Regime Switching Volatility Calibration by the Baum-Welch Method

Abstract

Regime switching volatility models provide a tractable method of modelling stochastic volatility. Currently the most popular method of regime switching calibration is the Hamilton filter. We propose using the Baum-Welch algorithm, an established technique from Engineering, to calibrate regime switching models instead. We demonstrate the Baum-Welch algorithm and discuss the significant advantages that it provides compared to the Hamilton filter. We provide computational results of calibrating and comparing the performance of the Baum-Welch and the Hamilton filter to S&P 500 and Nikkei 225 data, examining their performance in and out of sample.

Key words: Regime switching, stochastic volatility, calibration, Hamilton filter, Baum-Welch.

1. Introduction and Outline

Regime switching (also known as hidden Markov models (HMM)) volatility models provide a tractable method of modelling stochastic volatility. Currently the most popular method of regime switching calibration is the Hamilton filter. However, regime switching calibration has been tackled in engineering (particularly for speech processing) for some time using the Baum-Welch algorithm (BW), where it is the most popular and standard method of HMM calibration. A review of the Baum-Welch algorithm can be found in [Lev05],[JR91]. The BW algorithm is increasingly being applied beyond engineering applications (for instance in bioinformatics [BEDE04]) but has been hardly applied to financial modelling, especially to regime switching stochastic volatility models.

Unlike the Hamilton filter, the BW algorithm is capable of determining the entire set of HMM parameters from a sequence of observation data. Furthermore, BW is a complete estimation method since it also provides the required optimisation method to determine the parameters by the maximum likelihood method (MLE).

The outline of the paper is as follows. Firstly, we introduce regime switching volatility models and the Hamilton filter. In the next section we introduce the Baum-Welch method, describing the algorithm not just for univariate but also multivariate Gaussian mixture observations. We then conduct numerical experiments to verify the Baum-Welch capability to detect regimes: we run numerical experiments on the S&P 500 and Nikkei 225 indexes and compare its performance against the Hamilton filter. We finally end with a conclusion.

2. Regime Switching Volatility Model and Calibration

2.1. Regime Switching Volatility

Wiener process driven stochastic volatility models capture price and volatility dynamics more successfully compared to other volatility models. Specifically, such models successfully capture the short term volatility dynamics; for a review on volatility models the reader is referred to [Mit09a]. However, for longer term dynamics and fundamental economic changes (e.g. "credit crunch"), no mechanism existed to address the change in volatility dynamics and it has been empirically shown that volatility is related to long term and fundamental conditions. Bekaert in [BHL06] claims that volatility changes are caused by economic reforms, for example on Black Wednesday the pound sterling was withdrawn from the ERM (European Exchange Rate Mechanism), causing a sudden change in value of the pound sterling [BR02]. Schwert [Sch89] empirically shows that volatility increases during financial crises.

A class of models that address fundamental and long term volatility modelling is the regime switching model (or hidden Markov model) e.g. as discusssed in [Tim00], [EvdH97]. In fact, Schwert suggests in [Sch89] that volatility changes during the Great Depression can be accounted for by a regime change such as in Hamilton's regime switching model [Ham89]. Regime switching is considered a tractable method of modelling price dynamics and does not violate Fama's "Efficient Market Hypothesis" [Fam65], which claims that price processes must be Markov process. Hamilton [Ham89] was the first to introduce regime switching models, which was applied to specifically model fundamental economic changes.

For regime switching models, generally the return distribution rather than the continuous time process is specified. A typical example of a regime switching model is Hardy's model [Har01]:

$$log((X(t+1)/X(t))|i) \sim \mathcal{N}(u_i, \varphi_i), i \in \{1, .., R\}, \tag{1}$$

where

- φ_i and u_i are constant for the duration of the regime;
- i denotes the current regime (also called the Markov state or hidden Markov state);
- R denotes the total number of regimes;
- transition probability matrix **A** is specified.

For Hardy's model the regime changes discretely in monthly time steps but stochastically, according to a Markov process.

Due to the ability of regime switching models to capture long term and fundamental changes, regime switching models are primarily focussed on modelling the long term behaviour, rather than the continuous time dynamics. Therefore regime switching models switch regimes over time periods of months, rather than switching in continuous time. Examples of regime switching models that model dynamics over shorter time periods are Valls-Pereira et al.[VPHS04], who propose a regime switching GARCH process, while Hamilton and Susmel [HS94] give a regime switching ARCH process. Note that economic variables other than stock returns, such as inflation, can also be modelled using regime switching models.

Regime switching has been developed by various researchers. For example, Kim and Yoo [KY95] develop a multivariate regime switching model for coincident economic indicators. Honda [Hon03] determines the optimal portfolio choice in terms of utility for assets following GBM but with continuous time regime switching mean returns. Alexander and Kaeck [AK08] apply regime switching to credit default swap spreads, Durland and McCurdy [DM94] propose a model with a transition matrix that specifies state durations. Mitra [Mit09b] applies regime switching to option pricing purely to capture the influence of long term dynamics (e.g. economic cycles) upon option prices.

