochastic matrix*[57]*. Similarly, scale up the remaining elements in* π *such that all components in the vector sum to unity. This is known as the "pruning" step.*

(4) *Set* $n = n - 1$ *and use the resulting modified parameters as initial parameters for the next run in step 1.*

Once the algorithm is complete, one will have values for $C^{(n)}$ *for* $n_{min} \leq n \leq n_{max}$*. There optimal value for* n*:* n^**, is that value which satisfies* $n^* = max_n[C^{(n)}]$*.*

There are subtle advantages to sequential pruning over standard model selection. Standard model selection involves simply running the BW algorithm for different states under the same initial conditions and choosing the model corresponding to the highest information criterion. However, using this method under some information criterions such as the BIC or MDL can underestimate the true number of states (Figueiredo, Leitao, & Jain, 1999). Sequential pruning allows us to use a "nearly good" set of parameters as intialisation for the next

[55] The very first run will use a data based estimates for the initial run. $M^{(n)}$ refers to a model with n states.
[56] This corresponds to the state that is least likely to be visited over the long run.
[57] The sum of all elements in the same row is equal to 1.

iteration of the BW algorithm (Bicego, Murino, & Figeuiredo, 2003). This is logically very appealing and Bicego, Murino, & Figeuiredo (2003) show that it leads to less iterations of the BW algorithm and so more frequently predicted the correct number of states than the standard method in their training data.

Having both logical appeal backed up by some empirical studies, the method chosen for state selection in Section 11 will involve the modified MMDL criterion and a sequential pruning method.

7. Complete vs. Incomplete Markets

7.1. Definitions

We begin by defining what it means for a market to be complete.

Definition: Complete Markets

A market is said to be complete if any claim or option maturing at time T can be hedged, that is, there exists a self-financing portfolio of market assets whose value at time T is the value of the claim, in all possible states of the world.

The concept of market completeness is an important one for pricing options. It was instrumental in laying the foundation for the BS formula as seen in Section 2.

Market completeness allows us to create a portfolio of market instruments at an option's inception such that we can exactly replicate an option payment at expiry. If we assume that no arbitrage exists[58], then the current price of the self-financing portfolio must be the current price of the option. This unique price is often referred to as the no-arbitrage price. If the prices are not equal, market players are able to create an arbitrage portfolio. This is defined as follows:

Definition: Arbitrage Portfolio

An Arbitrage Portfolio is a self-financing portfolio, $V(t)$ that satisfies the following:

1) $V(0) = 0$
2) *There exists a* $T > 0$ *such that*
    a. $\Pr[V(T) \geq 0] = 1$ *and*
    b. $\Pr[V(T) > 0] > 0$

Creation of these portfolios usually involves selling instruments or portfolios that are above the no-arbitrage price and buying instruments that are below the price. The absence of arbitrage leads to the *law of one price* which states that all identical goods must have the same price.

The construction of a replicating portfolio was possible under the BS environment by selling and buying combinations of stock and cash to hedge out changes in the stock price.

[58] This is a very realistic assumption. If an instrument is priced such that an arbitrage opportunity exists, market players will react by exploiting this opportunity. This action brings prices back to equilibrium (the correct price) so that the arbitrage opportunity vanishes.

Loosely speaking, we need one asset for every source of randomness. In an MSM case, there are two sources of randomness, the Brownian Motion driving the volatility and the underlying Markov Chain that changes the parameters. To complete the market, we need a second financial instrument. One such instrument is a change-of-state security which is discussed in the next section.

This leads us to the fundamental theories of asset pricing discussed by Harrison & Pliska (1981). The theorem relates the absence of arbitrage to the presence of the existence of martingales that are equivalent to the real-world measure and cause the discounted underlying stock price to be a martingale. The First Fundamental Theorem of Asset Pricing is as follows:

Theorem: First Fundamental Theorem of Asset Pricing

A probability measure $\mathbb{Q}$ is said to be a Risk Neutral Measure (RNM) or Equivalent Martingale Measure (EMM) for a market with probability measure $\mathbb{P}$ if:

(i) $\mathbb{P} \sim \mathbb{Q}$ *($\mathbb{P}$ is an equivalent measure to $\mathbb{Q}$)*
(ii) *The discounted stock price is a $\mathbb{Q}$-martingale*

Furthermore we have the following equivalent statements:

(1) A market is complete if and only if there exists a unique equivalent martingale measure
(2) If the market is complete, then every derivative has a unique no-arbitrage price

In terms of incomplete markets, the following statements are equivalent:

(1) In an incomplete market, there exists infinitely many equivalent martingale measures
(2) In an incomplete market, there may be infinitely many "no-arbitrage prices"

In summary, we see that we need to first identify market completeness. If the market is complete, then there is only one unique EMM. On the other hand, if the market is incomplete we need to select an EMM from the set of allowable EMM's.

Markets are usually incomplete because there are more sources of risk than tradable assets but incompleteness may also exist because of illiquidity, transaction costs or portfolio constraints[59]. Traders are forced to partially replicate contingent claims and then take some risks at the end resulting in a "no-arbitrage price" interval within which any option could be priced at without the market exploiting it.

Our market in which MSM's govern the share price is clearly incomplete; there are more sources of risk than assets. In Section 9.1, we present a discrete "binomial" MSM model and we can easily see that there is no combination of stock and bonds that will allow us to hedge the option for all possible next step outcomes. Since the continuous time model can be

[59] For example: inability to short sell, borrowing constraints or position limits.

thought of as a limiting model of the discrete one, it provides an alternative framework for the understanding of market incompleteness.

In the section that follow, we briefly touch on some possible techniques to deal with market incompleteness and expand more fully on the chosen technique, Esscher transforms in more detail.

7.2. Variance-Optimal Martingale Measure and Mean-Variance Hedging

The goal in this technique is to select a self-financing portfolio to minimize the "hedging risk" at the time of the claim. The argument is as follows (Schweizer, 1992):

A call option has payoff at time T is $H = (S_T - K)^+$. Assume this is valued as V_0 at time 0. Assume we follow a self-financing strategy whose final value is g_T. We then undertake the following exercise: write a call option and receive V_0 now and invest this in risk free asset B_t (the risk-free bond, assuming $B_0 = 1$). At the end of the term, the "payoff" is $V_0 B_T + g_T - H$ which costs us g_0^* at the beginning. Our profit at time T is thus $V_0 B_T + g_T - H - g_0^* B_T$.

Now note that:

$$\mathrm{Var}[(V_0 - g_0)B_T + g_T - H] = \mathrm{Var}[g_T - H] \tag{7.1}$$

.

We first wish to find the optimal value for g such that it minimizes the variance of the profit. Once we find g^* then the optimal value for V_0, V_0^* is that value which satisfies $\mathbf{E}[(V_0^* - g_0^*)B_T + g_T^* - H] = 0$.

Rearranging the above leads gives us:

$$V_0^* = g_0^* + \mathbf{E}\left[\frac{H}{B_T} - \frac{g_T^*}{B_T}\right] \tag{7.2}$$

In the BS framework, it's possible to find a strategy g such that $g_T^* = H$ which results in $V_0^* = g_0^*$ for any value of the underlying stock price model[60]. This also causes the variance of the profit to be zero.

In incomplete markets, profit is not zero under all conditions and so we get a profit distribution depending on the values of the underlying stock price (and self-financing strategy chosen).

The variance-optimal martingale measure is then that measure, $\mathbb{Q}^V$, equivalent to the real-world measure $\mathbb{P}$, that satisfies:

[60] It is necessary that H be a square integrable claim.

$$\mathbf{E}_{\mathbb{Q}^V}\left[\frac{H}{B_T}\right] = g_0^* + \mathbf{E}\left[\frac{H}{B_T} - \frac{g_T^*}{B_T}\right] \tag{7.3}$$

Where g_0^* and g_T^* are determined from equation (7.1).

7.3. Minimal Martingale Measure and Local-Risk Minimizing

The Minimal Martingale Measure (MMM) is a specific EMM that arises when one follows the process of local-risk minimizing. The actual specification of the MMM is quite technical and readers are referred to Follmer & Schweizer (1991) for more details. Below I outline the simple concept of risk minimization as a basis for the MMM in the continuous time case.

Consider a portfolio strategy denoted as φ. θ represents the number of risky assets and V_t represents the value of the portfolio at time t.

We now define a cost process of portfolio strategy φ, $C_t(\varphi)$ as follows:

Definition: Cost Process of portfolio strategy φ

$$C_t(\varphi) = V_t - \int_0^t \theta_u dS_u \tag{7.4}$$

If $C_t(\varphi)$ is constant under the real-world measure $\mathbb{P}$ a.s, then it is self-financing. It is called mean self-financing if its cost process $C_t(\varphi)$ is a $\mathbb{P}$-martingale (Pham, 2000).

We know define a risk measure for portfolio strategy φ as follows:

Definition: Risk measure for portfolio strategy φ

$$R_t(\varphi) = \mathbf{E}[\left(C_T(\varphi) - C_t(\varphi)\right)^2 | \mathfrak{I}_t] \tag{7.5}$$

If we restrict ourselves to only portfolio strategies that can replicate[61] the claim H then the optimal risk minimizing strategy is that portfolio strategy φ^* that satisifies:

$$R_t(\varphi^*) \leq R_t(\varphi), \forall t \in [0, T] \tag{7.6}$$

[61] A strategy is known to be H-admissible if $V_T = H$ a.s

A risk minimization strategy may not always exist and the concept of local-risk minimization is introduced. Without delving into the technical details, roughly speaking a portfolio strategy φ is locally risk-minimising if, for any $t \in [0, T)$, $R_t(\varphi)$ is minimal under all infinitesimal perturbation of the strategy at time t (Pham, 2000).

The minimal martingale measure will then be that EMM, $\mathbb{Q}^M$ associated with the optimal locally risk-minimizing strategy. In other words:

$$V_0^* = \mathbf{E}_{\mathbb{Q}^M}[H] \tag{7.7}$$

7.4. Minimal Entropy Measure

Let $\mathbb{Q}$ be a probability measure equivalent to $\mathbb{P}$. We then define the relative entropy, $I(\mathbb{Q}, \mathbb{P})$ between measures $\mathbb{Q}$ and $\mathbb{P}$ as follows:

Definition: Relative entropy

$$I(\mathbb{Q}, \mathbb{P}) = \mathbf{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}} \ln\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \tag{7.8}$$

If $\mathbb{Q}$ is not equivalent to $\mathbb{P}$ then the relative entropy is ∞.

We now wish to find a measure $\mathbb{Q}^E \in \mathcal{Q}$ (where $\mathcal{Q}$ is the set of all probability measures equivalent to $\mathbb{P}$) such that $I(\mathbb{Q}^E, \mathbb{P}) = \min_{\mathbb{Q} \in \mathcal{Q}} I(\mathbb{Q}, \mathbb{P})$.

The probability measure $\mathbb{Q}^E$ is then known as the minimum entropy martingale measure (MEM). Entropy can loosely be thought of as the “distance” between measures (Fritelli, 2000).

Importantly, under certain technical conditions, the MEM is unique. Proof can be found in Fritelli (2000).

7.5. Esscher Transforms

Esscher transforms were first used in the context of modeling aggregate claims distributions (Gerber & Shiu, 1994) and is used in the context of continuous random variables (and in our case, continuous time processes). It arose as a transform of single random variables as follows:

Let $f_X(x)$ be a probability density function of random variable X. Let h be real number such that the following function exists:

$$g\big(h, f(x)\big) = \int_{-\infty}^{\infty} e^{hx} f(x)dx \tag{7.9}$$

Then $E_X(f(x), h)$ is known as the Esscher Transform of density $f(x)$ with parameter h. It is also a probability density function and is defined as follows:

$$E(f(x), h) = \frac{e^{hx} f(x)}{g(h, f(x))} \tag{7.10}$$

This is the Esscher transform of a single random variable. We could also similarly define an Esscher transform of a process $X(t)$. To do this, we need some preliminary results:

Let's assume $X(t)$ is an infinitely divisible continuous random variable with stationary and independent increments. Define $M(z,t) = \mathbf{E}[e^{zX(t)}]$ which represents the moment generating function of process $X(t)$. It can be proven that $M(z,t) = [M(z,1)]^t$ if the function is continuous at $t = 0$ (Gerber & Shiu, 1994). If the density of $X(t)$ is $f(x,t)$, then the moment generating function is:

$$M(z,t) = \int_{-\infty}^{\infty} e^{zx} f(x,t)dx \tag{7.11}$$

The above has similar form to $g\big(h, f(x)\big)$ from above. Suppose the above function exists, then there is at least one value z for which the function exists. Further, if there is a value of t for which the function exists then it exists for all t by the property $M(z,t) = [M(z,1)]^t$. Under these circumstances, we may equate the two functions $g\big(h, f(x)\big) = M(h,t)$.

We can then define the Esscher transform of a process $X(t)$:

Definition: Esscher transform on process $X(t)$

$$E_{X(t)}(f(x,t), h) = \frac{e^{hx} f(x,t)}{M(h,t)} \tag{7.12}$$

This is again a probability density function of a process with stationary and independent increments. This essential means we have modified or transformed the probability measure.

Furthermore this measure is equivalent to the original measure since the exponential function is always non-negative and both functions agree on the sets of zero probability.

Depending on the choice of h we may derive different measures equivalent to the original measure. We now seek a measure such that the Esscher transform is an EMM. I.e. we seek to find an $h = h^*$ such that the expected discounted value of an asset is a martingale. Gerber & Shiu (1994) have proved that h^* is unique.

$E_{X(t)}(f(x,t), h^*)$ is then called the risk neutral Esscher Transform and is unique. There is a close link between the MEM and the Esscher transform (Miyahara, 2001).

In the next section, we apply a simple method to complete the market, which ignores the necessity of selecting an EMM and presents theory on how to price vanilla options.

In Section 9 consideration towards the application and practical methods for pricing MSM's in incomplete markets is given. Initially this is done in discrete time where the notion of complete and incomplete markets is more intuitively understood.

8. Option Pricing Under Change-of-State Securities

8.1. Introduction

As we have noted in the previous Section, a market where the stock price is governed by an MSM is incomplete.

Guo X. (1999) proposed a security that could be used to complete the market and in particular hedge away the regime risk. It is called a "Change-of-State" security (CoS). It works as follows:

At each time t there exists a contract that pays out 1 unit at the next time the process changes its state. Where that waiting time is $\tau(t) = \inf\{u > t | X(u) \neq X(t)\}$. Since $\tau(t)$ is a waiting time, it is plausible to define $\tau(t)$ as an exponential random variable with rate parameter λ_i. Here the rate parameter is state dependent as time to the next state change is wholly dependent on the current state - a typical Markov Chain property.

Once that payout is over, the next payout is at the next change of state. Thus, the security pays out "dividends" at each change of state.

We further assume that this security is infinitely divisible and of unlimited supply in the market so that if one would like to hedge a known loss R but where the *timing* of the loss is unknown. One could then purchase a CoS security to create a replicating portfolio. Such a payout does exist in the Insurance world. One example is a whole-life life assurance contract. Usually this involves paying a policyholder a fixed amount known as the sum assured on his or her death. Hence the amount is known but the timing is not.

Although we find that the CoS security is mathematically very appealing (as it completes the market), there are some very major practical drawbacks:

Firstly, it is in direct contradiction of our understanding of the model. The state changes are governed by a latent process and we are only able to observe the stock price. Thus the value of the underlying Markov process and hence the timing of the state changes are, by their very definition, unknown. An investor could only ever estimate when the states change.

These leads us on to the second major drawback, namely that security like this does not exist in any major financial markets. In my opinion, the closest example of this is a derivative that pays out when a volatility index reaches a certain level. The volatility index, VIX can be an indicator of investor confidence (CBOE, 2011) and this fits nicely with our earlier understanding of why the stock market volatility is governed by a Hidden Markov Process. Thus, this derivative could act as a good proxy for a CoS security.

Again there are some drawbacks, namely that the volatility index will always be an average volatility over some past time period which means that true state changes will be reflected at a later stage (there will be a lag). And secondly, since the stock has general stock volatility, the regime effect on volatility may be swamped by general randomness. The larger the "window period" the more the regime switches are easily observed. However, by the same token, the larger the window period the longer the lag time will be (it will take longer to recognize regime shifts from when they actually occur). Thus an unfortunate trade off exists.

Besides the obvious drawbacks, use of the CoS security may give us an idea of a good lower limit of the option price. The idea being that the existence of only quasi-CoS securities would add a margin to the price due to the added risks discussed above.

Because COS securities eliminate arbitrage in the market, there is unique probability measure $\mathbb{Q}$ (a unique EMM[62]) under which all securities can be valued (Harrison & Pliska, 1981).

We thus have the value of this security $CoS_t(X_t)$ at time t, and currently in state i, defined as follows:

Definition: Change of State (CoS) security

$$CoS_t(X_t) = \mathbf{E}_{\mathbb{Q}}[1.e^{-r(\tau(t)-t)}|\mathfrak{I}_t] \tag{8.1}$$

8.2. Completing the market using CoS securities

What follows is a brief review of the use of CoS securities sourced mostly from the work of Guo X. (1999) in his Ph.D dissertation.

We can define a supplementary stochastic process $N(t)$, a counting process which counts the number of times a state change occurs. Using the rate matrix, $\boldsymbol{Q}$, we can specify the rate parameter under some technical conditions[63] under measure $\mathbb{Q}$:

$$\lambda_i = -q_{ii} \tag{8.2}$$

We can then define the waiting times between jumps away from state i, τ_i using the exponential distribution:

$$\Pr[\tau_i = t] = \lambda_i e^{-\lambda_i t} \tag{8.3}$$

Note that we can express this in terms of the general waiting time variable $\tau(t)$ conditional on being in state i:

[62] Equivalent to the real world probability measure $\mathbb{P}$
[63] The Markov Chain must be right-continuous with respect to t: $\Pr[\lim_{t\downarrow s} X(t) = X(s)] = 1$ for $\forall s < t$ and $q_{ii} < \infty$

$$\tau_i \equiv \tau(t) - t | X(t) = i \tag{8.4}$$

Knowing the distribution of this process allows us to price CoS securities:

$$\begin{aligned} CoS_t(X_t) &= \mathbf{E}_{\mathbb{Q}}[1. e^{-r(\tau(t)-t)} | \Im_t] \\ &= \mathbf{E}_{\mathbb{Q}}[e^{-r\tau_i} | \Im_t] \\ &= \left(1 - \frac{-r}{\lambda_i}\right)^{-1} \\ &= \frac{\lambda_i}{\lambda_i + r} \end{aligned} \tag{8.5}$$

Here we have used the general expression for a moment generating function of an exponential random variable to solve the expectation.