The theory of Markov models (MM) and Hidden Markov models (HMM) are methods of mathematically modelling time varying dynamics of certain statistical processes, requiring a weak set of assumptions yet allow us to deduce a significant number of properties. MM and HMM model a stochastic process (or any system) as a set of states with each state possessing a set of signals or observations. The models have been used in diverse applications such as economics [SSS02], queuing theory [SF06], engineering [TG01] and biological modelling [MGPG06]. Following Taylor [TK84] we define a Markov model:

Definition 1. *A Markov model is a stochastic process $X(t)$ with a countable set of states and possesses the Markov property:*

$$p(q_{t+1} = j \mid q_1, q_2, .., q_t = i) = p(q_{t+1} = j \mid q_t = i), \tag{2}$$

where

- q_t *is the Markov state (or regime) at time t of $X(t)$;*
- i *and j are specific Markov states.*

As time passes the process may remain or change to another state (known as state transition). The state transition probability matrix (also known as the *transition kernel or stochastic matrix*) $\mathbf{A}$, with elements a_{ij}, tells us the probability of the process changing to state j given that we are now in state i, that is $a_{ij} = p(q_{t+1} = j \mid q_t = i)$. Note that a_{ij} is subject to the standard probability constraints:

$$0 \leq a_{ij} \leq 1, \forall i, j, \tag{3}$$

$$\sum_{j=1}^{\infty} a_{ij} = 1, \forall i. \tag{4}$$

We assume that all probabilities are stationary in time. From the definition of a MM the following proposition follows:

Proposition 1. *A Markov model is completely defined once the following parameters are known:*

- R*, the total number of regimes or (hidden) states;*
- *state transition probability matrix $\mathbf{A}$ of size $R \times R$. Each element is $a_{ij} = p(q_{t+1} = j|q_t = i)$, where i refers to the matrix row number and j to the column number of $\mathbf{A}$;*
- *initial (t=1) state probabilities $\pi_i = p(q_1 = i), \forall i$.*

A hidden Markov model is simply a Markov model where we assume that (as a modeller) we do not observe the Markov states. Instead of observing the Markov states (as in standard Markov models) we detect observations or time series data where each observation is assumed to be a function of the hidden Markov state, thus enabling statistical inferences about the HMM. Note that in a HMM it is the states which must be governed by a Markov process, not the observations and throughout we will assume one observation occurs after one state transition.

Proposition 2. *A hidden Markov model is fully defined when the parameter set $\{\mathbf{A}, \mathbf{B}, \pi\}$ are known:*

- R*, the total number of (hidden) states or regimes;*

- ***A***, *the (hidden) state transition matrix of size* $R \times R$*. Each element is* $a_{ij} = p(q_{t+1} = j|q_t = i)$*;*
- *initial (t=1) state probabilities* $\pi_i = p(q_1 = i), \forall i$*;*
- ***B***, *the observation matrix, where each entry is* $b_j(O_t) = p(O_t|j)$ *for observation* O_t*. For* $b_j(O_t)$ *is typically defined to follow some continuous distribution e.g.* $b_j(O_t) \sim \mathcal{N}(u_j, \varphi_j)$*.*

2.2. Current Calibration Method: Hamilton Filter

In financial mathematics or economic literature the standard calibration method for regime switching models is the Hamilton filter [Ham89], which works by maximum likelihood estimation (MLE). MLE is a method of estimating a set of parameters of a statistical model (Θ) given some time series or empirical observations $O_1, O_2, ..., O_T$. MLE determines Θ by firstly determining the likelihood function $\mathcal{L}(\Theta)$, then maximising $\mathcal{L}(\Theta)$ by varying Θ through a search or an optimisation method.

A statistical model with known parameter values can determine the probability of an observation sequence $O = O_1 O_2 ... O_T$. MLE does the opposite; we numerically maximise the parameter values of our model Θ such that we maximise the probability of the observation sequence $O = O_1 O_2 ... O_T$. To achieve this the MLE method makes two assumptions:

1. In maximising $\mathcal{L}(\Theta)$ the local optimum is also the global optimum (although this is generally not true in reality). The optimal values for Θ are in a search space of the same dimensions as Θ. Hamilton in [Ham94] gives a survey of various MLE maximisation techniques such as the Newton-Raphson method;
2. The observations $O_1, O_2, ..., O_T$ are statistically independent. Note that for Markov models we assume the conditional observations $(O_t|O_{t-1}), (O_{t-1}|O_{t-2}), (O_{t-2}|O_{t-3})...$ are independent.

For a regime switching process the general likelihood function $\mathcal{L}(\Theta)$ is:

$$\begin{aligned} \mathcal{L}(\Theta) &= f(O_1|\Theta) f(O_2|\Theta, O_1) f(O_3|\Theta, O_1, O_2) \\ &\cdots f(O_T|\Theta, O_1, O_2, ..., O_{T-1}), \end{aligned}$$

where $f(O_{(.)}|\Theta)$ is the probability of $O_{(.)}$, given model parameters Θ. Now by properties of logarithms we have:

$$\begin{aligned} log(\mathcal{L}(\Theta)) &= log(f(O_1|\Theta)) + log(f(O_2|\Theta, O_1)) + \cdots && (5) \\ &+ log(f(O_T|\Theta, O_1, O_2, ..., O_{T-1})). && (6) \end{aligned}$$

Hamilton proposes a likelihood function for regime switching models, which we refer to as the Hamilton filter. As an example, if we assume we have a two regime model with each regime having a lognormal return distribution, we wish to determine parameters $\Theta = \{u_1, u_2, \varphi_1, \varphi_2, a_{12}, a_{21}\}$. Note that in this simple HMM $a_{22} = 1 - a_{12}$ and $a_{11} = 1 - a_{21}$ therefore we do not need to estimate a_{22}, a_{11} in Θ.