We also need to find the stock price process under the measure $\mathbb{Q}$. For this we need to find a measure such that the discounted stock price is a martingale. The most important condition is that the following equation is satisfied, $\forall\, T \geq t$:

$$\mathbf{E}_{\mathbb{Q}}\left[e^{-(r+k(X_t))(T-t)} S_T \frac{d\mathbb{Q}}{d\mathbb{P}} | \Im_t\right] = S_t \tag{8.6}$$

Here we recognize that the presence of the regime risk creates some additional risk premium $k(X_t)$ which is state dependant. However, the actual value of $k(X_t)$ is unimportant here. Note here that we have assumed that r is not state dependant. It is feasible that one can allow for r to vary according the latent process.

Guo X. (1999) then states that under the measure $\mathbb{Q}$, we have the following stock price dynamics:

$$\frac{dS_t}{S_t} = \left(r - d_{X_t}\right)dt + \sigma(X_t)dW_t \tag{8.7}$$

where d_{X_t} is a random quantity which alters the drift depending on the current state.

8.3. Option price under a CoS security-completed market (2 states)

We state here, without proof, the call option price, $O(C)(T-t, S_t, \sigma, r, X_t)$for a two-state continuous-time homogenous MSM with a constant risk free rate r and known initial state.

Due to the time homogeneity, only the time to expiry matters, $T-t$. Readers are referred to the appendix in Guo X. (1999) for details:

Definition: Vanilla Call Option Price under a market completed by CoS securities.

$$O(C)(T-t,S_t,\sigma,r,X_t) = \mathbf{E}_{\mathbb{Q}}[e^{-r(T-t)}(S_T-K)^+|X_t] \tag{8.8}$$

$$= e^{-r(T-t)}\int_0^{\infty}\int_0^{T-t} y\rho(\ln(y+k),m(b),v(b)f_i(b,T-t)dbdy \tag{8.9}$$

where $\rho(ln(y+k),m(b),v(b))$ is the value of a normal or Gaussian probability density function at point $ln(y+k)$ for a normal random variable with mean $m(b)$ and variance $v(b)$. i is the state that X_t occupies. This mean and variance are defined as follows:

$$m(b) = \left(d_1 - d_2 - \frac{1}{2(\sigma_2^2-\sigma_1^2)}\right)b + \left(r - d_1 - \frac{1}{2\sigma_1^2}\right)(T-t) \tag{8.10}$$

$$v(b) = \sigma_1^2(T-t) + (\sigma_2^2-\sigma_1^2)b \tag{8.11}$$

$f_i(b,T-t)$ is the total time spent between time b and $T-t$ during which $X_t = i$ if $X_t = i$ at time zero. Writing it as $f(t,T)$ to simply notation, it is defined as follows for $i = 1{,}2$:

$$f_1(t,T) = e^{-\lambda_1 T}e^{(\lambda_1-\lambda_2)t}\left(\left(\frac{T-t}{\lambda_1\lambda_2}\right)^{\frac{1}{2}} J_{-1}\left(2(-\lambda_2\lambda_1 Tt + \lambda_2\lambda_1 T^2)^{\frac{1}{2}}\right)\right) + \lambda_2 J_0\left(2(-\lambda_2\lambda_1 Tt + \lambda_2\lambda_1 T^2)^{\frac{1}{2}}\right) \tag{8.12}$$

$$f_2(t,T) = e^{-\lambda_2 T}e^{(\lambda_2-\lambda_1)t}\left(\left(\frac{T-t}{\lambda_1\lambda_2}\right)^{\frac{1}{2}} J_{-1}\left(2(-\lambda_2\lambda_1 Tt + \lambda_2\lambda_1 T^2)^{\frac{1}{2}}\right)\right) + \lambda_1 J_0\left(2(-\lambda_2\lambda_1 Tt + \lambda_2\lambda_1 T^2)^{\frac{1}{2}}\right) \tag{8.13}$$

$J_a(z)$, is a Bessel function. It is defined as:

$$J_a(z) = \left(\frac{1}{2z}\right)^a \sum_{n=0}^{\infty} \frac{(-1)^n \left({}^{z}/_{2}\right)^{2n}}{n!\,\Gamma(a+n+1)} \tag{8.14}$$

Since the argument in the gamma function is always an integer in the definitions above, we note that $\Gamma(n) = (n-1)!$

The formulation above for equations (8.9)-(8.14) appears complex. Equation (8.14) has a summation that sums to infinity; Equation (8.9) has no closed form expression implying that there is no neat way to compute the option price. One would need to approximate the double integral via integration approximation techniques, for example Simpsons rule.

Although we now have a preference free, objective option price, its theoretical underpinning and practical application is poor. As already discussed, CoS securities simply do not exist in the market. Even more importantly, we simply do not know what the value of the initial hidden state is as implied by these functions. The latter problem could be practically overcome by applying the approximation:

$$\pi_1 O(T-t, S_t, \sigma, r, X_t = 1) + \pi_2 O(T-t, S_t, \sigma, r, X_t = 2) \tag{8.15}$$

However, this is just a rough approximation since all proofs are based on the assumption that the initial state is known.

Despite its shortcomings, it is well known for its innovative way to create a unique objective option price. The next section looks at option prices in an incomplete market which is the case in reality.

9. Markov Modulated Option Prices

9.1. Introduction

In this Section we build on the theory developed in prior sections to present some option pricing methodology in incomplete markets. Firstly, we explore the case where we can use a special case discrete-time approximation to a continuous-time process to find option prices. In discrete-time we can again clarify that the market is incomplete. The discrete-time model that is presented could be used to price more exotic options such as American or Asian Options. We then explore the main models in continuous-time and we find that there are two important models that present themselves. One is when the additional “switching” risk is taken into account (or “priced”) in the model. The other is when switching risk is ignored or “not” priced. Theory and results are developed for each type.

9.2. Discrete-time Lattices

9.2.1. Introduction

As is the case with using Binomial and Trinomial trees to approximate the standard BS type options, we may apply a similar idea to price options where the underlying is governed by a regime switching process.

In the standard “Black Scholes” Binomial tree, the sample period is divided into a finite number of points. At the start node, the tree branches in two directions: “up” or “down” such that if there are n points then we have:

Definition: Binomial Stock Price Model/Tree

$$S_{n+1} = \begin{cases} S_n u, & if\ up\ jump \\ S_n d, & if\ down\ jump \end{cases} \tag{9.1}$$

If we consider the very simple case of one step binomial tree, calculating the value of an option, where the underlying process is binomial, amounts to finding the value of a replicating (and, for more than one time-step, self-financing) portfolio. We can set up a portfolio at time zero consisting of stocks and cash, $V_0 = \phi S_0 + \psi$; where S_0 represents the stock price at time 0, ϕ, the amount of stock and ψ, the amount of cash.

Assume that, under this one-step environment, an option exists with the payoff, O, at time 1 (the terminal time) as follows:

$$O_1 = \begin{cases} f_u, & \textit{if up jump} \\ f_d, & \textit{if down jump} \end{cases} \tag{9.2}$$

It's now possible to choose values for ϕ and ψ now such that portfolio V exactly replicates the option payoff at time 1, regardless of the route taken. Under a constant continuously compounding risk free rate r per period, we can solve the following equations simultaneously:

$$\phi S_0 u + \psi e^r = f_u \tag{9.3}$$

$$\phi S_0 d + \psi e^r = f_d \tag{9.4}$$

Since the portfolio exactly replicates the option payout, by a no-arbitrage argument, we can see that the value of a vanilla option at time 0, $O(C)_0$ must be the value of the portfolio today. This turns out to be:

$$O_0 = V_0 = e^{-r}\left[\left(\frac{e^r - d}{u - d}\right) f_u + \left(1 - \frac{e^r - d}{u - d}\right) f_d\right] \tag{9.5}$$

Extending this to cases where there are multiple steps, we simply start at the end time point and work our way backwards through the tree, rebalancing the portfolio at each time step through choice of ϕ and ψ until the start point is reached so that V_o (and hence O_0) is determined.

As is observed in many mathematical finance texts, the most startling observation is that the probabilities of an up or down move is irrelevant to the option price calculation. In fact it can be seen from (9.5) (and proved) that in general:

Definition: Binomial Tree Option Price

$$O_0 = e^{-rT}\mathbf{E}_{\mathbb{Q}}[Option\ Payoff\ at\ time\ T] \tag{9.6}$$

where $\mathbb{Q}$ is a probability measure (RNM or EMM) governing the binomial process under which the probability of an up jump is $(e^r - d)/(u - d)$

Cox, Ross, & Rubinstein (1979) show that by choosing appropriate values of u and d and letting the number of steps tend to infinity, the binomial process converges in distribution to a lognormal process and the option price approximates the BS price. Firstly this is done by ensuring that the first two moments of the processes match as the number of time steps approaches infinity and secondly showing that the binomial process converges to a lognormal distribution.

By parameterising $u = e^{\alpha}$ and $d = e^{-\alpha}$, we seek to find a value for α such that the mean and variance of the binomial process, under some probability measure, match that of the lognormal process. Here $dt = {}^{T}/_{n}$, where T is the length of the sample period and n, the number of nodes.

If we assume that when S is lognormal, then $logS_{t+dt} \sim N\left(logS_t + \left(\mu - \frac{1}{2}\sigma^2\right)dt, \sigma^2 dt\right)$ under an arbitrary measure $\wp$. Further, let p be the probability of an up jump. Then, we are required to solve:

$$pS_t e^{\alpha} + (1-p)S_t e^{-\alpha} = S_t e^{\mu dt} \quad (9.7)$$

$$p\alpha^2 + (1-p)(-\alpha)^2 - \mu^2 dt^2 = \sigma^2 dt \quad (9.8)$$

This yields:

$$p = \frac{e^{\mu dt} - e^{-\alpha}}{e^{\alpha} - e^{-\alpha}} \quad (9.9)$$

$$\alpha = \sqrt{\sigma^2 dt + \mu^2 dt^2} \quad (9.10)$$

Note that when dt is small, we could assume that $dt^2 \approx 0$ and we have the Cox, Ross and Rubenstein approximation with $u = e^{\sigma\sqrt{dt}}$.

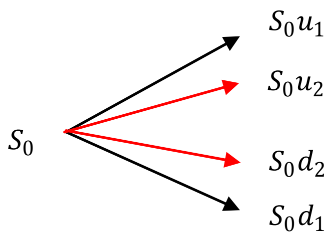
9.2.2. Making the step into a Markov-switching world

A key requirement in the above option pricing algorithm was the ability to choose a portfolio such that it replicates the option price no matter which route the process takes. Mathematically, this meant that we were able to find values of ϕ and ψ such that V was determinable (previsible) at the time step prior to the option payout.

If we go back to our one-step example, we could attempt to define a binomial model in a Markov-switching environment. Let's assume for the time being that there are only two states.

A model of the stock prices then becomes more complicated, since we have two values for u and d for each state: u_1, u_2, d_1, d_2. The stock price then takes four possible values at the next time step.

Figure 9.1: Quadrinomial Lattice (1)



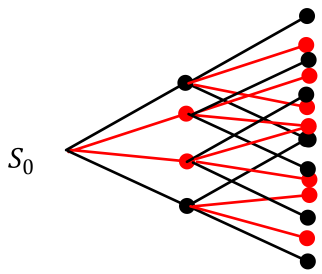
Therefore an option could have four different payouts at time 1. In the above figure we have a quadranomial lattice, with the inner two branches representing a low volatility regime and the outer two branches representing a high volatility regime. We could then ask if it's possible to set up a portfolio of stocks and cash such that it replicates the option. This would involve solving for two variables in four equations, which may have multiple solutions.

If there were other tradable securities in the market, then that would introduce other parameters and we could then create a hedging portfolio as discussed in the previous Section. In other words, with more sources of independent random variation, the more outcomes we could replicate.

With only stocks and cash available, the market is incomplete and no unique arbitrage price exists. Simply put, not all claims are replicable. We will have to choose another criterion/method to determine the option value. This statement is equivalent to choosing an appropriate probability measure for which to compute payoff expectations (i.e. an EMM)

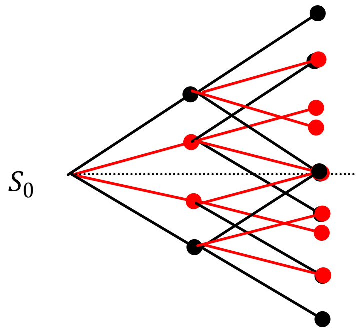
Furthermore, the branches of the lattice do not recombine well at all unless in general $u_1 - 1 = 2(u_2 - 1)$ and $u_i d_i = 1$. For example, in general a 2 step tree may resemble the following:

Figure 9.2: Quadrinomial Lattice (2)



This leads to 14 nodes at time 2. If we apply the restriction on u and d as above, then the tree combines more efficiently (the figure below is shown with $u_i d_i = 1$ for $i = 1,2$):

Figure 9.3: Recombinant Quadranomial Lattice



This reduces the number of nodes at each point to n^2.

9.2.3. Discretising the Hidden Markov Process

If the Hidden Markov process driving state determination is a continuous-time process, a fundamental requirement of pricing any options will be to determine (unconditional) probabilities of transitioning between states after each time step.

We would expect that the smaller the time step the greater the probability of not transitioning between states would be. Therefore, it is clear that the magnitude of these probabilities will rely on the size of the time step.

Calculating discrete transition probabilities based on a continuous-time process is discussed in Section 5. Returning to the two state example, we assume that over the entire sample period the probability of being in state 1 at the end (time T), given that the process was in state 1 at the start (time 0) is j_1^T. Similarly, for state 2, the probability is j_2^T.

Assume there is an intermediate time between time 0 and T, time $K = T/2$. Define similar probabilities for the states with the end time at time K and denote these as j_1^K and j_2^K. The following can then be deduced (with probabilities conditional on being at state 1 at the start):

$$j_1^T = \Pr[Being\ at\ state\ 1\ at\ time\ K\ and\ T] + \Pr[Being\ at\ state\ 2\ at\ time\ K] \times \Pr[Being\ at\ state\ 2\ at\ time\ K\ and\ State\ 1\ at\ time\ T] \quad (9.11)$$

We can then get the following set of equations:

$$j_1^T = (j_1^K)^2 + (1 - j_1^K)(1 - j_2^K) \quad (9.12)$$

$$j_2^T = (j_2^K)^2 + (1 - j_2^K)(1 - j_1^K) \quad (9.13)$$

The solution to these two equations produces two possibilities; the higher value is used which is in line with the thinking that the “staying” probabilities increase as the time step decreases.

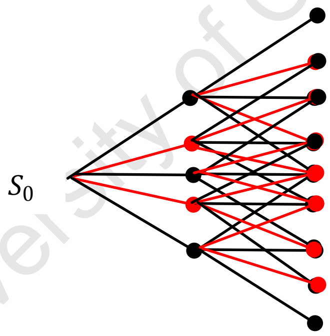
Thus by knowing the transition probabilities over the entire sample period we may calculate transitional probabilities for smaller time steps by subdividing the interval. In the special case where both probabilities are 0.5, equation (9.13) shows us that the smaller time step probabilities will remain at 0.5 and hence are independent of time.

9.2.4. Introduction to the Pentanomial Lattice

Bollen (1998) suggests a pentanomial lattice as opposed to a quadranomial lattice.

In Bollen's pentanomial lattice, each regime tree is represented by a trinomial lattice. Bollen refers to α as "the step size" of each tree. These are initially calculated to match the moments of the underlying conditional distributions. One step size is then adjusted such that one of the branches is shared between the two regimes. This is done such that the step size of one of the trees is increased such that there is 1:2 ratio between the step sizes. Conditional probabilities of the adjusted branch are then adjusted so that moment matching is still retained.

Figure 9.4: Pentanomial Lattice



In the above figure, the pentanomial lattice has $4n - 3$ nodes. This is a linear growth as opposed the exponential growth of the quadranomial lattice. This is a significant computational reduction.

9.2.5. Pentanomial Lattice construction (for 2 states)

Definition and Algorithm: Pentanomial Lattice

The return under each state $i = 1,2$ is assumed to be $\sim N(\mu_i dt, \sigma^2 dt)$.

We then follow the following algorithm for tree construction:

STEP 1 – Initial size and probability calculation:

For each state i, calculate p_i and α_i from equations (9.9) and (9.10) . Denote the greater of α_i as α_h and the lesser as α_l.

Denoting the conditional probabilities (given a regime i) of the regime's trinomial tree as the vector $(p_{i,d}, p_{i,m}, p_{i,u})$, initially set $p_{i,d} = (1 - p_i)$, $p_{i,m} = 0, p_{i,u} = p_i$ for each i.

STEP 2: - Re-proportioning step size and retaining moment matching

If $\alpha_h > 2\alpha_l$ then:

- *Set:*

$$p_{l,u} = \frac{e^{\mu_l dt} - e^{-\alpha_h/2} - p_{l,m}\left(1 - e^{-\alpha_h/2}\right)}{e^{\alpha_h/2} - e^{-\alpha_h/2}} \tag{9.14}$$

$$p_{l,m} = 1 - 4(\alpha_l/\alpha_h)^2 \tag{9.15}$$

$$p_{l,u} = 1 - p_{l,m} - p_{l,d} \tag{9.16}$$

- *Then set $\alpha_l = \alpha_h/2$.*

If $\alpha_h < 2\alpha_l$ then

- *Set:*

$$p_{h,u} = \frac{e^{\mu_h dt} - e^{-2\alpha_l} - p_{h,m}(1 - e^{-2\alpha_l})}{e^{2\alpha_l} - e^{-2\alpha_l}} \tag{9.17}$$

$$p_{h,m} = 1 - 0.25(\alpha_h/\alpha_l)^2 \tag{9.18}$$

$$p_{h,u} = 1 - p_{h,m} - p_{h,d} \tag{9.19}$$

- *Then Set $\alpha_h = 2\alpha_l$*

Performing the above steps at each node guarantees efficiently recombinant branches and fully defined conditional probabilities for each branch while retaining moment matching.

The methods in this section describe how state transition probabilities can be calculated for each time step and thus the pentanomial lattice is fully defined.

9.2.6. Pentanomial Lattice Option Valuation (for 2 states)

We know the market is incomplete and there is no unique no-arbitrage price. Another way to effectively skip over this issue is to make the very simplifying assumption that regime risk is not priced. The additional "regime risk" cannot be hedged away. However if, in the minds of market participants, regime switching risk is not "priced" then only pure stock volatility is priced.

In this case, we will apply standard risk neutral pricing, by assuming that the stock price drifts at the risk free rate and calculating discounted expected payoffs.

In the pentanomial lattice structure, there is usually multiple paths and regime switches that can take place to reach the same node. Consider a pentanomial branching process $S(t)$ and a two state underlying Markov chain $X(t)$, similar to as in Figure 9.4; i.e. with state 1 the "high volatility" state and state 2 the "low volatility" state. Then the vanilla call price, O_t at an arbitrary time t can be defined in terms of the call values at time $t + dt$ from the 5 nodes stemming from the current node:

Definition and Proof: Pentanomial Vanilla Option Price

$$
\begin{aligned}
O_t &= \mathbf{E}_{\mathbb{Q}}[e^{-r_i dt} O_{t+dt}] \\
&= j_1^t \mathbf{E}_{\mathbb{Q}}[e^{-r_1 dt} O_{t+dt} | X(t) = 1] + j_2^t \mathbf{E}_{\mathbb{Q}}[e^{-r_2 dt} O_{t+dt} | X(t) = 2] \\
&= j_1^t e^{-r_1 dt} [p_{1,u} O_{t+dt}(node\ 1) + p_{1,m} O_{t+dt}(node\ 3) \\
&\qquad + p_{1,d} O_{t+dt}(node\ 5)] + j_2^t e^{-r_2 dt} [p_{2,u} O_{t+dt}(node\ 2) \\
&\qquad + p_{2,m} O_{t+dt}(node\ 3) + p_{2,d} O_{t+dt}(node\ 4)]
\end{aligned} \tag{9.20}
$$

Since the call values at the terminal time is known for each terminal node, the option price can be pulled back recursively to time zero.

If we know the current regime at time zero, then we do not need to condition on the regime in the above equation.

The above lays out the framework for a method of option pricing in discrete time. It has some noticeable restrictions but preserves moment matching which is a good approximation to reality.

9.3. Continuous Time

9.3.1. Introduction

We turn our attention now to continuous time models. The technique used is the Regime-Switching Esscher transform. The basic idea behind Esscher Transforms was explored in Section 7.5

This Section largely outlines the papers by Elliot, Chan, & Kuen Sui (2005) and Kuen Siu & Yang (2009).

To recap, the risk-free rate, the drift and volatility are all stochastic and are denoted by r_t, μ_t and σ_t respectively. They are dependent on the underlying MSM X_t. Define a function $Z_t \equiv R(0,t)$ discussed in Section 3 as:

Definition: $\boldsymbol{Z_t}$

$$Z_t = \int_0^t \left(\mu_s - \frac{1}{2}\sigma_s^2\right) ds + \int_0^t \sigma_s dW_s \tag{9.21}$$

This implies that:

$$S_t = S_u e^{Z_t - Z_u} \tag{9.22}$$

for all $0 \leq u \leq t$.

We turn to some technical details in order to accurately define the Regime-Switching Esscher Transform (RSET):

Definition: Regime Switching Esscher Transform (RSET)

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be the complete probability space. Then let $\{\mathfrak{F}_t^X\}_{0 \leq t \leq T}$ and $\{\mathfrak{F}_t^Z\}_{0 \leq t \leq T}$ be the natural filtrations generated by the processes X_t and Z_t respectively. Then define a new σ-algebra, $\mathcal{G}_t \equiv \mathfrak{F}_t^X \vee \mathfrak{F}_t^Z$ for every t.

Define a RSET parameter, $\theta(t, X_t)$, which is stochastic and so is dependent on t and the current state. This defines a new probability measure $\mathbb{Q}^\theta$ which is equivalent to $\mathbb{P}$. For shorthand, we write $\theta(t, X_t) \equiv \theta_t$ and have:

$$\left.\frac{d\mathbb{Q}^\theta}{d\mathbb{P}}\right|_{\mathcal{G}_t} = \frac{\exp\left(\int_0^t \theta_s dZ_s\right)}{\mathbf{E}_{\mathbb{P}}[\exp(\int_0^t \theta_s dZ_s)|\mathfrak{F}_t^X]} \tag{9.23}$$

Note that since Z_t is normally distributed, the exponent of the exponential function in the argument in the expectation in the denominator is a classic Ito integral and its distribution is easily specified, $\int_0^t (\theta_s dZ_s)|\mathfrak{F}_t^X \sim N(\int_0^t \theta_s \left(\mu_s - \frac{1}{2}\sigma_s^2\right) ds, \int_0^t \theta_s^2 \sigma_s^2 ds)$. We can then use the properties of the moment generating function for normal distributions to rewrite the denominator:

$$\mathbf{E}_{\mathbb{P}}\left[\exp\left(\int_0^t \theta_s dZ_s\right)|\mathfrak{F}_t^X\right] = \exp\left[\int_0^t \theta_s \left(\mu_s - \frac{1}{2}\sigma_s^2\right) ds + \frac{1}{2}\int_0^t \theta_s^2 \sigma_s^2 ds\right] \tag{9.24}$$

We can therefore rewrite (9.23) as follows:

$$\begin{aligned}
\left.\frac{d\mathbb{Q}^\theta}{d\mathbb{P}}\right|_{\mathcal{G}_t} &= \frac{\exp\left(\int_0^t \theta_s dZ_s\right)}{\mathbf{E}_{\mathbb{P}}[\exp(\int_0^t \theta_s dZ_s)|\mathfrak{F}_t^X]} \\
&= \frac{\exp\left(\int_0^t \theta_s dZ_s\right)}{\exp\left[\int_0^t \theta_s \left(\mu_s - \frac{1}{2}\sigma_s^2\right) ds + \frac{1}{2}\int_0^t \theta_s^2 \sigma_s^2 ds\right]} \\
&= \exp\left[\int_0^t \theta_s dZ_s - \int_0^t \theta_s \left(\mu_s - \frac{1}{2}\sigma_s^2\right) ds - \frac{1}{2}\int_0^t \theta_s^2 \sigma_s^2 ds\right] \\
&= \exp\left(\int_0^t \theta_s \sigma_s dW_s - \frac{1}{2}\int_0^t \theta_s^2 \sigma_s^2 ds\right)
\end{aligned} \tag{9.25}$$

We now need to create an EMM. We can make a simplifying assumption as we did with the trinomial trees in the previous section. We can assume that the regime risk is not priced which is equivalent to the statement that the market knows the current state of the Markov Chain. Essentially the hidden process becomes just an observable Markov Process.

We now need to find an appropriate set of RSET parameters such that the EMM is satisfied. Call this set, $\{\tilde{\theta}_s\}_{0 \le t \le T}$ and the associated measure $\mathbb{Q}^{\tilde{\theta}}$.

$$S_0 = \mathbf{E}_{\mathbb{Q}^{\tilde{\theta}}}\left[\exp\left(-\int_0^t r_s ds\right) S_t \middle| \mathfrak{F}_t^X\right] \tag{9.26}$$

The filtration $\mathfrak{F}_t^X$ clarifies knowledge of the Markov Chain. Define an intermediate function $b_t = u_t - \frac{1}{2}\sigma_s^2$. We now use (9.25) and (9.26) to rewrite the martingale condition in terms of the real world measure.

$$\begin{aligned}
S_0 &= \mathbf{E}_{\mathbb{Q}^{\tilde{\theta}}}\left[\exp\left(-\int_0^t r_s ds\right) S_t \middle| \mathfrak{F}_t^X\right] \\
\Rightarrow \quad 1 &= \mathbf{E}_{\mathbb{P}}\left[\frac{\exp\left(-\int_0^t r_s ds\right) S_t}{S_0} \left.\frac{d\mathbb{Q}^\theta}{d\mathbb{P}}\right|_{\mathcal{G}_t} \middle| \mathfrak{F}_t^X\right] \\
\Rightarrow \quad 1 &= \mathbf{E}_{\mathbb{P}}\left[\exp\left(\int_0^t b_s - r_s ds + \int_0^t \sigma_s dW_s + \int_0^t \tilde{\theta}_s \sigma_s dW_s - \frac{1}{2}\int_0^t \tilde{\theta}_s^2 \sigma_s^2 ds\right) \middle| \mathfrak{F}_t^X\right]
\end{aligned}$$

$$\Rightarrow \qquad 1 = \mathbf{E}_{\mathbb{P}}\left[\exp\left(\int_0^t \left(b_s - r_s - \frac{1}{2}\tilde{\theta}_s^2\sigma_s^2\right)ds + \int_0^t (\tilde{\theta}_s + 1)\sigma_s dW_s\right)\middle|\mathfrak{F}_t^X\right] \qquad (9.27)$$

The exponent in the expectation is normally distributed by virtue of the normality of the Ito integral. We can again make use of the moment generated function to rewrite the expectation. Continuing from above, we have:

$$\Rightarrow \qquad 1 = \exp\left(\int_0^t \left(b_s - r_s - \frac{1}{2}\tilde{\theta}_s^2\sigma_s^2\right)ds + \frac{1}{2}\int_0^t (\tilde{\theta}_s + 1)^2\sigma_s^2 ds\right)$$

$$\Rightarrow \qquad 1 = \exp\left(\int_0^t b_s - r_s + \frac{1}{2}(2\tilde{\theta}_s + 1)\sigma_s^2 ds\right)$$

$$\Rightarrow \qquad 0 = \int_0^t b_s - r_s + \frac{1}{2}(2\tilde{\theta}_s + 1)\sigma_s^2 ds$$

$$\Rightarrow \qquad 0 = b_t - r_t + \frac{1}{2}(2\tilde{\theta}_t + 1)\sigma_t^2$$

$$\Rightarrow \qquad \tilde{\theta}_t = \frac{r_t - \mu_t}{\sigma_t^2} \qquad (9.28)$$

This takes a very familiar form to the market price of risk which is generally defined as $\mu_t - r_t/\sigma_t$.

9.3.2. The Non-Priced Regime Risk Option Formula

By knowing the RSET that uniquely determines $\mathbb{Q}^{\tilde{\theta}}$, we can write down the option pricing formula. We need some more technicalities here: knowing the entire path of the Markov Chain (i.e. up to time T) but knowing only the current value of Z_t will allow us to write down an option pricing formula. For this we need an appropriate σ-algebra which is double-indexed. This is $\mathcal{G}_{T,t} \equiv \mathfrak{F}_t^X \vee \mathfrak{F}_T^Z$, for all $0 \le t \le T$.

Since we now have an EMM, we can write down the vanilla call option price at time t

$$O(C)(t,T,S_t,P_{t,T},U_{t,T},K) = \mathbf{E}_{\mathbb{Q}^{\tilde{\theta}}}\left[\exp\left(-\int_t^T r_s ds\right)(S_T - K)^+\middle|\mathcal{G}_{T,t}\right]$$

$$\Leftrightarrow \qquad O(C)(t,T,S_t,P_{t,T},U_{t,T},K) = \mathbf{E}_{\mathbb{Q}^{\tilde{\theta}}}\left[\exp\left(-\int_t^T r_s ds\right)(e^{Z_T - Z_t} - K)^+\middle|\mathcal{G}_{T,t}\right] \qquad (9.29)$$

Here, K is the strike price, $P_{t,T} = \int_t^T r_s(X_t)ds$ and $U_{t,T} = \int_t^T \sigma^2(X_t)ds$. In the expression, I use the argument X_t to highlight the parameters dependency on the latent process. I use $P_{t,T}$ and $U_{t,T}$ as arguments in the option price, as we shall see that Z_T is a function of $P_{t,T}$ and $U_{t,T}$ and that the distribution of Z_T is easily determinable once we have knowledge of $P_{t,T}$ and $U_{t,T}$. This is of course a mathematical convenience, since we do not know anything about the latent process and so $P_{t,T}$ and $U_{t,T}$ are also random variables and hence so is the option price.

We need to take a second expectation of this price with respect to the distributions of $P_{t,T}$ and $U_{t,T}$. Before we do this, we first need Z_t given $\mathcal{G}_{T,t}$ under $\mathbb{Q}_{\tilde{\theta}}$. This is achieved via a simple Girsanov transformation since we know the Radon-Nikodym derivative of $\mathbb{Q}_{\tilde{\theta}}$ with respect to $\mathbb{P}$. By (9.25) we have:

$$\left.\frac{d\mathbb{Q}^{\theta}}{d\mathbb{P}}\right|_{\mathcal{G}_t} = \exp\left(\int_0^t \theta_s\sigma_s dW_s - \frac{1}{2}\int_0^t \theta_s^2\sigma_s^2 ds\right)$$

$$\Rightarrow \quad \left.\frac{d\mathbb{Q}^{\tilde{\theta}}}{d\mathbb{P}}\right|_{\mathcal{G}_t} = \exp\left(\int_0^t \frac{r_s-\mu_t}{\sigma_s^2}\sigma_s dW_s - \frac{1}{2}\int_0^t \left(\frac{r_s-\mu_s}{\sigma_s^2}\right)^2 \sigma_s^2 ds\right)$$

$$\Rightarrow \quad \left.\frac{d\mathbb{Q}^{\tilde{\theta}}}{d\mathbb{P}}\right|_{\mathcal{G}_t} = \exp\left(\int_0^t \frac{r_s-\mu_t}{\sigma_s} dW_s - \frac{1}{2}\int_0^t \left(\frac{r_s-\mu_s}{\sigma_s}\right)^2 ds\right) \tag{9.30}$$

With reference to Girsanov's theorem[64], the expression represents $\partial(t)$ with $q(t) = {(r_s - \mu_s)}/{\sigma_s}$.

So that where we previous had the following under measure $\mathbb{P}$,

$$dS_t = \mu_t S_t dt + \sigma_t S_t dW_t \tag{9.31}$$

We now have the following under measure $\mathbb{Q}_{\tilde{\theta}}$:

$$dS_t = \mu_t S_t dt + \sigma_t S_t d\left(W^{\tilde{\theta}} + \frac{r_t-\mu_t}{\sigma_t}dt\right)$$

$$= r_t S_t dt + \sigma_t S_t dW^{\tilde{\theta}} \tag{9.32}$$

This is the standard lognormal form to give us the EMM condition. So we are able to determine the distribution of $Z_T|\mathcal{G}_{T,t}$ simply using standard SDE techniques. This turns out to be: $Z_T|\mathcal{G}_{T,t} \sim N\left(\int_t^T \left(r_s - \frac{1}{2}\sigma_s^2\right)ds, \int_t^T \sigma_s^2 ds\right)$. Or better expressed in terms of our new functions, we have $Z_T|\mathcal{G}_{T,t} \sim N\left(P_{t,T} - \frac{1}{2}U_{t,T}, U_{t,T}\right)$. We can now re-express equation (9.29):

$$O(C)(t,T,S_t,P_{t,T},U_{t,T},K) = \mathbf{E}_{\mathbb{Q}^{\tilde{\theta}}}\left[\exp\left(-\int_t^T r_s ds\right)(e^{Z_T - Z_t} - K)^+ \middle| \mathcal{G}_{T,t}\right]$$

$$= \mathbf{E}_{\mathbb{Q}^{\tilde{\theta}}}\left[\exp\left(-P_{t,T}\right)\left(S_t e^{P_{t,T} - \frac{1}{2}U_{t,T} + \sqrt{U_{t,T}}\,W_t^{\tilde{\theta}}} - K\right)^+ \middle| \mathcal{G}_{T,t}\right] \tag{9.33}$$

[64] Refer to Section 2

The above has a startling resemblance to equation (2.13) Section 2. In fact, if $r_t \equiv r$ and $\sigma_t \equiv \sigma$ as is the case with the standard BS world, then $r(T-t) \equiv P_{t,T}$ and $\sigma\sqrt{(T-t)} \equiv U_{t,T}$. We can then proceed with the integration steps and produce a similar result.

Definition: Vanilla Call Option price when the regime risk is not priced

$$O(C)\big(t,T,S_t,P_{t,T},U_{t,T},K\big) = S_t N(d_1) - K\exp\big(-P_{t,T}\big)N(d_2) \tag{9.34}$$

where

$$d_1 = \big(U_{t,T}\big)^{-\frac{1}{2}}\left(\ln\left(\frac{S_t}{K}\right) + P_{t,T} + \frac{1}{2}U_{t,T}\right) \tag{9.35}$$

$$d_2 = d_1 - \big(U_{t,T}\big)^{\frac{1}{2}} \tag{9.36}$$

The final step is to find the distributions of $P_{t,T}$ and $U_{t,T}$. For this Elliot, Chan, & Kuen Sui (2005) propose a convenient alternative of expressing this quantities: Define $J_k(t,T)$ to be the total amount of time spent in state k between time t and T. $J_k(t,T)$ can thus be thought of as an total occupational time. Note that this is not a "waiting time". In other words, we are not measuring the time in a state until the next jump because later the latent process may return to the same state. Further, $\sum_{k=1}^{n} J_k(t,T) = T - t$.

Under this definition we can express $P_{t,T}$ and $U_{t,T}$ differently:

$$P_{t,T} = \int_t^T r_s ds = \sum_{k=1}^{n} r(k) J_k(t,T) \tag{9.37}$$

$$U_{t,T} = \int_t^T \sigma_s^2 ds = \sum_{k=1}^{n} \sigma^2(k) J_k(t,T) \tag{9.38}$$

Each of the above expressions contain n quantities of the occupational time and as such we need to find the joint distribution of all these occupational times. The derivation is complex and for details the reader is referred to Sericola (2000).

I now state the characteristic function of occupational times for Markov Chains without proof. The characteristic function is known since we know the parameters of rate-matrix $\boldsymbol{Q}$.

Definition: Characteristic Function of $\boldsymbol{J}(t,T)$

Define the dummy variables $\boldsymbol{C} = (C_1, C_2 \dots . C_n)$, a vector and a matrix $\boldsymbol{D}$ where matrix $\boldsymbol{D}$ is of dimension n with the entries of $\boldsymbol{C}$ on its diagonal and zeros everywhere else. Then the characteristic function, $\phi_{\boldsymbol{J}(t,T)}(\boldsymbol{C})$ of $\boldsymbol{J}(t,T) \equiv (J_1(t,T), J_2(t,T) \dots, J_n(t,T))$ is defined as follows:

$$\phi_{\boldsymbol{J}(t,T)}(\boldsymbol{C}) = \mathbf{E}[\exp(i < \boldsymbol{C},\boldsymbol{J}(t,T) >)|\mathfrak{F}_t^Z]$$

$$= < \exp[(\boldsymbol{Q}+i\boldsymbol{D})(T-t)]\, X_t, \boldsymbol{I} > \quad (9.39)$$

Here, $\boldsymbol{I}$ is not the identity matrix but a row vector of length n with each element equal to 1; X_t is a n-length vector corresponding to the hidden process's state at time t with a 1 at the state that the hidden process is currently at an zeros elsewhere.

Due to the one-to-one nature of a characteristic function and its cumulative distribution function, we can completely determine the distribution of $\boldsymbol{J}(t,T)$ from just the characteristic function. This is done via an Inverse Fourier Transform. The density function of a univariate random variable can be recovered via the inverse Fourier transform of the conjugate characteristic function: $f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{itx}\overline{\phi_X(t)}dt$.

Many computer packages have efficient built in Fourier and inverse Fourier transforms which allows simpler calculation. Call this multivariate density function $f\big(J_1(t,T), J_2(t,T), \dots, J_n(t,T)\big)$. We can now define the new option price based on a second expectation of equation (9.34) with respect to the joint distribution of the occupational times:

$$O(C)(t,T,S_t,K) = \int_0^T\int_0^T \dots \int_0^T O\big(t,T,S_t,P_{t,T},U_{t,T},K\big)\, f\big(J_1(t,T),J_2(t,T),\dots,J_n(t,T)\big)dJ_1 dJ_2 \dots dJ_n \quad (9.40)$$

There is unfortunately no closed form solution to the above and approximations need to be made such as the use of Simpson's rule to approximate each integral. This would be a painstaking process.