To obtain $f(O_t|\Theta)$ in equation (6) for $t > 1$, Hamilton showed that it could be calculated by a recursive filter. We observe the relation:

$$f(O_t|\Theta, O_1, O_2, ..., O_{t-1}) = \sum_{q_t=1}^{2} \sum_{q_{t-1}=1}^{2} f(q_t, q_{t-1}, O_t|\Theta, O_1, ..., O_{t-1}). \tag{7}$$

Now using the relation:[1]

$$p(O, Q|\Theta) = p(O|\Theta, Q)p(Q|\Theta), \tag{8}$$

where $Q = q_1 q_2 ...$ represents some arbitrary state sequence, we make the substitution $f(q_t, q_{t-1}, O_t|\Theta, O_1, ..., O_{t-1})$

$$= p(q_{t-1}|\Theta, O_1, ..., O_{t-1}) \times p(q_t|q_{t-1}, \Theta) \times f(O_t|q_t, \Theta). \tag{9}$$

Therefore
$f(O_t|\Theta, O_1, O_2, ..., O_{t-1})$

$$= \sum_{q_t=1}^{2} \sum_{q_{t-1}=1}^{2} p(q_{t-1}|\Theta, O_1, ..., O_{t-1}) \times p(q_t|q_{t-1}, \Theta) \times f(O_t|q_t, \Theta), \tag{10}$$

where

- $p(q_t|q_{t-1}, \Theta) = p(q_t = j|q_{t-1} = i, \Theta)$ represents the transition probability a_{ij} we wish to estimate;
- $f(O_t|q_t = i, \Theta) = p_i(O_t)$ where $p_i(\cdot) \sim \mathcal{N}(u_i, \varphi_i)$ the Gaussian probability density function for state i, whose parameters u_i, φ_i we wish to estimate.

The parameters $\Theta = \{u_1, u_2, \varphi_1, \varphi_2, a_{12}, a_{21}\}$ are obtained by maximising the likelihood function using some chosen search method.

To calculate $f(O_t|\Theta, O_1, O_2, ..., O_{t-1})$ we require the probability $p(q_{t-1}|\Theta, O_1, O_2, ..., O_{t-1})$ in equation (10) (summed over two different values of q_{t-1} in

[1] Note: following discussions with Prof.Rabiner [Rab08] on the equation for $p(O, Q|\Theta)$ it was concluded that the equation for $p(O, Q|\Theta)$ in Rabiner's paper [Rab89] is incorrect.

the summations in equation (10)). This can be achieved through recursion, that is the probability $p(q_{t-1}|\Theta, O_1, .., O_{t-1})$ can be obtained from $p(q_{t-2}|\Theta, O_1, .., O_{t-2})$:

$$p(q_{t-1}|\Theta, O_1, O_2, ..., O_{t-1}) = \frac{\left(\sum_{i=1}^{2} f(q_{t-1}, q_{t-2} = i, O_{t-1}|\Theta, O_1, ..., O_{t-2})\right)}{f(O_{t-1}|\Theta, O_1, ..., O_{t-2})}. \tag{11}$$

The denominator of equation (11) is obtained from the previous period of $f(O_t|\Theta, O_1, O_2, ..., O_{t-1})$ (in other words $f(O_{t-1}|\Theta, O_1, O_2, ..., O_{t-2})$) so by inspecting equation (10) we can see it is a function of $p(q_{t-2}|\Theta, O_1, ..., O_{t-2})$. The numerator of equation (11) is obtained from calculating equation (9) for the previous time period, which is also a function of $p(q_{t-2}|\Theta, O_1, ..., O_{t-2})$.

To start the recursion of equation (11) at $p(q_1 = i|O_1, \Theta)$ we require $f(O_1|\Theta)$. Hamilton assumes the Markov chain has been running sufficiently long enough so that we can make the following assumption about our observations $O_1, O_2, ..., O_T$. Technically, Hamilton assumes the observations $O_1, O_2, ..., O_T$ are all drawn from the Markov chain's invariant distribution. If a Markov chain has been running for a sufficiently long time, the following property of Markov chains can be applied:

$$\eta_j = \lim_{t\to\infty} p(q_t = j|q_1 = i), \qquad \forall i, j = 1, 2, .., R, \tag{12}$$

$$\text{where} \sum_{j=1}^{R} \eta_j = 1, \eta_j > 0. \tag{13}$$

The probability η_j tells us in the long run ($t \to \infty$) the (unconditional) probability of being in state j and this probability is independent of the initial state (at time t=1). An important interpretation of η_j is as the fraction of time spent in state j in the long run. Therefore the probability of state j is simply η_j and so:

$$f(O_1|\Theta) = f(q_1 = 1, O_1|\Theta) + f(q_1 = 2, O_1|\Theta), \tag{14}$$

$$\text{where } f(q_1 = i, O_1|\Theta) = \eta_i p_i(O_1). \tag{15}$$

We can therefore calculate $p(q_1 = i|O_1, \Theta)$:

$$p(q_1 = i|O_1, \Theta) = \frac{f(q_1 = i, O_1|\Theta)}{f(O_1|\Theta)}, \tag{16}$$

$$= \frac{\eta_i p_i(O_1)}{\eta_1 p_1(O_1) + \eta_2 p_2(O_1)}. \tag{17}$$

Furthermore it can be proved for a two state HMM that:

$$\eta_1 = a_{21}/(a_{12} + a_{21}),$$

$$\eta_2 = 1 - \eta_1.$$

Therefore $p(q_1 = i|O_1, \Theta)$ can be obtained from estimating the parameter set $\Theta = \{u_1, u_2, \varphi_1, \varphi_2, a_{12}, a_{21}\}$, which is obtained by a chosen search method.

The advantages of Hamilton's filter method are firstly we do not need to specify or determine the initial probabilities, therefore there are fewer parameters to estimate (compared to the alternative Baum-Welch method). Therefore the MLE parameter optimisation will be over a lower dimension search space. Secondly, the MLE equation is simpler to understand and so easier to implement compared to other calibration methods. The disadvantages of the Hamilton filter will be discussed in the next section and these shortcomings will be addressed by the Baum-Welch algorithm.

3. Baum-Welch Algorithm

The Baum-Welch (BW) is a complete estimation method since it also provides the required optimisation method to determine the parameters by MLE. We will now explain the BW algorithm and to do so we must first explain the forward algorithm, which we will do now.

3.1. Forward Algorithm

The forward algorithm calculates $p(O|M)$, the probability of a fixed or observed sequence O=$O_1O_2...O_T$, given all the HMM parameters denoted by $M = \{\mathbf{A}, \mathbf{B}, \pi\}$. We recall from the definition of HMM that the probability of each observation $p(O_t)$ will change depending on the state at time t (q_t). Hence the most straightforward way to calculate $p(O|M)$ is:

$$\begin{aligned} p(O|M) &= \sum_{\text{all Q}} p(O, Q|M), && (18)\\ &= \sum_{\text{all Q}} p(O|M, Q).p(Q|M), && (19)\\ &= \sum_{\text{all Q}} \pi_{q_1} b_{q_1}(O_1).a_{q_1 q_2} b_{q_2}(O_2)....a_{q_{T-1} q_T} b_{q_T}(O_T), && (20)\\ \text{where } p(O|M, Q) &= b_{q_1}(O_1).b_{q_2}(O_2).....b_{q_T}(O_T). && (21) \end{aligned}$$

Here "all Q" means all possible state sequences $q_1q_2...q_T$ that could account for observation sequence O, $b_{(.)}(O_{(.)})$ is defined in proposition 2, $p(O|M, Q)$ is the probability of the observed sequence O, given it is along one single state sequence $Q = q_1q_2...q_T$ and for HMM M. We must sum equation (20) over all possible Q state sequences, requiring R^T computations and so this is computationally infeasible even for small R and T.

To overcome the computational difficulty of calculating $p(O|M)$ in equation (20) we apply the forward algorithm, which uses recursion (dynamic programming). The forward algorithm only requires computations of the order R^2T and so is significantly faster than calculating equation (20) for large R and T.

Let us define the forward variable $\kappa_t(i)$:

$$\kappa_t(i) = p(O_1O_2...O_t, q_t = i|M). \tag{22}$$

Given the HMM M, $\kappa_t(i)$ is the probability of the joint observation upto time t of $O_1O_2...O_t$ and the state at time t is i i.e. $q_t = i$. If we can determine $\kappa_T(i)$ we can calculate $p(O|M)$ since:

$$p(O|M) = \sum_{i=1}^{R} \kappa_T(i). \tag{23}$$

Now $\kappa_{t+1}(j)$ can be expressed in terms of $\kappa_t(i)$, therefore we can calculate $\kappa_{t+1}(j)$ by recursion:

$$\kappa_{t+1}(j) = \left[\sum_{i=1}^{R} \kappa_t(i)a_{ij}\right] b_j(O_{t+1}), 1 \leq t \leq T-1. \tag{24}$$

The variable $\kappa_{t+1}(j)$ in equation (24) can be understood as follows: $\kappa_t(j)a_{ij}$ is the probability of the joint event $O_1....O_t$ is observed, the state at time t is i and state j is reached at time t+1. If we sum this probability over all R possible states for i, we get the probability of j at t+1 accompanied with all previous observations from $O_1O_2...O_t$ only. Thus to get $\kappa_{t+1}(j)$ we must multiply by $b_j(O_{t+1})$ so that we have all observations $O_1...O_{t+1}$.

Therefore the recursive algorithm is as follows:

1. Initialisation:

$$\kappa_1(i) = \pi_i b_i(O_1), 1 \leq i \leq R. \tag{25}$$

2. Recursion:

$$\kappa_{t+1}(j) = \left[\sum_{i=1}^{R} \kappa_t(i)a_{ij}\right] b_j(O_{t+1}), 1 \leq t \leq T-1. \tag{26}$$

3. Termination: t+1=T.
4. Final Output:

$$p(O|M) = \sum_{i=1}^{R} \kappa_T(i). \tag{27}$$

At t=1 no sequence exists but we initialise the recursion with π_i to determine $\kappa_1(i)$.

3.2. Baum-Welch Algorithm

Having explained the forward algorithm we can now explain the BW algorithm. Using observation sequence O, the BW algorithm iteratively calculates the HMM parameters $M = \{\mathbf{A}, \mathbf{B}, \pi\}$. Specifically, BW estimates $M = \{a_{ij}, b_j(\cdot), \pi_i\} \forall i, j$, denoted respectively by $\overline{M} = \{\overline{a}_{ij}, \overline{b}_j(\cdot), \overline{\pi}_i\}$, such that it maximises the likelihood of $p(O|\overline{M})$. No method of analytically finding the globally optimal $\overline{M}$ exists. However it has been theoretically proven BW is guaranteed to find the local optimum [Rab89].