However, it is perhaps better to rather simulate $\boldsymbol{J}(t,T)$ via Monte Carlo Methods. Simulation of total occupation times to has already been dealt with in Section 10.3. It is merely a logical extension of the theory of simulating waiting times to state jumps.

To determine $O(C)(t,T,S_t,K)$ via Monte Carlo we need to simulate possible multiple values of $O(C)\big(t,T,S_t,P_{t,T},U_{t,T},K\big)$ and then average our results to determine its expectation (remembering of course that this function is a random variable dependant on $\boldsymbol{J}(t,T)$). We do this via the following algorithm:

Algorithm: Vanilla call option price when the regime risk is not priced.

0) Start by setting a dummy variable, say $sum = 0$. In each iteration, we will add the result to this variable.

Iterate the following algorithm exactly m times, where m is the number of simulations performed[65]:

a) Generate $\boldsymbol{J}(t,T) \equiv (J_1(t,T), J_2(t,T), \dots, J_n(t,T))$ via the methods in Section 10.3
b) Determine $P_{t,T} = \sum_{k=1}^{n} r(k) J_k(t,T)$ and $U_{t,T} = \sum_{k=1}^{n} \sigma^2(k) J_k(t,T)$
c) Determine $O\left(t,T,S_t,P_{t,T},U_{t,T},K\right)$ via equations (9.34), (9.35) and (9.36).
d) Add the result to the dummy variable sum

1) Set $O(t,T,S_t,K) = {}^{sum}/_{m}$

Numerical experiments using the above algorithm are carried out in Section 11. The next section explores the case when the regime risk is priced.

9.3.3. The Priced Regime Risk Option Formula

This section is largely based on the work of Kuen Siu & Yang (2009) who has adopted a modified approach to the one proposed in the section above. It attempts to assume the regime-switching risk is priced by the market. We would thus expect that the resultant call option price should be slightly higher the call price above.

A modified RSET is proposed in an attempt to price the regime-switching risk. Kuen Siu & Yang (2009) defined a RSET as follows:

Definition: Regime Switching Esscher Transform when the regime-switching risk is priced.

$$\frac{d\mathbb{Q}^{\theta}}{d\mathbb{P}} = \frac{\exp\left(\int_0^T \theta_s dZ_s\right)}{\mathbf{E}_{\mathbb{P}}[\exp(\int_0^T \theta_s dZ_s)|X_0]} = \Lambda_T \qquad (9.41)$$

65 m will usually be a very high number, possibly of the order of 10,000 or 20,000. The higher the number, the lower the variability in the answer and the number should be chosen such that the final answer has an acceptable level of variability. For example, it may be acceptable to choose a number such that the first three significant figures stay constant. The choice of m will also depend on the variance of the random variable(s) to be simulated. Lower variance random variables require less simulations to achieve answer stability than those with a higher variance.

The above is only slightly different from equation (9.23) in that the denominator holds the unconditional expectation rather than one conditioned on $\mathfrak{F}_t^X$. Not conditioning on $\mathfrak{F}_t^X$, is equivalent to saying that the regime risk ***is*** priced.

By defining $\Lambda_t = \mathbf{E}_{\mathbb{P}}[\Lambda_T|\mathcal{G}_t]$ we have that Λ_t is a $\mathbb{P}$-martingale by the tower rule. Kuen Siu & Yang (2009) then show that we can rewrite Λ_t in terms of a vector of rate parameters $\{\lambda_i\}_{0\leq i\leq n}$ which depend on the RSET parameters θ_t defined in the above equation. I will state his theorem without proof[66]:

Definition: $\boldsymbol{\lambda_i(\theta_i)}$

Let

$$\lambda_i(\theta_i) = \theta_i\mu_i - \frac{1}{2}\theta_i\sigma_i^2 + \frac{1}{2}\theta_i^2\sigma_i^2 \tag{9.42}$$

for $i = 1,2\ldots,n$

and $\boldsymbol{\lambda}(\boldsymbol{\theta}) \equiv \big(\lambda_1(\theta_1), \lambda_2(\theta_2), \ldots, \lambda_n(\theta_n)\big) \in \mathbb{R}^n$. *Then*

$$\Lambda_t = \exp\left(\int_0^T \theta_s dZ_s\right)\frac{\langle e^{\boldsymbol{Q}+\mathrm{diag}\,(\boldsymbol{\lambda}(\boldsymbol{\theta}))(T-t)}X_t, \mathbf{1}_n\rangle}{\langle e^{\boldsymbol{Q}+\mathrm{diag}\,(\boldsymbol{\lambda}(\boldsymbol{\theta}))T}X_0, \mathbf{1}_n\rangle} \tag{9.43}$$

wheret <·> is the dot product function[67], diag is the diagonal matrix function[68] and $\mathbf{1_n}$ *is a vector containing* n *ones.*

The above equations allow us to represent the Radon-Nikodym derivative, Λ_T explicitly:

$$\frac{d\mathbb{Q}^\theta}{d\mathbb{P}} = \Lambda_T = \exp\left(\int_0^T \theta_s dZ_s\right)\frac{1}{\langle e^{\boldsymbol{Q}+\mathrm{diag}\,(\boldsymbol{\lambda}(\boldsymbol{\theta}))T}X_0, \mathbf{1}_n\rangle} \tag{9.44}$$

Next comes determination of $\boldsymbol{\theta}$ and its associated function $\boldsymbol{\lambda}$ such that $\mathbb{Q}^\theta$ is a RNM. Call these parameters $\widetilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\lambda}}$ respectively. Here it is necessary to allow the discounted stock price to be a $\mathbb{Q}^\theta$-martingale on the enlarged filtration $\mathcal{G}_t$ - so that the stock price is a martingale if both the current price and the current underlying state is known.

I again state a lemma which enables determination of $\widetilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\lambda}}$ without proof, which involves a simple application of Bayes Rule[69].

[66] Please refer to the original paper for the full proof.
[67] The dot product function is where two vectors components are multiplied and then summed. I.e. $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} a_i b_i$.
[68] The diagonal matrix function is one whose argument is an n-element vector and whose result is a $n \times n$ square matrix with the vector down its main diagonal and zeros elsewhere.
[69] Once again, for further information please consult the original text.

Lemma: $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{i}}(\boldsymbol{\theta}_{\boldsymbol{i}})$

Let

$$\tilde{\lambda}_i(\theta_i) = -r_i + (\theta_i + 1)\mu_i - \frac{1}{2}(\theta_i + 1)\sigma_i^2 + \frac{1}{2}(\theta_i + 1)^2\sigma_i^2 \tag{9.45}$$

and $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \equiv \left(\tilde{\lambda}_1(\theta_1), \tilde{\lambda}_2(\theta_2), \dots, \tilde{\lambda}_n(\theta_n)\right) \in \mathbb{R}^n$*. Then, the martingale condition is satisfied if and only if:*

$$\langle e^{\boldsymbol{Q}+\mathrm{diag}\left(\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right)(t-u)} X(u), \mathbf{1}_n\rangle - \langle e^{\boldsymbol{Q}+\mathrm{diag}\left(\boldsymbol{\lambda}(\boldsymbol{\theta})\right)(t-u)} X(u), \mathbf{1}_n\rangle = 0 \tag{9.46}$$

for all values of $X(u)$ *with* $u \le t \le T$*.*

Note that since our latent process is time homogenous, we are able to simplify the above equation. Furthermore, there are only n possible values for $X(u)$ corresponding to the number of states of the latent process. Before we can restate the equation simply, recall the vector $\boldsymbol{e_i}$, a vector with a 1 in the i'th element and zeros elsewhere. With this in mind, we can restate the equation:

$$\langle e^{\left(\boldsymbol{Q}+\mathrm{diag}\left(\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right)\right)\boldsymbol{T}} \boldsymbol{e_i}, \mathbf{1}_n\rangle - \langle e^{\left(\boldsymbol{Q}+\mathrm{diag}\left(\boldsymbol{\lambda}(\boldsymbol{\theta})\right)\right)\boldsymbol{T}} \boldsymbol{e_i}, \mathbf{1}_n\rangle = 0 \tag{9.47}$$

for all $\mathrm{i} = 1,2,\dots,\mathrm{n}$

The solution to these equations gives rise to the values for $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ and hence $\boldsymbol{\theta}$**.** One issue we've overlooked around is computation of the matrix exponential. For a matrix $\boldsymbol{A}$, we have $e^{\boldsymbol{A}} = \sum_{i=1}^{\infty} \frac{\boldsymbol{A}^i}{i!}$, that is the exponential can be written as its Maclaurin series. In practical terms, we are obviously prevented in summing to ∞ and hence need come up with a suitable maximum limit. It is interesting to note that when we only use the linear approximation, a simpler result emerges:

$$\begin{aligned}
\langle e^{\left(\boldsymbol{Q}+\mathrm{diag}\left(\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right)\right)\boldsymbol{T}} \boldsymbol{e_i}, \mathbf{1}_n\rangle &\approx \left\langle \left(\boldsymbol{I} + \begin{pmatrix} q_{11} + \tilde{\lambda}_1 & q_{12} & \cdots & q_{1n} \\ q_{21} & \ddots & & \vdots \\ \vdots & & \ddots & q_{(n-1)n} \\ q_{n1} & \cdots & q_{n(n-1)} & q_{nn} + \tilde{\lambda}_n \end{pmatrix} T \right) \boldsymbol{e}_i, \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\rangle \\
&= \left\langle \begin{pmatrix} q_{1i}T \\ \vdots \\ 1 + (q_{ii} + \tilde{\lambda}_i) \\ \vdots \\ q_{ni}T \end{pmatrix}, \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\rangle \\
&= 1 + \tilde{\lambda}_i T + T \sum_{k=1}^{n} q_{ki}
\end{aligned} \tag{9.48}$$

Therefore we have that:

$$\langle e^{\left(\boldsymbol{Q}+\mathrm{diag}\left(\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right)\right)\boldsymbol{T}}\boldsymbol{e_i}, \mathbf{1}_n\rangle - \langle e^{\left(\boldsymbol{Q}+\mathrm{diag}\left(\boldsymbol{\lambda}(\boldsymbol{\theta})\right)\right)\boldsymbol{T}}\boldsymbol{e_i}, \mathbf{1}_n\rangle \approx (\tilde{\lambda}_i - \lambda_i)T \tag{9.49}$$

Now writing we can write this in terms of θ only using equations (9.42) and (9.45) :

$$(\tilde{\lambda}_i - \lambda_i)T \quad = \left(\mu_i - r_i + \sigma_i^2\theta_i\right)T \tag{9.50}$$

Therefore the solution to equation (9.47) is:

$$\theta_i = \frac{r_i - \mu_i}{\sigma_i^2} \tag{9.51}$$

This result is the same as per equation (9.28) under which the regime risk is not priced. This is surprising as there seems to be no intuitive reason for this, the authors note that this is the case but do not expand as to why these results are the same. One could hazard a guess by proposing that the simple linear approximation of the matrix exponential simplifies the martingale equation too much since the pricing of regime switching risk is of a higher order.

This notion has not been explored or tested since its result is not relevant to the argument. A mathematical coincidence or based on some solid reasoning, the reasoning for the equivalence of the subject of this thesis.

The introduction of a higher order exponential matrix approximation introduces some further problems, namely that multiple solutions begin to present themselves. Even when we use the quadratic approximation, the authors show that we can have at most three possible sets of values for $\boldsymbol{\theta}$**.** In other words we may have m sets of solutions to equation (9.47). We can therefore define $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(m)}\}$ as the set of all possible sets of solutions to equation (9.47). This highlights that even after choosing the Esscher transform as a Risk Neutral measure, we still have multiple measures that we can choose from.

In order to choose the appropriate set of values for $\boldsymbol{\theta}$ the authors choose that set that minimizes the maximum entropy between Esscher transform measure and the real-world measure. This was discussed in Section 7.4. Recall that the minimum entropy measure is that measure $\mathbb{Q}$ such that $I(\mathbb{Q}, \mathbb{P}) = \mathbf{E}_{\mathbb{P}}[\frac{d\mathbb{Q}}{d\mathbb{P}} \ln\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)]$ is at a minimum. In this case we begin by restricting ourselves to the set of measures $\mathbb{Q}^{\theta}$ where the values of $\boldsymbol{\theta}$ arise from the solution to equation (9.47). We seek to find the "conditional" minimum entropy of each set, i.e. the minimum entropy given the initial state only, call this $I\left(\mathbb{Q}^{\theta}, \mathbb{P}|X_0\right)$

The author shows that $I\left(\mathbb{Q}^{\theta}, \mathbb{P}|X_0\right)$ can be written in a similar exponential and dot product form by defining an intermediate function $\lambda_z(\theta_i)$ which is the derivative of $\lambda(z\theta_i)$ with respect to z calculated as:

$$\lambda_z(\theta_i) = \theta_i\mu_i - \frac{1}{2}\theta_i\sigma_i^2 + \theta_i^2\sigma_i^2 \tag{9.52}$$

The author then shows that:

Definition: Relative/Conditional entropy in terms of $\lambda(z\boldsymbol{\theta}_i)$ and $\lambda_z(\boldsymbol{\theta}_i)$

$$I\left(\mathbb{Q}^{\theta},\mathbb{P}|X_0\right)=\frac{\langle e^{\left(\boldsymbol{Q}+\text{diag}\,(\boldsymbol{\lambda_z}(\boldsymbol{\theta}))\right)^{\boldsymbol{T}}}X_0,\mathbf{1}_n\rangle}{\langle e^{\left(\boldsymbol{Q}+\text{diag}\,(\boldsymbol{\lambda}(\boldsymbol{\theta}))\right)^{\boldsymbol{T}}}X_0,\mathbf{1}_n\rangle}-\ln\langle e^{\left(\boldsymbol{Q}+\text{diag}\,(\boldsymbol{\lambda}(\boldsymbol{\theta}))\right)^{\boldsymbol{T}}}X_0,\mathbf{1}_n\rangle \tag{9.53}$$

We may know the initial value of the hidden process, X_0, in which case we just need to find the set of parameters, $\widehat{\boldsymbol{\theta}}$, that minimises the conditional entropy. $\widehat{\boldsymbol{\theta}}$ arises from the set

$$\widehat{\boldsymbol{\theta}}=\min_{\boldsymbol{\theta}\in\Theta}I\left(\mathbb{Q}^{\theta},\mathbb{P}|X_0\right) \tag{9.54}$$

On the other hand, if we do not know the initial state, then we first determine the maximum of conditional entropy in equation (9.53) over all possible initial states:

$$I\left(\mathbb{Q}^{\theta},\mathbb{P}\right)=\max_{i=1,2,\dots,n}I\left(\mathbb{Q}^{\theta},\mathbb{P}|X_0=e_i\right) \tag{9.55}$$

We then seek to find $\widehat{\boldsymbol{\theta}}$ similarly to above:

$$\widehat{\boldsymbol{\theta}}=\min_{\boldsymbol{\theta}\in\Theta}I\left(\mathbb{Q}^{\theta},\mathbb{P}\right) \tag{9.56}$$

Computation of $\widehat{\boldsymbol{\theta}}$ now allows us to write the option pricing formula out:

$$\begin{aligned} O(C)(t,T,S_t,K,X_t) &= \mathbf{E}_{\mathbb{Q}^{\hat{\theta}}}\left[\exp\left(-\int_t^T r_u\,du\right)(S_T-K)^+\middle|S_t,X_t\right] \\ &= \mathbf{E}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\exp\left(-\int_t^T r_u\,du\right)(S_T-K)^+\middle|S_t,X_t\right] \\ &= \frac{\mathbf{E}\left[\exp\left(\int_t^T\hat{\theta}_u dZ_u\right)\exp\left(-\int_t^T r_u\,du\right)(S_T-K)^+\middle|S_t,X_t\right]}{\mathbf{E}\left[\exp\left(\int_t^T\hat{\theta}_u dZ_u\right)\middle|X_t\right]} \end{aligned} \tag{9.57}$$

This can again be expressed in terms of occupational times by writing $\int_t^T r_u du$ as $P_{t,T}$ as per equation (9.37). The question arises how to write $\int_t^T \theta_u dZ_u$ in terms of total occupational times. As per Section 10.3, recall that T_k^i are the waiting times for the k'th jump away from state i so that $J_i(t,T)=\sum_k T_k^i$. Firstly, we note that for successive periods between jumps, θ_i is constant. Furthermore that the stock movement increment, Z, are independent for each increment where the latent process is in the same state. Hence if we same all increments for which the latent process is in state i then it will have an identical distribution to $Z_{J_{i(t,T)}}|\{X_s=i\,\forall t\leq s\leq T\}$. If we define $Y_t(i)$ as:

$$Y_t(i) = Z_s|\{X_s = i \ \forall t\} \tag{9.58}$$

then it is clear that $\sum_{i=1}^{n} Y_{J_i(t,T)}(i)$ is identically distributed to $Z_T - Z_t$ for all $t \leq T$

We could therefore express $\int_t^T \theta_u dZ_u$ as follows:

$$\int_t^T \theta_u dY_u \quad = \sum_{i=1}^{n} \hat{\theta}_i \, Y_{J_i(t,T)}(i) \tag{9.59}$$

Our option equation is then:

Definition: Vanilla Call Option price when the regime risk is priced.

$$O(C)(t,T,S_t,K,X_t) = \frac{\mathbf{E}\left[\exp\left(\sum_{i=1}^{n} \hat{\theta}_i \, Y_{J_i(t,T)}(i)\right)\exp\left(-P_{t,T}\right)\left(S_t e^{\sum_{i=1}^{n} Y_{J_i(t,T)}(i)} - K\right)^+ \middle| S_t, X_t\right]}{\mathbf{E}\left[\exp\left(\sum_{i=1}^{n} \hat{\theta}_i \, Y_{J_i(t,T)}(i)\right)\middle| X_t\right]} \tag{9.60}$$

We can now write an algorithm out to determine $O(C)(t,T,S_t,K,X_t)$ using similar techniques as in the option price determination when the regime-switching risk is not priced. The difference here is that we need to evaluate each expectation (numerator and denominator) separately via Monte Carlo processes. This is a two stage process, firstly to determine the optimal parameter vector $\boldsymbol{\theta}$ and then to carry out Monte-Carlo Simulation to evaluate the expression in equation (9.60).

Algorithm: Vanilla Call Option price when the regime risk is priced

<u>Stage 1</u>

1) *Using the inputs $Q, \boldsymbol{\mu}, \boldsymbol{\sigma}, (T-t)$ solve equation (9.47) solve for values of $\boldsymbol{\lambda}(\boldsymbol{\theta})$ and hence determine $\boldsymbol{\Theta}$ – the set of all $\boldsymbol{\theta}$ that satisfy equation (9.47).*
2) *For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, determine $I\left(\mathbb{Q}^{\theta}, \mathbb{P}|X_0\right)$ using equation (9.53).*
3) *Determine $\widehat{\boldsymbol{\theta}} = min_{\boldsymbol{\theta} \in \Theta} I\left(\mathbb{Q}^{\theta}, \mathbb{P}|X_0\right)$*

<u>Stage 2</u>

Once $\boldsymbol{\theta}$ has been determined, we now turn to Monte Carlo simulation to evaluate equation (9.60). Using one loop, we can effectively evaluate both expectations simultaneous.