Let us define $\psi_t(i, j)$:

$$\psi_t(i, j) = p(q_t = i, q_{t+1} = j \mid O, M). \tag{28}$$

The variable $\psi_t(i, j)$ is the probability of being in state i at time t and state j at time t+1, given the HMM parameters M and the observed observation sequence O. We can re-express $\psi_t(i, j)$ as:

$$\begin{aligned} \psi_t(i, j) &= p(q_t = i, q_{t+1} = j \mid O, M), && (29) \\ &= \frac{p(q_t = i, q_{t+1} = j, O \mid M)}{p(O|M)}. && (30) \end{aligned}$$

Now we can re-express equation (30) using the forward variable $\kappa_t(i) = p(O_1 O_2 ... O_t, q_t = i|M)$ and using analogously the so called backward variable $\varrho_{t+1}(i)$:

$$\varrho_t(i) = p(O_{t+1} O_{t+2} ... O_T | q_t = i, M), \tag{31}$$

$$\text{so that } \varrho_{t+1}(i) = p(O_{t+2} O_{t+3} ... O_T | q_{t+1} = i, M). \tag{32}$$

The backward variable $\varrho_t(i)$ is the probability of the partial observed observation sequence from time t+1 to the end T, given M and the state at time t is i. It is calculated in a similar recursive method to the forward variable using the backward algorithm (see [Rab89] for more details). Hence we can rewrite $\psi_t(i, j)$ as

$$\psi_t(i, j) = \frac{\kappa_t(i) a_{ij} b_j(O_{t+1}) \varrho_{t+1}(j)}{p(O|M)}. \tag{33}$$

We can also rewrite the denominator $p(O|M)$ in terms of the forward and backward variables, so that $\psi_t(i, j)$ is entirely expressed in terms of $\kappa_t(i), a_{ij}, b_j(O_{t+1}), \varrho_{t+1}(j)$:

$$p(O|M) = \sum_{i=1}^{R} \sum_{j=1}^{R} \kappa_t(i) a_{ij} b_j(O_{t+1}) \varrho_{t+1}(j). \tag{34}$$

Now let us define $\Gamma_t(i)$:

$$\begin{aligned} \Gamma_t(i) &= p(q_t = i|O, M), && (35) \\ &= \sum_{j=1}^{R} \psi_t(i, j). && (36) \end{aligned}$$

Equation (36) can be understood from the definition of $\psi_t(i,j)$ in equation (29); summing $\psi_t(i,j)$ over all j must give $p(q_t = i|O,M)$, the probability in state i at time t, given the observation sequence O and model M. Now if we sum $\Gamma_t(i)$ from t=1 to T-1 it gives us $\Upsilon(i)$, the expected number of transitions made from state i:

$$\Upsilon(i) = \sum_{t=1}^{T-1} \Gamma_t(i). \tag{37}$$

If we sum $\Gamma_t(i)$ from t=1 to T it gives us $\vartheta(i)$, the expected number of times state i is visited:

$$\vartheta(i) = \sum_{t=1}^{T} \Gamma_t(i). \tag{38}$$

We are now in a position to estimate $\overline{M}$. The variable $\overline{a}_{ij}$ is estimated as the expected number of transitions from state i to state j divided by the expected number of transitions from state i:

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \psi_t(i,j)}{\Upsilon(i)}. \tag{39}$$

The variable $\overline{\pi}_i$ is estimated as the expected number of times in state i at time t=1:

$$\overline{\pi}_i = \Gamma_1(i). \tag{40}$$

The variable $\overline{b}_j(\tilde{s})$ is estimated as the expected number of times in state j and observing a particular signal $\tilde{s}$, divided by the expected number of times in state j:

$$\overline{b}_j(\tilde{s}) = \frac{\displaystyle\sum_{t=1}^{T} \Gamma_t(j)'}{\vartheta(j)}, \tag{41}$$

where $\Gamma_t(j)'$ is $\Gamma_t(j)$ with condition $O_t = \tilde{s}$.

We can now describe our BW algorithm:

1. Initialisation:
   Input initial values of $\overline{M}$ (otherwise randomly initialise) and calculate $p(O|\overline{M})$ using the forward algorithm.
2. Estimate new values of $\overline{M}$:
   Iterate until convergence:
   (a) Using current $\overline{M}$ calculate variables $\overline{\kappa}_t(i), \overline{\varrho}_{t+1}(j)$ by the forward and backward algorithm and then calculate $\overline{\psi}_t(i,j)$ as in equation (33).
   (b) Using calculated $\overline{\psi}_t(i,j)$ in (a) determine new estimates of $\overline{M}$ using equations (36)-(41).

(c) Calculate $p(O|\overline{M})$ with new $\overline{M}$ values using the forward algorithm.

(d) Goto step 3 if two consecutive calculations of $p(O|\overline{M})$ are equal (or converge within a specified range). Otherwise repeat iterations: goto (a).

3. Output $\overline{M}$.

The BW algorithm is started with initial estimates of $\overline{M} = (\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\pi})$. These estimates in turn are used to calculate the right hand side of equations (39),(40) and (41) to give the next new estimate of $\overline{M} = (\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\pi})$. We consider the new estimate $\overline{M_n}$ to be a better estimate than the previous estimate $\overline{M_p}$, if $p(O|\overline{M_n}) > p(O|\overline{M_p})$, with both probabilities calculated via the forward algorithm. In other words, we prefer the $\overline{M}$ that increases the probability of observation O occurring.

If $p(O|\overline{M_n}) > p(O|\overline{M_p})$ then the iterative calculation is repeated with $\overline{M_n}$ as the input. Note that at the end of step two, if the algorithm re-iterates then inputting the new $\overline{M}$ at step 2a means we will get a new set of $\overline{M}$ after executing 2b. The iteration is stopped when $p(O|\overline{M_n}) = p(O|\overline{M_p})$ or is arbitrarily close enough and at this point the BW algorithm finishes.

Since the BW algorithm has been proven to always converge to the local optimum, the BW will output the local optimum. We also note that correct choice of R is important since $p(O|M)$ changes as M changes for a fixed O, however this disadvantage is common to all MLE methods.

3.3. Multivariate Gaussian Mixture Baum-Welch Calibration

To account for the variety of empirical distributions possible for various assets and capturing asymmetric properties arising from volatility (such as fat tails), we model each regime's distribution by a two component multivariate Gaussian mixture (GM), which is a mixture of two multinormal distributions.

Definition 2. *(Multinormal Distribution) Let $\mathbf{X} = (X_1, X_2..., X_n)$ be an n-dimensional random vector where each dimension is a random variable. Let $\boldsymbol{u}$=$(u_1, u_2..., u_n)$ represent an n dimensional vector of means, $\mathbf{\Sigma}$ represent an $n \times n$ covariance matrix. We say $\mathbf{X}$ follows a multinormal distribution if*

$$\mathbf{X} \sim \mathcal{N}_{\mathbf{n}}(\boldsymbol{u}, \mathbf{\Sigma}), \tag{42}$$

which may be alternatively written as

$$\begin{pmatrix} X_1 \\ X_{...} \\ X_n \end{pmatrix} \sim \mathcal{N}_n \left(\begin{pmatrix} u_1 \\ u_{...} \\ u_n \end{pmatrix}, \begin{pmatrix} \varphi_{11} & \varphi_{...} & \varphi_{1n} \\ \varphi_{...} & \varphi_{...} & \varphi_{...} \\ \varphi_{n1} & \varphi_{...} & \varphi_{nn} \end{pmatrix} \right). \tag{43}$$

The probability of $\mathbf{X}$ is

$$p(\mathbf{X}) = \frac{1}{2\pi\sqrt{det(\mathbf{\Sigma})}} exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{u})^T \mathbf{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{u})\right), \tag{44}$$

where $det(\mathbf{\Sigma})$ *denotes the determinant of* $\mathbf{\Sigma}$.

Definition 3. *(Multivariate Gaussian Mixture) A multivariate Gaussian mixture consists of a mixture of K multinormal distributions, spanning n-dimensions. It is defined by:*

$$\mathbf{X} \sim c_1\mathcal{N}_{\mathbf{n}}(\mathbf{u}_1, \mathbf{\Sigma}_1) + ... + c_K\mathcal{N}_{\mathbf{n}}(\mathbf{u}_K, \mathbf{\Sigma}_K), \tag{45}$$

where c_k *are weights and*

$$\sum_{k=1}^{k=K} c_k = 1, c_k \geq 0. \tag{46}$$

The term $p_{gmm}(\mathbf{X})$ *denotes the probability of a multivariate Gaussian mixture variable* $\mathbf{X}$ *and is defined as*

$$p_{gmm}(\mathbf{X}) \quad = \quad \sum_{k=1}^{K} c_k p_k(\mathbf{X}), \tag{47}$$

where $p_k(\mathbf{X}) \sim \mathcal{N}_{\mathbf{n}}(\mathbf{u}_k, \mathbf{\Sigma}_k)$.

If we model a stochastic process $\mathbf{X}$ by a Gaussian mixture for each regime then for a given regime j we have:

$$\mathbf{X} \quad \sim \quad c_{j1}\mathcal{N}_{\mathbf{n}}(\mathbf{u}_{j1}, \mathbf{\Sigma}_{j1}) + ... + c_{jK}\mathcal{N}_{\mathbf{n}}(\mathbf{u}_{jK}, \mathbf{\Sigma}_{jK}). \tag{48}$$

The probability of $\mathbf{X}$ for a given regime j, $p_{gmm}(\mathbf{X})_j$, is:

$$p_{gmm}(\mathbf{X})_j \quad = \quad \sum_{k=1}^{k=K} c_{jk} p_{jk}(\mathbf{X}). \tag{49}$$

where

- $p_{jk}(\mathbf{X}) \sim \mathcal{N}_{\mathbf{n}}(\mathbf{u}_{jk}, \mathbf{\Sigma}_{jk})$;
- c_{jk} are weights for each regime j and

$$\sum_{k=1}^{k=K} c_{jk} = 1, \forall j. \tag{50}$$

Note that the dimensions of multivariate distribution n are independent of the number of mixture components K.

For an n-asset portfolio $\mathbf{X(t)} = (X_1(t), X_2(t), ..., X_n(t))$, where $X_i(t)$ represents the stock price of asset i, with each asset following a Gaussian mixture, the portfolio returns would be modelled by:

$$d\mathbf{X}/\mathbf{X} \sim c_{j1}\mathcal{N}_{\mathbf{n}}(\mathbf{u}_{j1}, \mathbf{\Sigma}_{j1}) + c_{j2}\mathcal{N}_{\mathbf{n}}(\mathbf{u}_{j2}, \mathbf{\Sigma}_{j2}). \tag{51}$$

For practial calibration purposes we set the multivariate observation *vector* $\mathbf{O_t}$ to annual log returns:

$$\mathbf{O_t} = log(\mathbf{X}(t+\Delta t)/\mathbf{X}(t)), \tag{52}$$

where Δt=1 year.