Start by setting a dummy variable, say $sum1 = 0$ and $sum2 = 0$. In each iteration, we will add the result to this variable.

Iterate the following algorithm exactly m times, where m is the number of simulations performed[70]*:*

a) *Generate two sets of $\boldsymbol{J}(t,T) \equiv (J_1(t,T), J_2(t,T), \dots, J_n(t,T))$ via the methods in Section 10.*
b) *Generate two sets of $\boldsymbol{Y}_{\boldsymbol{J}(t,T)} \equiv Y_{J_1(t,T)}, Y_{J_2(t,T)}, \dots, Y_{J_n(t,T)}$ using each set of $\boldsymbol{J}(t,T)$ and equations (9.58) and (9.21)*
c) *Determine $P_{t,T} = \sum_{k=1}^{n} r(k) J_k(t,T)$ using the first set of $\boldsymbol{J}(t,T)$.*
d) *With knowledge of S_t and X_t, evaluate the following (argument in the expectation in the numerator of equation (9.60) using the first set of generations for $\boldsymbol{J}(t,T)$ and $\boldsymbol{Y}(t,T)$:*

$$\exp\left(\sum_{i=1}^{n} \hat{\theta}_i \, Y_{J_i(t,T)}(i)\right) \exp\left(-P_{t,T}\right) \left(S_t e^{\sum_{i=1}^{n} Y_{J_i(t,T)}(i)} - K\right)^{+} \tag{9.61}$$

e) *Add the result to the dummy variable $sum1$*
f) *With knowledge of X_t, evaluate the following (argument in the expectation in the denominator of equation (9.60) using the second set of generations for $\boldsymbol{J}(t,T)$ and $\boldsymbol{Y}(t,T)$:*

$$\exp\left(\sum_{i=1}^{n} \hat{\theta}_i \, Y_{J_i(t,T)}(i)\right) \tag{9.62}$$

g) *Add the result to the dummy variable $sum2$*
h) *Return to step (a).*

4) *Set $O(C)(t,T,S_t,K,X_t) = \dfrac{sum\,1}{m} \Big/ \dfrac{sum\,2}{m} = \dfrac{sum\,1}{sum\,2}$*

The next section deals with some of the tools required to evaluate the above alogirthm such as estimation of $\boldsymbol{J}(t,T)$**.**

[70] m will usually be a very high number, possibly of the order of 10,000 or 20,000. The higher the number, the lower the variability in the answer and the number should be chosen such that the final answer has an acceptable level of variability. For example, it may be acceptable to choose a number such that the first three significant figures stay constant. The choice of m will also depend on the variance of the random variable(s) to be simulated. Lower variance random variables require less simulations to achieve answer stability than those with a higher variance.

10. Simulating hidden homogeneous, continuous time, discrete state Markov processes

10.1. Introduction

Simulation of the process is often done in two ways. The most accurate method involves simulating the time between state jumps. Essentially there are two variables to simulate, the time till the next jump and to which the state the process jumps to. This gives the one complete information of the sample path of the process over the entire sample period.

The second method involves calculating the value of the process at discrete points; known as "discretising" the process. This involves dividing the sample period into an arbitrary number of points and calculating the probability transition matrix. The simulating procedure then involves simulating only which state the process will be at the next time interval.

Firstly, we recall some initial relationships that exist between the transition matrix, the generator matrix and the jump time.

There exists a unique relationship between the generator matrix $\boldsymbol{Q}$ and the transition matrix $\boldsymbol{P}(t)$. Both matrices are of dimension n, the number of states. This unique relationship is described by the following equation:

$$\boldsymbol{P}(t) = e^{t\boldsymbol{Q}} = \sum_{k=0}^{\infty} \frac{(t\boldsymbol{Q})^k}{k!} \quad (10.1)$$

which we recall from Section 3.1. Since the exponential of a matrix is not well defined, the expansion is needed. This relationship is identical to the standard scalar Taylor expansion of the exponential function. Note that it is common to recognise that a matrix to the power of zero is equal to the identity matrix. Given that matrix multiplication and addition is well defined, computation of $\boldsymbol{P}(t)$ is simple. For sufficiently small t, $\boldsymbol{P}(t)$ may be approximated by the first few terms accurately enough for the purposes of simulation.

Thus, given either the generator matrix or the transition matrix, the other may be deduced. The elements of the generator matrix $\boldsymbol{Q}$ have a well-understood interpretation. The off diagonal elements of the matrix $q_{ij}, \forall\ 0 < i,j < n;\ i,j \in \mathbb{Z}$ represent the intensities of the *independent* Poisson processes of transition from state i to j.

The negative values of diagonal elements of the matrix $-q_{ii}$ represent the Poisson intensities of transition to any state from state i.

The next section describes the two types of simulation procedures mentioned above.

10.2. Type 1: Simulating Jump times and state destinations

By knowing the Poisson intensities we may thus simulate the time till the next transition from state i by simulating an exponential variable with rate parameter $-q_{ii}$.

Assuming we can generate uniform (0,1) random variables, we may simulate jump times as follows:

The exponential cumulative distribution for a random variable X with rate parameter λ is $F_X(x) = 1 - e^{-\lambda x}$. Thus we have $F_X^{-1}(x) = \ln(1-x)/(-\lambda)$. Since $0 \le x \le 1$, by standard simulation theory, we may generate a sequence of random numbers between 0 and 1 using a random number generator. Plugging these values into the inverse cumulative distribution one may simulate values for the exponential random variable X. It is worth noting that if $X \sim U(0,1)$ then $1 - X \sim U(0,1)$. Therefore simulation of the exponential variable X is simplified slightly and we may generate exponentially distributed waiting times, t, according to the following equation:

$$t = \frac{\ln U}{q_{ii}} \tag{10.2}$$

where U is a Uniform(0,1) random variable.

The next step is simulate to which state the jump occurs. The probability that the jump from state i to j occurs is $q_{ij} / \sum_{k \ne i} q_{ik}$.

If we define Y as a random variable describing the next state to which the process jumps to from state i, then Y is a discrete random variable taking $n-1$ values. Firstly partition the unit interval into n partitions. Let $x_i(k)$ describe the k'th partition point on the unit interval given the jump occurs from state i. Firstly, define $x_i(0) = 0$ and then define $x_i(k)$ as follows:

$$x_i(k) = \begin{cases} \displaystyle\sum_{1 \le j \le k} \frac{q_{ik}}{\sum_{l \ne i} q_{il}} & ; k < i \\ x_i(k) = x_i(k-1) & ; k = i \\ x_i(i) + \displaystyle\sum_{i < j \le k} \frac{q_{ik}}{\sum_{l \ne i} q_{il}} & ; k > i \end{cases} \tag{10.3}$$

This implies $0 \le x_i(1) \le x_i(2) \le \cdots \le x_i(n) = 1$. Defining $x_i(0) = 0$ and considering U_2 to be another Uniform(0,1) random variable, then the expression $\Pr[x_i(k) < U_2 \le x_i(k+1)]$ exactly represents the probability of a jump to state $k+1$. Note that since $x_i(i) = x_i(i-1)$, this formulation implies that a jump from a state can never be back to that same state almost surely[71]

We continue with this, repeating each step by generating successive jump times T_k^i (the k'th jump away from state i) and jump destinations j. Eventually the jump time will exceed the

[71] It is possible that a uniform generation will equal that same single value, i.e. $U = x_i(i)$ but this occurs with zero probability even though the event is possible.

observed sample period. When this occurs, it is assumed that no jump occurs in the final period leading up to T and the cycle ends.

10.3. Simulating Total Occupation Times

Total occupation times, defined as $J_i(t,T)$ being the total time that the Markov Chain is in state i between t and T. This function is particular useful when describing time homogenous Markov Chains. To examine the benefits, take for example our MSM. Knowledge of the initial state and vector $\boldsymbol{J}(t,T) \equiv (J_1(t,T), J_2(t,T), \dots, J_n(t,T))$ is enough information from the Markov Chain to give the distribution of the stock value at the end of the period (at time T). This is implies that the order of jumps and to which state is unimportant, only the total time spent in each state. This is a direct consequence of the time homogeneity of the process.

Some useful identities follow from this. Firstly, it is clear that $T - t = \sum_{i=1}^{n} J_i(t,T)$. Secondly, is easy to express $J_i(t,T)$ as:

$$J_i(t,T) = \sum_k T_k^i \tag{10.4}$$

That is, we just sum the successive waiting times for which a jump is away from state i in the period t to T.

10.4. "Discretising" the process

We could also divide the sample period T into J subperiods of length $\Delta t = T/J$. We then need to calculate $\boldsymbol{P}(\Delta t)$ representing the probability transition matrix per subperiod Δt. From equation (10.1), we may approximate $\boldsymbol{P}(\Delta t) \approx I + \Delta t\boldsymbol{Q} + {}^1/_2\,(\Delta t\boldsymbol{Q})^2$. We then follow a similar method that we used to simulate a discrete time Markov Chain. The time period is fixed and so only one simulation is required to represent the next state destination. In this case, there is also the possibility that the process remains in the same state at the next subperiod. Now define $x_i(k)$ as follows:

$$x_i(k) = \sum_{j=1}^{k} p_{ij}(\Delta t) \tag{10.5}$$

where $p_{ij}(\Delta t)$ is the (i,j)'th entry of matrix $\boldsymbol{P}(\Delta t)$. By generating another Uniform(0,1) random variable we may again simulate the destination state by observing which interval the simulated value falls in.

In comparison, Type 1 simulation is more accurate than Type 2 as the timing of state switches (and destinations) are measured to the highest degree of accuracy.

Computational efficiency is less obvious. Although two generates of random variables are needed to complete one iteration of a Type 1 simulation and one generation for Type 2, it is likely that less iterations will be needed for Type 1 than for Type 2. This will be even more true the smaller the diagonal elements of $\boldsymbol{Q}$ are in absolute value and the larger $\boldsymbol{J}$ is.

However, It could thus be argued that in most scenario's Type 1 is superior to Type 2.

11. A South-African Study of TOP40 Data

11.1. Introduction

South Africa's equity market is owned by the Johannesburg Stock Exchange Limited but is usually known just as the JSE. The JSE Limited and the FTSE Group entered in a joint-venture to create the FTSE/JSE Africa Index Series Indices. The two most relevant indices that represent overall market movements are the FTSE/JSE All-share Index (or simply ALSI for short) and the FTSE/JSE Top 40 Index (or TOP40 for short).

The JSE is characterized by a large concentration of resource stocks. Both the TOP40 and ALSI characterize a very large proportion of the total market capitalization and therefore are broadly represented of overall market movements.

The derivatives market of the JSE has options on the TOP40 futures index. There are many published index-traded derivatives and also a considerable over-the-counter (OTC) market.

In this section we perform the analysis detailed in Section 4 to 10 to develop a Hidden Market Model on the TOP40 data. The estimation procedure will determine the number of hidden states, the transitional probabilities and the distribution parameters (μ_i and σ_i) associated with the MSM. Thereafter, various option prices are calculated and assessed for reasonability. Throughout this Section, reference to "Option price" refers to a vanilla call option on the stock.

11.2. Empirical Analysis (1): Introduction

The Study period is 03/01/2005 to 20/07/2011. The study period was chosen to encompass the early "bull" period as well as the "bear" period and the market crash that was present during 2008.

It must be noted that the study period chosen is expected to drastically affect the estimation results. This is known as *sample selection bias.* This is especially true for the transitional probability matrix. For example, should the study period chosen be such that it solely encompasses the MSM while it is in one state, then the transitional probabilities are likely to be underestimated. In the extreme case, the estimation procedure may assign a zero transitional probability to other states. These are the problems associated with using a period that is too short.

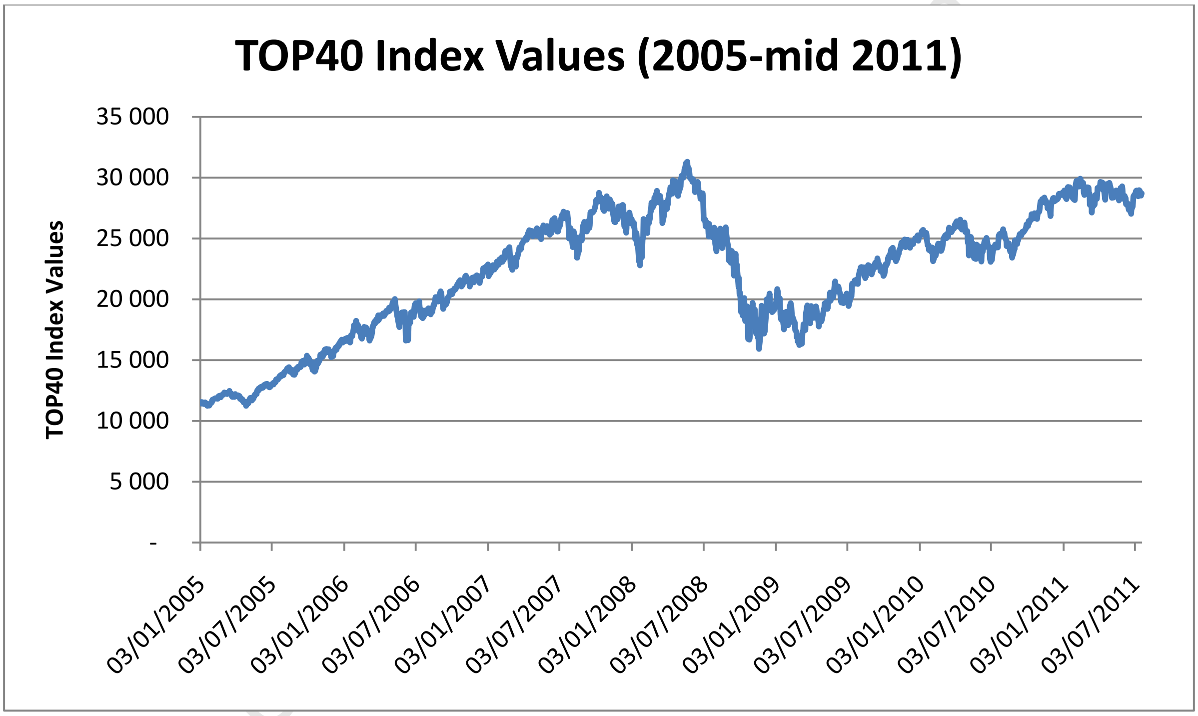
Of course, we must realize that if we are attempting to create a model that will effectively "forecast" or rather capture the inherent future volatility and trends in the TOP40, then estimates based on past-period figures are not a reliable predictor of future stock market movements. Therefore, using too long a study period implies that the model is not "robust" enough to estimate MSM parameters should they be time varying. Essentially, we would prefer the model to be more biased towards the most recent stock values as this is most likely give parameter estimates similar to those that future stock values will give.

Thus, the choice of the sample period is a delicate affair and needs to be chosen to match the purpose. Since we are essentially presenting a framework for using MSM's in stock market data. Here, we have chosen a sample period that best represents potential state transitions.

Data was extracted from *INET Bridge* over the past year.

The following graph summaries the trends in the data over the sample period[72]:

Figure 1.1: Top40 Index Values



In the graphic above, we clearly see the upward trend in market values over the 2005 to late 2007 period. The market was commonly considered to be in a "bull" period in this phase. Some excessive volatility occurs between 2007 and mid 2008 before the market values sharply fell. The TOP40 was at a maximum in the sample period of 31,315.34 on 22 May 2008. Between then and 9 March 2009, the market fell 14,981.24 index points, a drop of roughly 48%.

[72] Note that many days in the sample period are not "trading" days as a result of them either falling on a weekend or on a public holiday. In essence, we essentially assume the HMM is merely "paused" for those days and resumes on the trading days. Thus *actual* time does NOT correspond with market trading times. Furthermore, due to this anomaly, not every "trading" year is equivalent so care must be taken in calculating the correct time to expiry of any options that may be valued.

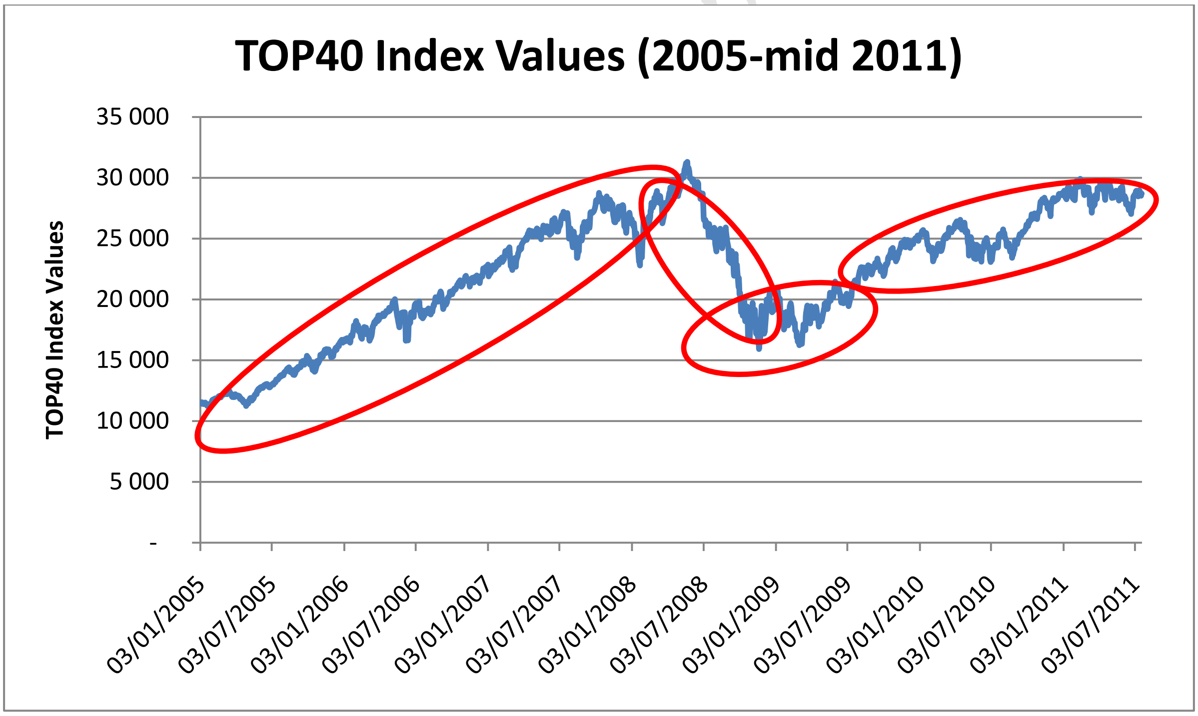
This period and the excessive volatility that appears in the subsequent six-months is reflective of overall global market turmoil. September 2008 saw the demise of the Lehman Brothers bank and is known to some as "Black September".