Combining GM with HMM gives us a GM-HMM (Gaussian mixture HMM) model and the BW algorithm can be adapted to it: Gaussian mixture BW (GM-BW). For $\mathbf{O}_t$ our observation (vector) at time t we model $b_j(\mathbf{O})$ by GM:

$$b_j(\mathbf{O}) = p_{gmm}(\mathbf{O})_j. \tag{53}$$

The BW algorithm for calculating $\mathbf{A}, \pi_i$ remains the same; for $\mathbf{B}$ we have a GM. We would like to obtain the GM mixture coeffficents c_{jk}, mean vectors $\boldsymbol{u}_{jk}$ and covariance matrices $\boldsymbol{\Sigma}_{jk}$ whose estimates are $\overline{c}_{jk}$, $\overline{\boldsymbol{u}}_{jk}$ and $\overline{\boldsymbol{\Sigma}}_{jk}$ respectively. These can be incorporated within the BW algorithm as detailed by Rabiner [Rab89]:

$$\begin{aligned}
\overline{\boldsymbol{u}}_{jk} &= \frac{\sum_{t=1}^{T}\Gamma_t(j,k).\mathbf{O}_t}{\sum_{t=1}^{T}\Gamma_t(j,k)}, && (54)\\
\overline{\boldsymbol{\Sigma}}_{jk} &= \frac{\sum_{t=1}^{T}\Gamma_t(j,k).(\mathbf{O}_t-\overline{\boldsymbol{u}}_{jk})(\mathbf{O}_t-\overline{\boldsymbol{u}}_{jk})^T}{\sum_{t=1}^{T}\Gamma_t(j,k)}, && (55)\\
\overline{c}_{jk} &= \frac{\sum_{t=1}^{T}\Gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{K}\Gamma_t(j,k)}, && (56)\\
\text{where } \Gamma_t(j,k) &= \left[\frac{\kappa_t(j)\varrho_t(j)}{\sum_{j=1}^{N}\kappa_t(j)\varrho_t(j)}\right]\left[\frac{c_{jk}p_{jk}(\mathbf{O}_t)}{p_{gmm}(\mathbf{O}_t)_j}\right]. && (57)
\end{aligned}$$

Here $\Gamma_t(j,k)$ is the probability at time t of being in state j with the k mixture component accounting for $\mathbf{O}_t$. Using the same logic as in section 3 (for non mixture distributions) we can understand equations (54)-(56), for example $\overline{c}_{jk}$ is the expected number of times the HMM k-th component is in state j divided by the expected number of times in state j.

It is worth noting that a well known problem in maximum likelihood estimation of GM is that observations with low variances give extremely high likelihoods, in which case the likelihood function does not converge [MT07]. To overcome this problem in the univariate case Messina and Toscani [MT07] implement Ridolfi's and Idier's [RI02] penalised maximum likelihood function, which limits the likelihood value of observations. This is beneficial in [MT07] because the observation time scales are of the order of days and therefore the variance of samples may approach zero. For our applications we calibrate the GM-HMM to annual return data, therefore the samples are unlikely to approach variances anywhere near zero.

3.4. Advantages of Baum-Welch Calibration

The BW algorithm has significant advantages over the Hamilton filter. Firstly, the Hamilton filter requires observation data to be taken from the invariant distribution in order to estimate the parameters (see equation (12)). To obtain observations from the invariant distribution implies the number of state transitions approaches a large limit, so is not suited to Markov chains that have run for a short time. Furthermore, the time to reach the invariant distribution increases with the number of regimes R and the number of Gaussian mixtures K.

Psaradakis and Sola [PS98] investigated the finite sample properties of the Hamilton filter for financial data. They concluded that samples of at least 400 observations are required for a simple two state regime switching model where each state's observation is modelled by a normal distribution.

Secondly, the Hamilton filter has no method of estimating the initial state probabilities whereas the BW is able to take account of and estimate initial state probabilities. This has a number of important consequences:

1. BW does not require observations from the invariant distribution and so can be calibrated to data of any observation length.
2. the Hamilton filter cannot fully define the entire HMM model since the initial state probabilities are one of the key HMM parameters in the definition (see HMM definition in section 2).
3. we cannot determine the probability of observation sequences $p(O|M)$, since we require the initial state probabilities. This can be understood from the forward algorithm.
4. we cannot determine the most likely state sequence that accounts for a given observation sequence and HMM, which can be obtained by the Viterbi algorithm. The Viterbi algorithm tells us the most likely state sequence for a given observation sequence and HMM parameters M (see Forney [FJ73] for more information).
5. without the initial state probabilities, we cannot simulate state sequences since the initial state radically alters the state sequence and its influence on the state sequence increases as the sequence size decreases. Consequently we cannot validate a model's feasibility by simulation.

Note that BW estimates initial state probabilities independently of the transition probabilities, whereas in the Hamilton filter η_i is a function of estimated transition probabilities. Hence BW is able to independently estimate more HMM parameters than the Hamilton filter.