The period until the end of the sample period is representative of steady growth together with higher-than-usual volatility. Fears of a double-dip recession in current times outside of the sample period (during late 2011) are rife and this has translated into more volatility in the index values. Hints of this are reflected in the index values near the end of the sample period.

Before delving into the process of applying the mathematical techniques we have already developed, we may be able to get a reasonable idea of the number of states we may expect and furthermore what kind of parameters we may expect.

For instance, we may reasonably expect the early bull periods to be representative of a single state, where both μ and σ appear reasonably constant. The following graph represents possible states.

Figure 9.2: Top40 stock trends and possible states



This is of course only one possible representation. For example, the rightmost circle may just be a revisit to the state in the first circle (although the upward trend appears lower). The second circle from the right may simply be representative of the same state as that in the rightmost circle, but perhaps just in a period of excessive volatility.

We may therefore, reasonably expect 2-4 states and hope that the mathematical analysis is reflective of this or if not, there are not much more than 4 states.

Stock values were converted to daily (or rather one-period) log-returns, or R_t in the format we have been accustomed to. The following table outlines the sample statistics were computed based on the log-returns:

Sample Statistics		**Gaussian Statistics**	
Minimum	-7.9594%	Kurtosis	6.0892
Maximum	7.7069%	Skewness	-0.1380
Mean	0.0557%	Drift (mu)	0.0678%
Median	0.1455%		
Upper Quartile	0.9084%		
Lower Quartile	-0.7414%		
Variance	0.0243%		
Standard Deviation	1.5585%		
Number of Observations	1,636		

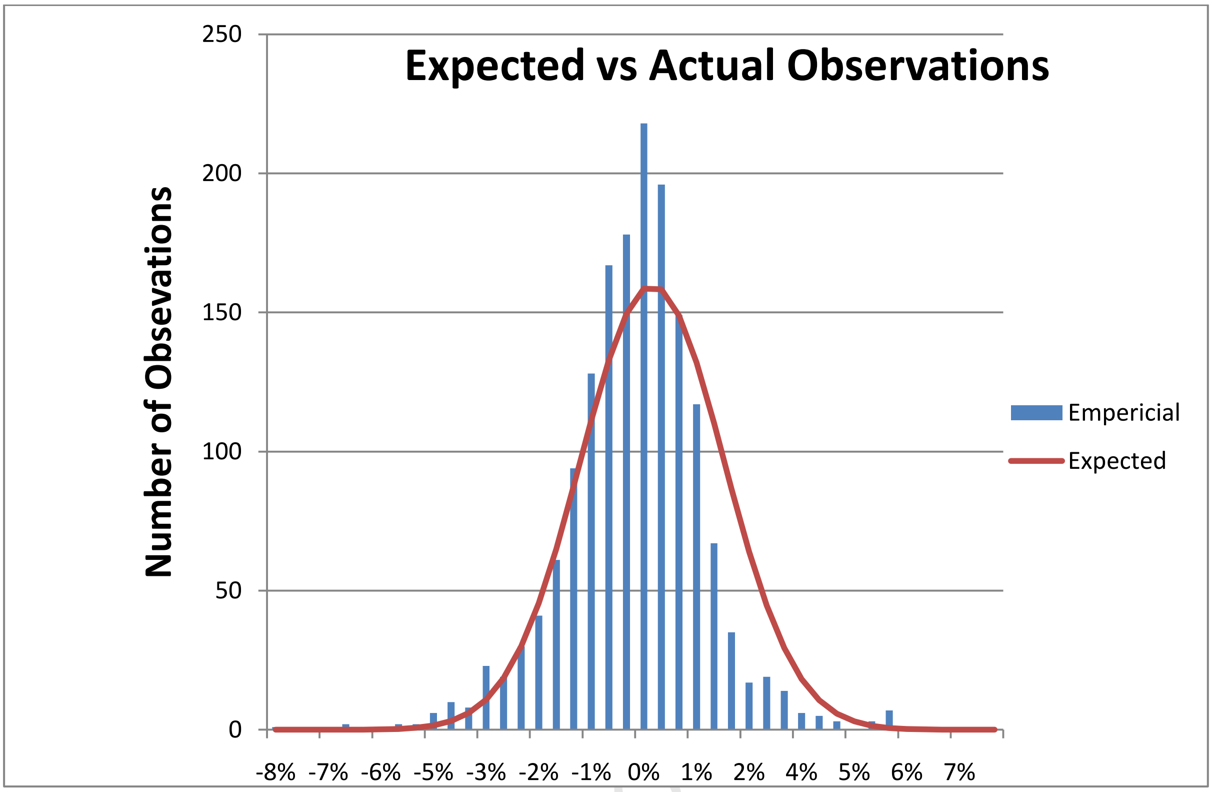
Some stand out comments that can be made is that the median differs from the mean by about ten or so basis points[73]. This is reflective in the skewness statistic which indicates a negatively skewed sample distribution. The kurtosis is greater than the usual 3 that is expected in normal or Gaussian distributions in general and quite excessively so.

The above statistics on first glance appear to represent sample values that possibly have excess kurtosis and negative skew and hence represent a deviation from the normal distribution.

Data was then organized into bins to create a histogram and an expected normal distribution plot (with the sample mean and sample standard deviation as parameters) has been overlayed for a graphical comparison.

[73] 100 basis points refers to 1% and hence 1 basis point refers to 0.01%.

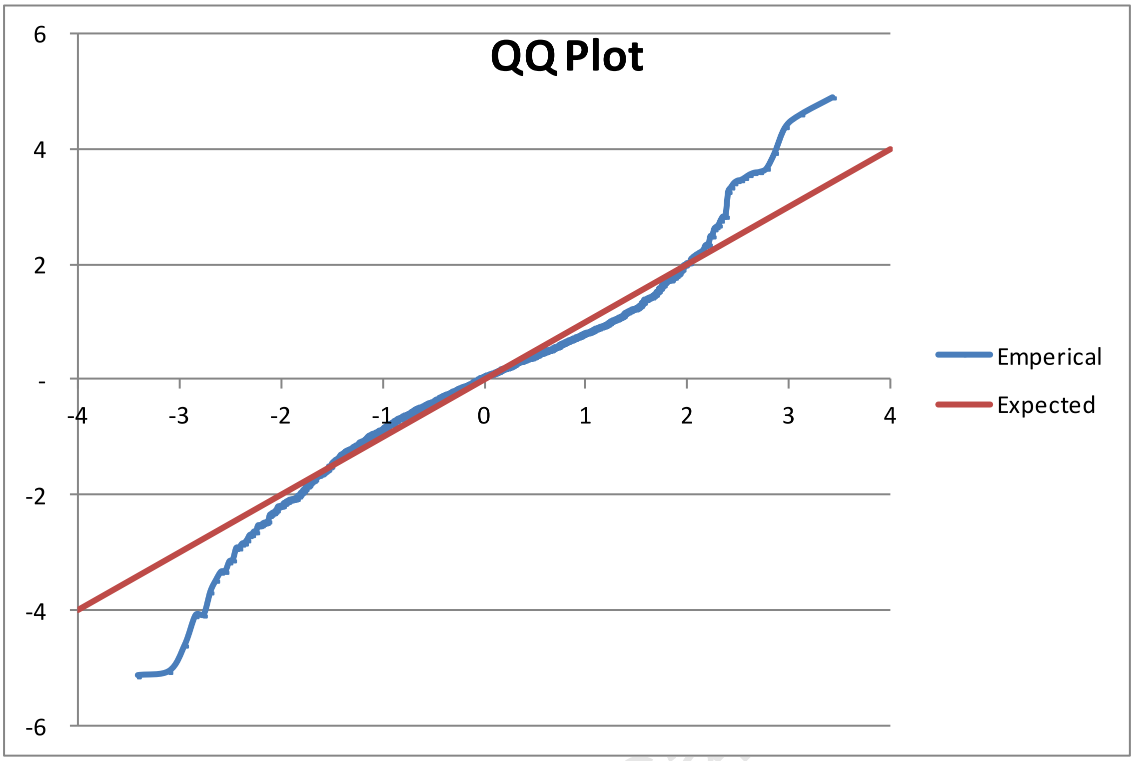
Figure 9.3: Gaussian Expectations vs. Actual Observations



The graph confirms what we can deduce from the sample statistic: The data is negatively skewed away from the theoretical symmetrical normal distribution which should have an equal distribution about the mean.

The empirical data is also has much more data clumped around the mean than expected which has resulted in excess kurtosis. To further confirm the deviation from the normal distribution, I have plotted a quantile-to-quantile plot (or QQ-plot) below.

Figure 9.4: QQ-Plot of Emperical Observations



The straight line shows where we may expect the various points to lie should they arise from a normal distribution. The figure above shows significant variation from that. A Shapiro-Wilks test was performed. The null-hypothesis being that the data comes from a normal distribution and the alternative hypothesis is that it arises from some other distribution.

The test-statistic was 0.96 with an associated p-value of less than 0.01, a highly significant result pointing towards the premise of non-normality.

Given the above results, it is plausible that we deviate away from the standard GBM stock model. In the following section we attempt to fit our MSM using the methods described in previous sections of the report.

11.3. Empirical Analysis (2): Fitting an MSM

We begin with choosing the number of states, a pre-requisite to using some of the estimation methods already discussed. We follow the Sequential Pruning method together with the MMDL criterion discussed in Section 6.

The intial process is to select a n_{max} and n_{min}. Based on the above, we can conclude that $n_{max} = 4$ and $n_{min} = 1$ are appropriate. With $n_{min} = 1$, it reduces to the standard normal distribution or GBM model.

A tolerance of 0.000001 was chosen for the log-likelihood function: this means that should the difference between successive log-likelihood values (i.e. for successive iterations) be less than 0.000001 then the process stops. I have chosen 0.000001 due to some empirical studies that have shown me that there is no variation in parameter estimates after the 4th decimal place which is suitable to fit the MSM and for use in option pricing.

The sequential pruning process begins by initializing the process with $n = n_{max} = 4$. We also need an initial guess for both $\boldsymbol{\pi}$ and $\boldsymbol{P}$. For the purposes of simplicity, we will assume no prior knowledge about the likelihood of any long-term probabilities and assume that $\boldsymbol{P} = \{\frac{1}{4}\}$.

The initial μ and σ are chosen to be representative of the sample mean and standard deviation. These were also doubled and halved to test whether the BW algorithm had perhaps encountered local maximum, but there were no changes to the result. The following was used:

$$\boldsymbol{\mu} = \begin{pmatrix} 0.0678\% \\ \vdots \\ 0.0678\% \end{pmatrix} \tag{11.1}$$

$$\boldsymbol{\sigma} = \begin{pmatrix} 1.5585\% \\ \vdots \\ 1.5585\% \end{pmatrix} \tag{11.2}$$

<u>4 States Results</u>

$$\boldsymbol{P} = \begin{pmatrix} 95.0823\% & 4.9177\% & 0.0000\% & 0.0000\% \\ 62.6495\% & 24.7661\% & 12.5944\% & 0.0000\% \\ 2.0277\% & 0.0000\% & 97.3158\% & 0.6566\% \\ 0.0000\% & 0.0000\% & 3.2073\% & 96.7927\% \end{pmatrix} \tag{11.3}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.2254\% \\ -1.7513\% \\ 0.0139\% \\ -0.1457\% \end{pmatrix} \tag{11.4}$$

$$\boldsymbol{\sigma} = \begin{pmatrix} 0.8942\% \\ 0.8577\% \\ 1.9507\% \\ 3.6874\% \end{pmatrix} \tag{11.5}$$

$$\boldsymbol{\Pi} = \begin{pmatrix} 64.3329\% \\ 4.2046\% \\ 26.1162\% \\ 5.3462\% \end{pmatrix} \tag{11.6}$$

The log-likelihood function was 4,773.899 with an associated MMDL value of 4,627.824.

The results are interesting, especially the nature of the transition matrix with there being six entries of zero transition probabilities. All states do not fully communicate with each other with Transition to state 2 being only possible from state 1 for example.

The next step is the pruning process whereby we remove the associated probabilities relating to the state with the lowest steady-state probability which is state 2. The remaining transitional probabilities are then scaled up such the rows in the resultant matrix still sum to unity. The resultant matrix and the previous estimates for μ and σ are then used as inputs in our estimation scheme for the next state in the process, which is $n = 3$.

<u>3-states results</u>

$$\boldsymbol{P} = \begin{pmatrix} 95.7736\% & 4.2264\% & 0.0000\% \\ 1.4661\% & 97.7146\% & 0.8193\% \\ 0.0000\% & 1.6633\% & 98.3367\% \end{pmatrix} \quad (11.7)$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.2205\% \\ 0.0964\% \\ -0.1012\% \end{pmatrix} \quad (11.8)$$

$$\boldsymbol{\sigma} = \begin{pmatrix} 0.6854\% \\ 1.1383\% \\ 2.4689\% \end{pmatrix} \quad (11.9)$$

$$\boldsymbol{\Pi} = \begin{pmatrix} 18.8579\% \\ 54.3632\% \\ 26.7789\% \end{pmatrix} \quad (11.10)$$

The associated log-likelihood function is 4,754.356 and the MMDL value is 4,665.241

The above results are very intuitive. State 1 could represent a bull-state, of positive growth and low volatility. State 3 could represent a bear-state characterized by negative growth and high volatility. State 2 could represent an intermediate state characterized by average drift and moderate volatility.

The transition probability matrix also shows some interesting results. One cannot transition straight from a bear-market to a bull-market or vice versa. You would need to pass through the intermediate "average" state in order to do so. This also makes sense when we consider some theories that the stock market is cyclical, where we seesaw between boom and recessionary periods. State transitions are rare due to the high persistence.

Since the MMDL result is higher than the previous MMDL under the 4-state scenario, we can conclude that the 3-state model is a better fit than the 4-state model and the process continues. We remove state two and scale up the transition probabilities as before.

<u>2-states</u>

$$\boldsymbol{P} = \begin{pmatrix} 99.2399\% & 0.7601\% \\ 1.7659\% & 98.2341\% \end{pmatrix} \quad (11.11)$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.1360\% \\ -0.0931\% \end{pmatrix} \quad (11.12)$$

$$\boldsymbol{\sigma} = \begin{pmatrix} 0.9978\% \\ 2.3897\% \end{pmatrix} \quad (11.13)$$

$$\boldsymbol{\Pi} = \begin{pmatrix} 69.9096\% \\ 30.0904\% \end{pmatrix} \quad (11.14)$$

The associate log-likelihood function is 4,736.040 with an associated MMDL value of 4,717.24.

This model reduces to the classic case of the bull and bear states. The bull-state (state 1) exhibits positive drift and low volatility while the bear state (state 2) exhibits negative drift with high volatility. The model is further characterized by high persistence and a low number of expected transitions (state switching is rare and once it occurs, it remains in that same state for a long-period of time). This makes intuitive sense, switching should be rare as we do not expect to get frequent switches between bull and bear phases on a daily basis.

Since the MMDL value is higher than was previously calculated and we can favour the 2-state model over the 3-state model.

For completeness sake, we also need to consider the 1-state model. Essentially, this reduces to the standard GBM model and we can easily calculate the log-likelihood function using standard likelihood theory by $LogL_{Gauss}\,(\boldsymbol{\theta})$:

$$logL_{Gauss}\,(\theta) = \sum_{i=1}^{T} \ln 1 - \ln(\sqrt{2\pi\sigma}) + \left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \quad (11.15)$$

The result comes to 4,487.267 and the associated MMDL value is 4,483.567.

Since the MMDL value is lower for the one-state model, we can conclude that one-state model is <u>not</u> better than the two-state model. Therefore, the above analysis shows that the 2-state model is the optimal model to describe the sample and we proceed forward with the stated parameters.

11.4. Option Pricing using Empirical MSM

11.4.1. Introduction

The estimation of the risk-free-rate is important in the context of option pricing. As discussed in Section 4.9, risk-free rate estimation for each state is not explored in this thesis.

The approach for this section is very simplistic for illustration purposes. The risk-free yield curve (sourced from the Bond Exchange of South Africa) has exhibited rates roughly between 5% and 9% depending on the term and point in time.

Given this, and the negative correlation between bonds and equity, we will assume that the risk-free vector:

$$\boldsymbol{r} = \begin{pmatrix} 7.0000\% \\ 7.0000\% \end{pmatrix} \quad (11.16)$$

In other words, a low risk-free rate is associated with higher drift and lower volatility. NB: The above vector is the parameter vector in for a *per annum* return.

This simplistic approach is of course not accurate but is used for illustration purposes as risk-free bond prices become simple to estimate allowing for seemless put-call parity calculations.

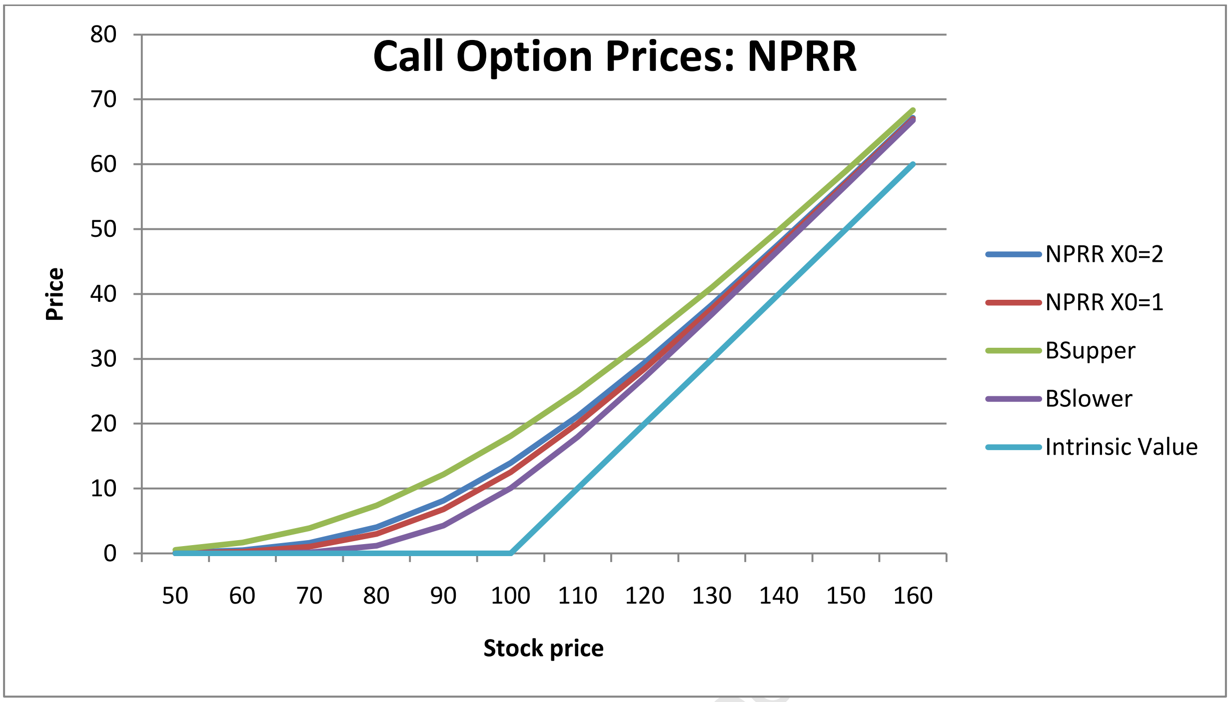
For the following Sections, we assume that there are 250 days in a year and specify time in years. This allows us to represent all parameters in a daily format.

11.4.2. The Non-Priced Regime Risk Option Study

We will now employ the methods outlined in Section 9.3.2. The requisite inputs are X_0, $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{P}$ and $\boldsymbol{r}$. It is only X_0 that is unknown. For completeness, we will show option prices under both state 1 and state 2. 10,000 simulations were used in order to calculate the option prices.

In the non-priced regime risk (NPRR) option formula, we have clear boundaries where we expect option prices to lie. Given our outstanding of the BS formula, the high volatility, scenario will yield the highest price. On the other hand, the lowest price will be for the low volatility scenario. Since there is no additional premium associated with state-switching, this gives us useful boundaries for the vanilla call option price. The figure is presented on the following page.

Figure 9.5: Non-priced regime risk vanilla call option prices



We have shown both the intrinsic price[74] as well as the two option prices for the two different starting states. As expected, the option price lies between the two Black-Scholes boundaries and so the answers are reasonable.

The option price when the starting state is state 2 appears to lie slightly above that of state 1 and this is also intuitive given that we start in a high volatility environment, however there appears to be little difference. As expected, the four price lines converge at two ends of the graph. This graphic makes sense.

For completeness, we also use put-call parity[75] to sketch a vanilla put option which we get by put call parity is:

$$Put(K, S_t, \boldsymbol{r}, \boldsymbol{\sigma}, X_o, T) = Call(K, S_t, \boldsymbol{r}, \boldsymbol{\sigma}, X_0, T) - S_t + Ke^{-rT} \tag{11.17}$$

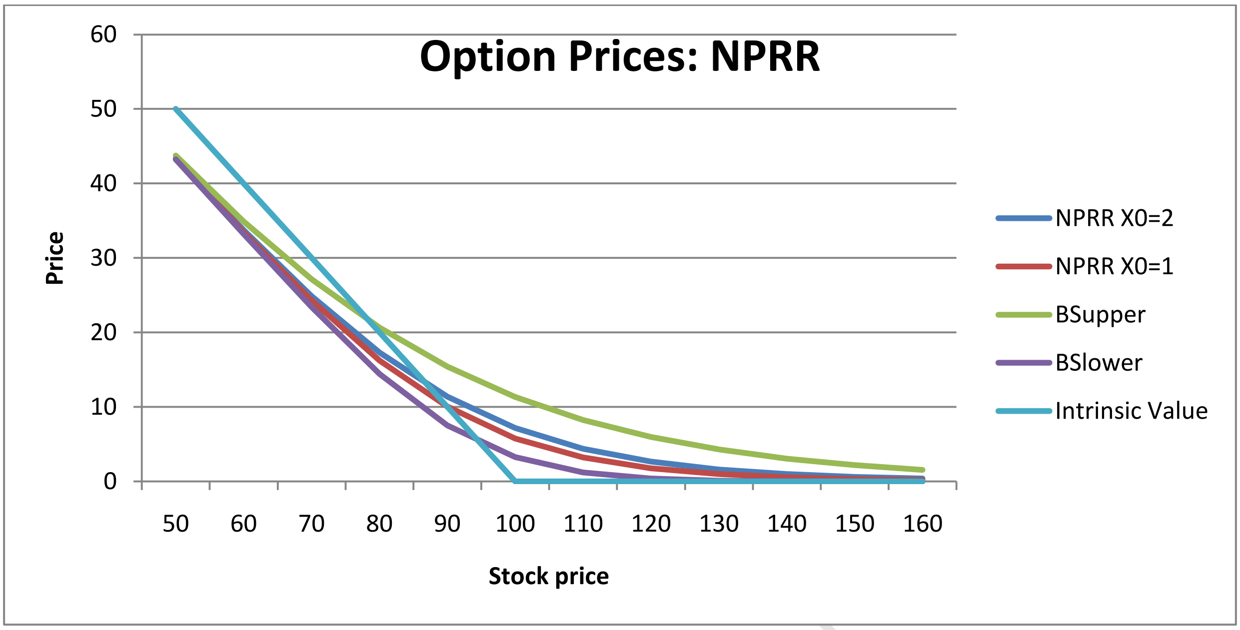
Figure vanilla for puts is as follows:

---

[74] The Intrinsic price for a vanilla call option is the maximum between zero and the stock price less the strike price.

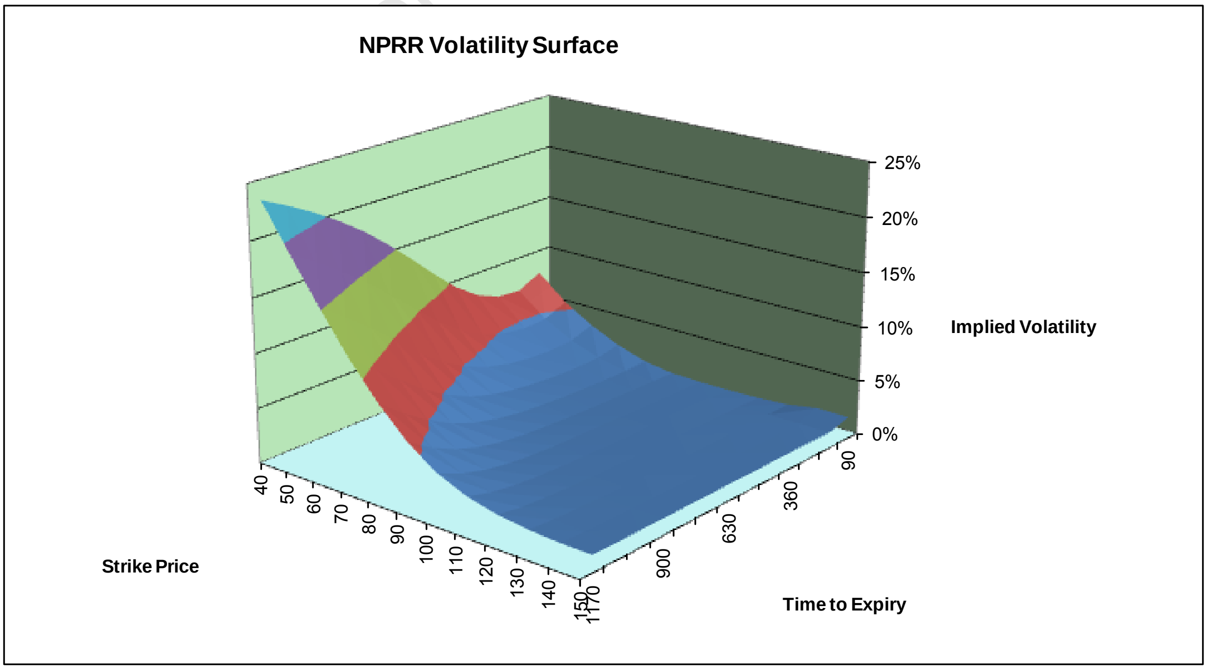
[75] Put-call parity states that for vanilla call and put options: The price of a vanilla call and stock (underlying the vanilla call option) is equal to the price of a corresponding vanilla put (same strike, underlying and expiry) plus a zero coupon bond with redemption amount equal to the strike price and the same expiry date of the options.

Figure: 9.6: Non-priced regime risk vanilla put prices



In order to really ascertain the appropriateness of the method and whether it exhibits the kind of option price patterns we might expect, it is perhaps better to examine a volatility surface instead. Below is the figure for when the initial state is state 1[76], with the stock price at 100.

Figure 9.7: Non-priced regime risk implied volatility surface.



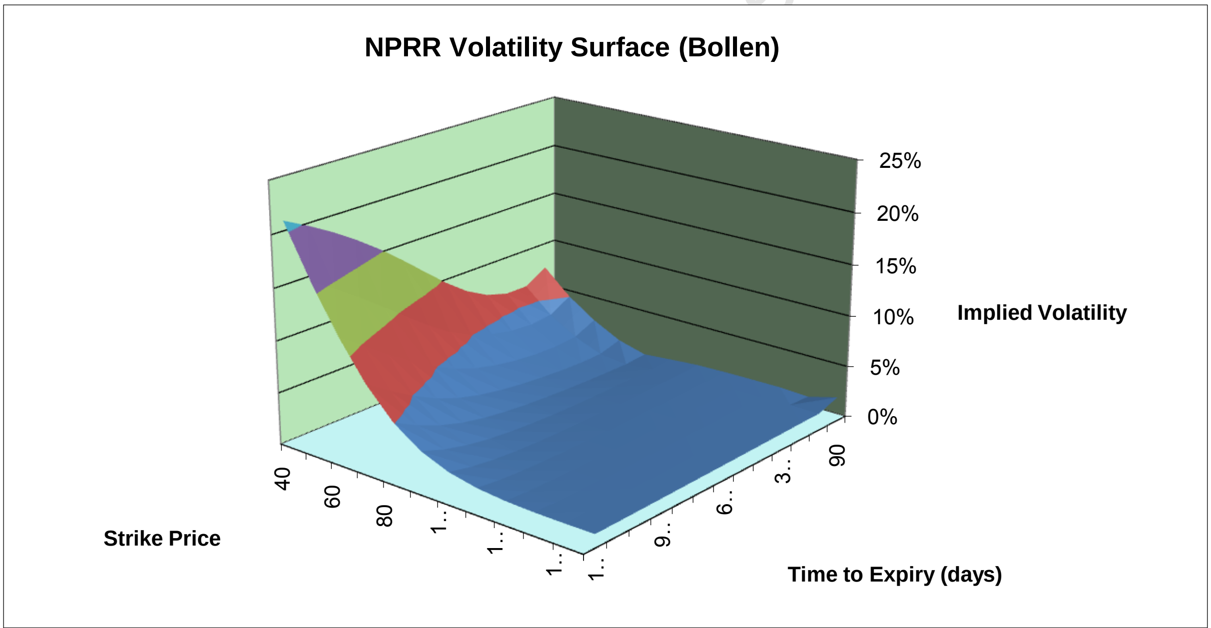
[76] We will omit the figure for when the initial state is state 2 as there is negligible difference.

This figure is gives a very intuitive appeal. If we ignore the time-dimension for a moment and examine the strike price-implied volatility plane then we notice that there is a distinct volatility skew evolving. In fact, this is sometimes called a “reverse skew” or volatility smirk, common for equity indices. Far in-the-money calls (or far out-the-money puts) have higher implied volatility than far out-the-money calls (and in-the-money puts). A common interpretation of this is that investors are concerned about market-crashes and protective puts are in demand.

Adding the time dimension to the analysis reveals that longer-dated options tend to exhibit a more extreme skew.

For completeness sake, we can also examine Bollen’s pentanomial lattice using our parameters. We will set the step size to a single day (i.e. a single observation such that $dt = 1$). Bollen’s lattice allows one to avoid the difficulty of selecting an initial state. We choose an equal probability of being in either state and arrive at the following volatility surface (again with stock price = 100)

Figure 9.8: Implied Volatility surface for Bollen Pentanomial Lattice Method



The figure has a striking resemblance to the continuous-case model as we expected.

11.4.3. The Priced Regime Risk Option Study

We now turn to the methods discussed in Section 9.3.3

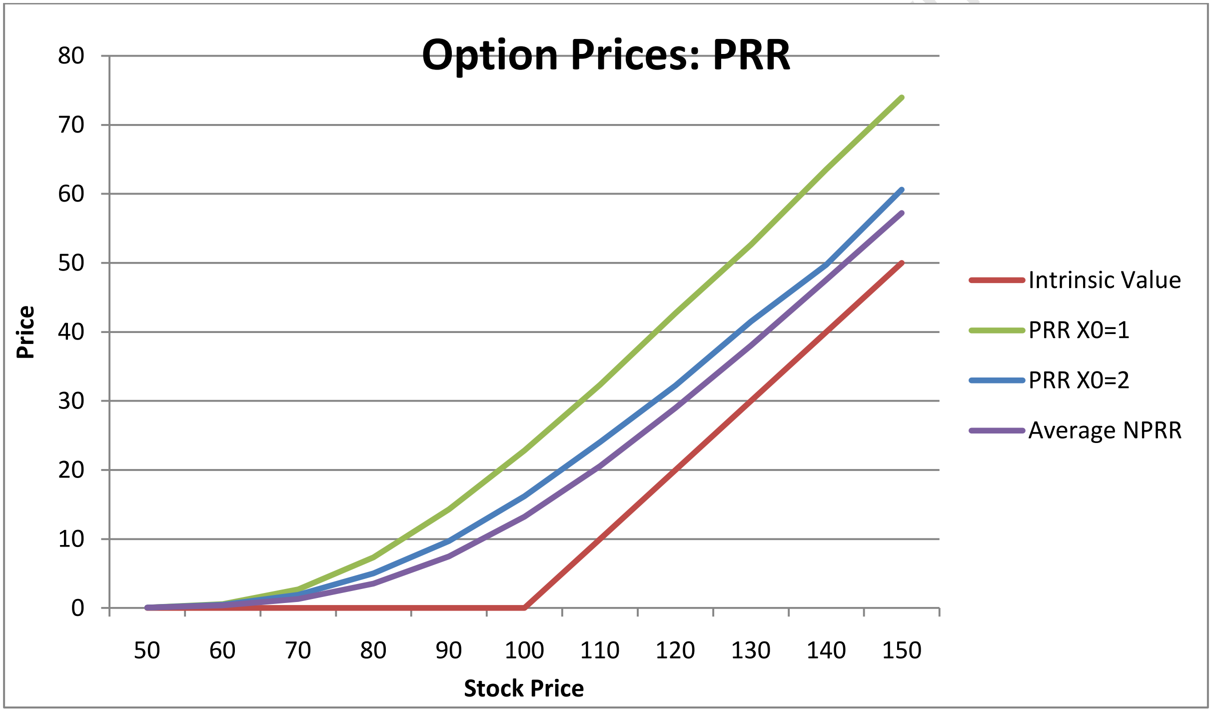
The first step is to calculate the RSET parameters needed. We will recall that the parameters depend on the level of approximation of the exponential matrix function. The higher the order of the Taylor approximation to the exponential function, the more potential solutions to equation (9.47) there are.

For simplicities sake, we will use the linear approximation which yields a time independent derivation for the RSET parameters. Recalling that $\theta_i = \frac{\mu_i - r_i}{\sigma_i^2}$ under the linear approximation, the RSET parameters turn out to be as follows:

$$\boldsymbol{\theta} = \begin{pmatrix} 10.9700 \\ -2.1227 \end{pmatrix} \quad (11.18)$$

Again 10,000 simulations are used to produce the following figure:

Figure 9.9: Priced regime risk vanilla call option prices



Here, the average NPRR price is the average of the two prices for the two potential starting states. The most noticeable feature is that for either of the states, the option price when the regime risk is priced (i.e. Priced-Regime Risk or PRR) is higher than that for the NPRR Option prices. This is again what we expect, as the additional switching risk has been incorporated into the price of the option. An interesting feature is also the sensitivity of the starting states for the PRR Option prices.

In the NPRR option price derivations, we see that the drift is not taken into account as with the general BS theory but in the PRR case, the RSET parameters depend on the drift and thus the PRR option price is dependent on the drift. There thus seems to be a significant difference depending on the current state. This is explored further by examining the volatility surfaces under the two possible initial states with stock price = 100.

Figure 9.10: Priced regime risk volatility surface under $X_0 = 1$

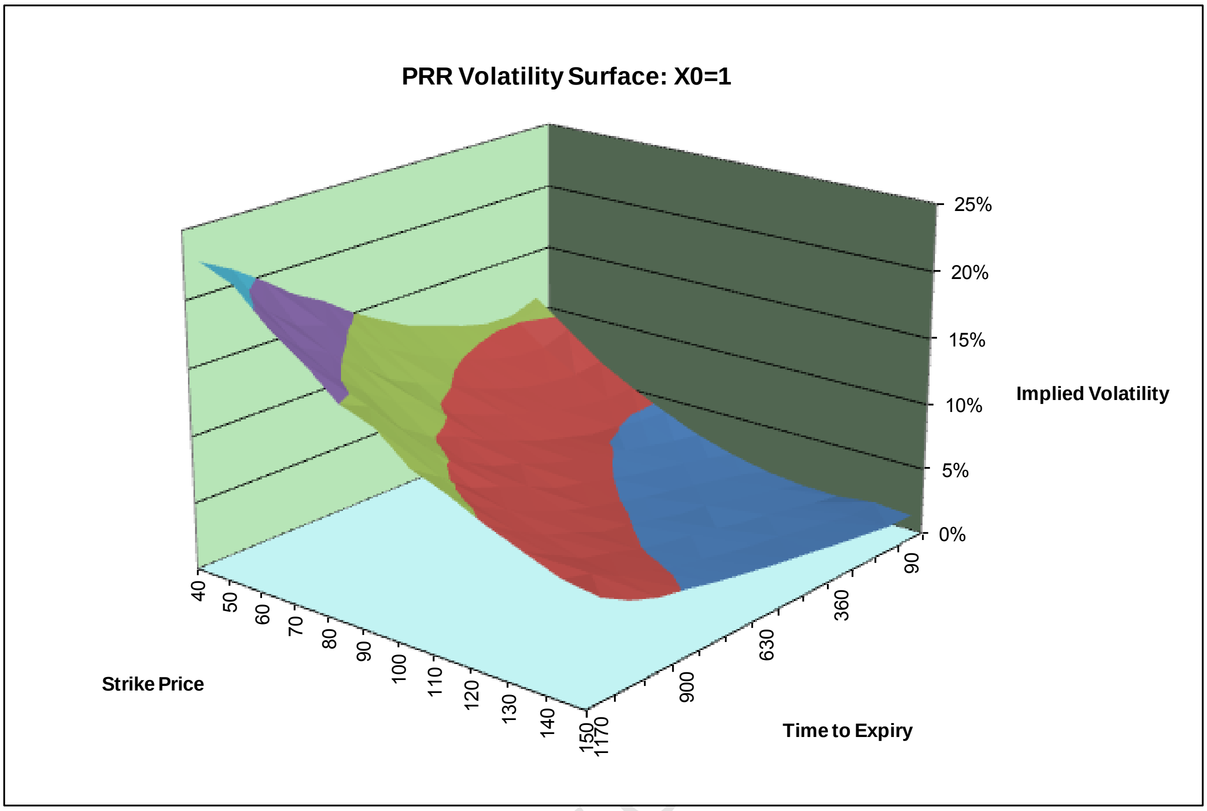
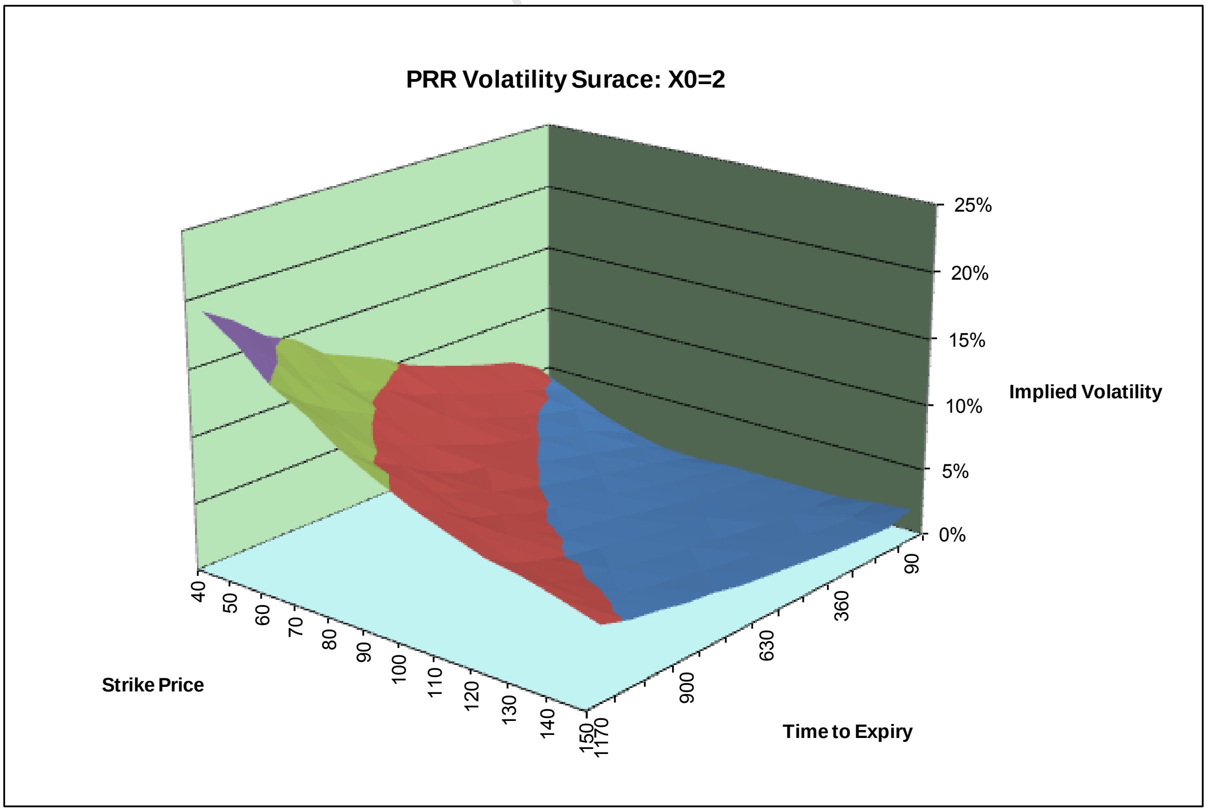


Figure 9.10: Priced regime risk volatility surface under $X_0 = 2$



It should be noted that the surfaces are not completely "smooth" and this indicates that 10 000 simulations may not be enough to accurately simulate this option price. One could consider 50 000 to better model the surface. However, time constraints would need to be considered and the increase in the required number of simulations can be a significant disadvantage.

For both surfaces, we see again a volatility smirk but it decidedly different from the one we see under the NPRR scenario. Here the volatility decreases with an increasing strike, albeit at a decreasing rate. Out-the-money calls have a lower implied volatility than at-the-money calls but both still exhibit a strong negative skew, although the skew is more extreme in the bull-market initial state scenario.

Again, adding the time dimension appears to enhance the skewing affect.

Overall, we see that the NPRR and PRR volatility surfaces have reasonably similar shapes but the scale and slope of the surface changes depending on the initial state. As discussed, the results are intuitive and represent volatility skews that we may reasonably expect in the market place.

This gives credibility to the idea that the MSM for stock prices is a suitable one for the purpose of option pricing. We cannot justify using the NPRR over the PRR model or vice versa but note that they both produce reasonable implied volatility surfaces which fits well with intuition implying that the MSM model has merit.

12. Conclusion and further avenues of research

12.1. Conclusion

The BS model broke new ground in option pricing by not only the creation of an objective option pricing formula but also paved the way for the introduction of risk-neutral valuation methods.

After the 1987 crash it became evident that the BS model was not being used in the market and the common explanation for this was that the stock price model proposed by BS was not entirely accurate as shown by implied volatility surfaces. The introduction of stochastic volatility appeared to remedy the situation of which the MSM is one such model.

We have presented the tools and theory required to accurately fit a MSM to empirical data as well as use the fitted model to estimate vanilla option prices. The method deemed best to do this is a modified[77] BW-algorithm to determine the parameters and a customized MMDL criterion with sequential pruning for selecting the number of states.

Option pricing theory was also developed and numerous potential methods were explored. In particular, the idea of priced vs. non-priced regime risk was introduced which produces differing option pricing formula. Non-priced regime risk may be just a theoretical convenience, since it seems unlikely that unexpected jumps in volatility and drift would not be taken into account by the market. Priced regime-risk brings plenty more complexities such as more difficult valuation models.

Once the model had been fitted, it was then used to price vanilla options and produce different implied volatility surfaces. The implied volatility surfaces under both the PRR and NPRR scenarios produce intuitive results leading us to the conclusion that an MSM model can produce realistic implied volatility surfaces. The most notable difference between the MSM model and the Heston model is that implied volatility tends to increase with time-to-expiry in the MSM case but appears to reduce with time-to-expiry in the Heston model.

Despite producing realistic implied volatility surfaces, the MSM model is backed up by the way we expect stock markets to work – that is, information-driven stock price thinking and behavioural finance dynamics. The idea of the presence of cyclical economic regimes affecting stock price expectations and volatility is consistent with an MSM model.

Some of the results contrast here to that seen in Anderson (2006) where he concluded that his sample of stock data and the resultant implied volatilities did not produce figures and surfaces expected. However, he did go on to say that this could be due to poor numerical optimization techniques rather than inherent failure of the model.

Besides option pricing, the MSM model for stock prices could be used for other avenues of research as discussed in the next section.

[77] A under/overflow resistant algorithm specified for Gaussian observations

12.2. Further research and Alternate Models

12.2.1. Improvements to the model

We have explored time homogeneous Markov chains as the latent variable process, however, we could also consider time heterogeneous (or time inhomogeneous) models where the transition probabilities and, in some cases, the values of μ_i and σ_i vary with time. This is a useful tool but the dangers are that these models are over specified and hence not robust should new information become available. Anderson (2006) points this out in his concluding paragraph, further mentioning that it was not a good fit to his JSE sample data. Readers are directed to Diebold, Lee, & Wienbach (1994) who present a modified Baum-Welch algorithm for determining the time-varying transition probabilities.

Another potential enhancement is to include adding an autoregressive component to the stock model. However, this violates the semi-strong form of the Efficient Market Hypothesis. Other models variants are double Markovian models, where the drift and volatility are governed by two MSM's that are correlated (Siu, Fung, & Ng, Option Valuation with a Discrete-Time Double Markovian Regime-Switching Model, 2008). Higher order models could also be considered, where the current value of a Markov Chain is dependent on the previous m say values of the chain. The addition of a jump component has also been explored (Yin, Song, & Zhang, 2005).

Semi-Markov models can also be considered (Yu, 2010). Semi-Markov models are one which the transition probabilities vary depending on the time spent (or duration) in the current state. This is in contrast to our HMM model where transition away from the current state is duration independent.

12.2.2. Alternative testing

We have seen that the MSM models given produce appealing volatility surfaces. To examine the true appropriateness of these models to a practitioner, it is perhaps best if option prices and implied volatility surfaces are compared to those currently seen in the market (or just after the sample period).

One further avenue is to explore techniques of finding the implied parameters such that the resultant implied volatility surface best fits the observed implied volatility surface. Assuming the stock model is correct, this will help us gain valuable insight into what the average market participant's outlook is on the likelihoods of bull and bear markets. It would furthermore allow us to gain insight into the magnitude of risk premium market participants are adding in the face of future uncertain stock price movements.

12.2.3. Further avenues

Although we have shown the benefits of this model under vanilla calls and puts, the ideas in this paper could be extended to other more exotic options. Anderson (2006) considers American and Bermudan options as well as Bermudan Swaptions. More exotic options such as Asian and Lookback options are explored by Boyle & Draviam (2007) where option price PDEs are given for each type of derivative.

If the MSM can be shown to produce implied volatility surfaces that are consistent with observed volatility surfaces, it opens the door to using this model for other applications in financial analysis such as portfolio selection (Chen, Yang, & Yin, 2008) and risk measurement (Siu, Ching, Fung, Ng, & Li, 2009).

MSM models can also be used in modeling interest rates and bond models (Sun, 2005). The former of which was side-stepped in this thesis.

The idea of a latent process could also be taken further to actual process switching as opposed to simply parameter switching. I could not find any literature on this, but the idea is intriguing. The change from a GBM process to, for example, a mean-reverting process (such as an OU process) could help possibly explain some of the behavioural phenomenon discussed in Section 1.4.

Ultimately, the application of these models opens up many doors to future research and the hope is that this thesis provides the basic tools for the creation and estimation of MSM models for use in option pricing and beyond.

REFERENCES

Alexander, C., & Lazar, E. (2005). Asymmetries and Volatility Regimes in the European Equity Markets. *ICMA Centre Discussion Papers in Finance, No 14*.

Anderson, M. (2006). Option Pricing using Hidden Markov Models. *Msc Thesis*.

Anderson, T. G., & Sorenson, B. E. (1996). Generalized Methods of Moments estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business & Economic Statistics*, 328-352.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics, Vol 49*, 307-343.

Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A maximisation technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist, Vol 41, No. 1*, 164-171.

Bicego, M., Murino, V., & Figeuiredo, M. A. (2003). A sequential pruning strategy for the selection of the number of states in Hidden Markov Models. *Pattern Recognition Letters, 24*, 1395-1407.

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy*, 637-654.

Bollen, N. P. (1998). Valuing Options in Regime-Switching Models. *Journal of Derivatives, Vol 6*, 38-49.

Borsboom, D., Mellenberg, G. J., & Van Heeden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review, Vol 110, No 2*, 203-219.

Boyle, P., & Draviam, T. (2007). Pricing exotic options under regime switching. *Insurance: Mathematics and Economics, Vol 40*, 267-282.

CBOE. (2011, November 3). *Introduction to VIX Options and Futures*. Retrieved November 3, 2011, from CBOE: http://www.cboe.com/micro/VIX/vixintro.aspx

Chambaz, A., Garivier, A., & Gassiat, E. (2005). A MDL approach to HMM with Poisson and Gaussian emissions. Application to order identification.

Chen, P., Yang, H., & Yin, G. (2008). Markowitz's mean-variance asset-liability management with regime-switching: A continious-time model. *Insurance: Mathematics and Economics, Vol 43*, 456-465.

Cont, R. (2001). Emperical properties of asset returns: stylised facts and statistical issues. *Quantitative Finance, Vol 1*, 223-236.

Cox, J. C., Ross, S. A., & Rubinstein, M. (1979). Option Pricing: A simplified Approach. *Journal of Financial Economics*.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 1-38.

REFERENCES

Derman, E. (1999). *Regimes of Volatility.* Goldman Sachs: Quantitative Strategies Research Notes.

Diebold, F. X., Lee, J.-H., & Wienbach, G. C. (1994). Regime switching with time-varying transition probabilities. *Advanced texts in economics*, 283-302.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical terms.* ISBN 0-19-850994-4.

Elliot, R. J., Chan, L., & Kuen Sui, T. (2005). Option Pricing and Esscher transform under regime switching. *Annals of Finance*, 423-432.

Elliot, R. J., Krishnamurthy, V., & Sass, J. (2008). Moment based regression algorithms for drift and volatility estimation in continious-time Markov switching models. *Econometrics Journal, Vol 1*, 1-26.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Emperical Work. *Journal of Finance, Vol 25, No 2*, 383-417.

Figueiredo, M. A., Leitao, J. M., & Jain, A. K. (1999). On fitting mixture models. *Energy minimisation methods in in Computer vision and pattern recognition*, 54-69.

Follmer, H., & Schweizer, M. (1991). Hedging of Contingent Claims under Incomplete Information. *Applied Stochastic Analysis, Stochastics Monographs, vol 5*, 389-414.

Franq, C., & Roussignol, M. (1997). On White Noise Driven by Hidden Markov Chains. *Journal of Time Series Analysis, 18*, 553-578.

Fritelli, M. (2000). The Minimal Entropy Martingale Measure and the Valuation Problem in Incomplete Markets. *Mathematical Finance, Vol 10, No1*, 39-52.

Gerber, H. U., & Shiu, E. S. (1994). Martingale Approach to Pricing Perpetual American Options. *Astin Bulletin, Vol 24, No 2*, 195-220.

Gerber, H. U., & Shiu, E. S. (1994). Option Pricing by Esscher Transforms. *Transactions of the Society of Actuaries, Vol 46*, 99-191.

Guidici, P., Ryden, T., & Vandekerkhove, P. (2000). Likelihood-Ratio Tests for Hidden Markov Models. *Biometrics, Vol 56, No 3*, 742-747.

Guo, R. (2006). *Markov Processes - An Introduction and Application.* Cape Town: University of Cape Town.

Guo, X. (1999). Information and Option Pricings. *Ph.D dissertation*.

Hahn, M., & Sass, J. (2009). Parameter Estimation in Continuous Time Markov Switching Models: A semi-Continious Markov Chain Monte Carlo Approach. *Bayesian Analysis*, 63-84.

Hahn, M., Fruhwirth-Schnatter, S., & Sass, J. (2007). Markov chain Monte-Carlo methods for parameter estimation in multidimensional continious time Markov switching models. *RICAM Report*.

REFERENCES

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica, Vol 57*, 357-484.

Hansen, L. P. (1982). Large Sample Properties of Generalised Methods of Moments Estimators. *Econometrica 50*, 1029-1054.

Harrison, M., & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their application 11*, 215-260.

Harrison, M., & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their application 11*, 215-260.

Harte, D. (2006). Mathematical Background Notes for Package "Hidden Markov". 2-9.

Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies, Vol 6, No 2*, 327-343.

Ho, M. S., Perraudin, W. R., & Sorenson, B. E. (1996). A continuous-time arbitrage pricing model with stochastic volatility and jumps. *Journal of Business and Economic Statistics, 14*, 31-43.

Israel, R. B., Rosenthal, J. S., & Wei, J. Z. (2001). Finding generators for Markov chains via emperical transition matrices, with applications to credit ratings. *Mathematical Finance*, 245-265.

Kuen Siu, T., & Yang, H. (2009). Option Pricing When the Regime-Switching Risk is Priced. *Acta Mathematicae, Vol 25, No 3*, 369-388.

Li, X., & Hao, Q. (2005). Determine the number of regimes in Markov Switching Model. *China-USA Business Review, Vol 4, No 8*, 80-81.

Liu, N., Davis, R., Lovell, B., & Kootsookos, P. (2004). Effect of Initial HMM choices in multiple sequence training for gesture recongition. *Information Technology: Coding and Computing, Vol 1*, 5-7.

MacKay, R. J. (2002). Estimating the Order of a Hidden Markov Model. *The Canadian Journal of Statistics, Vol 30, No 4*, 573-589.

Magnusson, M. A., & Wydick, B. (2002). How Efficient are Africa's emerging stock markets. *Journal of Development Studies*, 141-156.

Mitra, S. (2010). Regime Switching Volatility Calibration by the Baum-Welch Method.

Miyahara, Y. (2001). Geometric Levy process and MEMM: pricing model and related estimation problems. *Asia-Pacific Financial Markets*, 45-60.

Mlambo, C., & Biekpe, N. (2007). The Efficient Market Hypothesis. *Investment Analysts Journal, No 66*, 5-17.

REFERENCES

Neal, M. R. (1991). *Bayesian Mixture Modeling by Monte Carlo Simulation.* Toronto: Department of Computer Science, University of Toronto.

Oliver, J. J., Baxter, R. A., & Wallace, S. C. (1996). Unsupervised Learning Using MML. *Thirtheenth International Conference on Machine Learning* (pp. 364-372). San Francisco: Morgan Kaufman.

Otranto, E., & Gallo, G. M. (2001). *A Nonparametric Bayesian Approach to Detect the Number of Regimes in a Markov Switching Model.* Rome: Universita degli Studi di Firenze.

Pham, H. (2000). On quadratic hedging in continuous time. *Mathematical Methods of Operational Research, 51*, 315-339.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 257-286.

Richardson, S., & Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society. Series B, Vol 4, No 4*, 731-792.

Rios, R. (2008). Penalized estimate of the number of states in Gaussian linear AR with Markov regime. *Electronic Journal of Statistics, Vol 2*, 1111-1128.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica, Vol 14*, 465-471.

Schweizer, M. (1992). Risk Options Simplified. *International Journal of Theoretical and Applied Finance 2*, 59-82.

Sericola, B. (2000). Occupation times in Markov Processes. *Communications in Statistics - Stochastic Models, Vol 16, No 5*, 1-33.

Shleifer, A. (2000). *Inefficient Markets.* Oxford: Oxford University Press.

Simons, D., & Laryea, S. A. (2004). *Testing the Efficiency of Selected South African Stock Markets.* Canada.

Siu, T. K., Ching, W. K., Fung, E., Ng, M., & Li, X. (2009). A high-order Markov-Switching model for risk measurement. *Computers and Mathematics with Applications, Vol 58*, 1-10.

Siu, T. K., Fung, E. S., & Ng, K. M. (2008). Option Valuation with a Discrete-Time Double Markovian Regime-Switching Model. *Applied Mathematical Finance, Vol 0, Issue 0 (ahead of print)*, 1-18.

Smith, D. R. (2002). Markov-Switching and Stochastic Volatility Diffusion Models of Short-Term Interest Rates. *Journal of Business & Economic Statistics, Vol 20, No 2*, 183-198.

Sun, L. (2005). Regime shifts in interest rate volatility. *Journal of Emperical Finance, Vol 12*, 418-434.

REFERENCES

Titterington, D. M., Smith, A. F., & Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester, UK: Wiley.

Vigfusson, R. (1997). Switching Between Chartists and Fundamentalists: A Markov Regime-Switching Approach. *Journal of Financial Economics, 2*, 291-305.

Wainscott, C. B. (1990). The Stock-Bond Correlation and its implications for Asset Allocations. *Financial Analysts Journal*, 55-60.

Yin, G., Song, Q. S., & Zhang, Z. (2005). Numerical solutions for jump-diffusions with regime switching. *Stochastics: An International Journal of Probability and Stochastic Processes, Vol 77, No 1*, 61-79.

Yu, S.-Z. (2010). Hidden semi-Markov Models. *Artifical Intelligence, Vol 174*, 215-243.