Thirdly, to our knowledge the Hamilton filter cannot be applied to multivariate distributions, nor more complicated univariate distributions than Gaussians. Particularly for financial applications, we use multivariate data to model portfolios and multivariate stochastic volatility is becoming an increasingly important research area (see Bauwens et al. [BLR06] for a survey on multivariate GARCH). Hamilton has proposed a calibration method for univariate mixture distributions, the Quasi-Bayesian MLE approach [Ham91], yet this requires some prior knowledge regarding the reliability of observations. The GM-BW calibrates a multivariate Gaussian mixture to multivariate data, thereby capable of modelling most empirically observed distributions.

Fourthly, the GM-BW can calibrate time varying correlations. It is known that correlations amongst random variables tend to be unstable with time; for example Buckley et al. [BSS07] give evidence of covariances varying with time and model them as regime dependent. The GM-BW algorithm gives the covariance matrix for each regime and each regime is postulated to be linked to an economic state. Therefore, we can model and extract information on changing correlations with changing economic conditions. For instance, some stocks are considered to be strongly correlated with the economic cycle (known as cyclical stocks) e.g. British Airways, whereas other stocks are considered independent of the economy (known as defensive stocks) e.g. Tesco.

Finally, the BW algorithm is a complete HMM estimation method whereas the Hamilton filter is not. Hamilton's method provides no method or guidance as to the optimisation algorithm to apply for finding the parameters from the non-linear filter, yet the solutions can be significantly influenced by the non-convex optimisation method applied. The BW algorithm includes an estimation method for the full HMM and a numerical optimisation scheme. Additionally, the BW method is guaranteed to find the local optimum.

4. Numerical Experiments: Baum-Welch GM-HMM Calibration Results

In this section we compare the calibration of a two state, annually regime switching model by the Hamilton filter and the GM-BW method. We calibrate the model to annual returns to two different markets, over the period 1976-96, with data obtained from Datastream. The two markets are the American S&P500 index and the Japanese Nikkei 225 index; these markets were chosen because their time dynamics differ significantly (one can view their empirical return history later in the section). Hence we can gain a better understanding of the calibration performance under different markets.

The GM-BW method calibrates a 2-state regime switching model, with 2 Gaussian components to represent the observations of each state. The calibrated model is therefore

$$dX/X \sim c_{j1}\mathcal{N}_1(u_{j1}, \varphi_{j1}) + c_{j2}\mathcal{N}_1(u_{j2}, \varphi_{j2}), j \in \{1, 2\}, \tag{58}$$

where X(t) is some asset price or index value. The (original) Hamilton filter calibrates to a two state regime switching model, with a lognormal distribution for each state:

$$log((X(t+1)/X(t))|j) \sim \mathcal{N}(u_j, \varphi_j), j \in \{1, 2\}. \tag{59}$$

Both regime switching models were calibrated to annual returns data from 1976-1996. We therefore define our set of observations of annual log returns as:

$$O_t = log(X(t + \Delta t)/X(t)), \tag{60}$$

where Δt=1 year and X(t) is the index value.

To further validate the quality of the regime switching calibration (in addition to the calibration results), we generated the state sequence for each market. The state sequences were obtained for the in sample period 1976-96 (that is the data period to which the model was calibrated) and the out of sample period 1997-2007. As mentioned in section 3.4 the GM-BW provides the full HMM and so can provide the state sequence associated with some observation data. The state sequence is obtained by applying the Viterbi algorithm to the GM-BW model and the observation data. Hamilton's filter does not provide the full HMM, hence it cannot truly give state sequences associated with some observation data. However, Hamilton provides a method of "inferring" state sequences from observation data [Ham94] and this was applied to our data.

4.1. Procedure

We calibrated the regime switching model to each market as follows, discussing the GM-BW method first. The basic GM-BW software implementations available are numerous due to GM-BW's wide usage in engineering. We chose K. Murphy's Matlab implementation [Mur08] because it is considered one of the most standard and cited GM-BW programs. It also offers many useful features that are unavailable on other implementations e.g. the Viterbi algorithm for obtaining state sequences.

Since the GM-BW algorithm only finds the local optimum GM-HMM parameters that maximise the likelihood of the observations, this had to be addressed because the search space is nonconvex. In other words, the GM-BW maximisation of the likelihood of the observations does not necessarily determine the globally optimum paramters on first usage. Commonly, the global optimal parameters are obtained by initialising the GM-BW algorithm over every possible starting point. However, for our model this would involve initialising over a nonconvex solution search space of thirteen dimensions; this is because the GM-BW finds the parameters $\overline{M} = (\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\pi})$ (where $\overline{\mathbf{B}}$ is parameterised by $\overline{c}_{jk}, \overline{\varphi}_{jk}$ and $\overline{u}_{jk}$).

Due to the high dimensionality of the nonconvex solution search space, initialising GM-BW at different starting points was not practical. Instead, we obtained our GM-BW solutions by initialising GM-BW from good initial parameter estimates, therefore the locally optimum estimates found by GM-BW should be close to the global optimum. It is also worth noting that initialisation strongly influences the GM-BW optimisation [LDLK04], hence this suggests a better optimisation method than initialising from every possible start point. We will now describe the initialisation for each GM-BW parameter:

$\bar{\mathbf{A}}$ *Initialisation*

We can initialise $\overline{\mathbf{A}}$ based on our expectations of the economic regimes that each state represents. We will now explain our initialisation:

$$\bar{\mathbf{A}} = \begin{pmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{pmatrix}. \tag{61}$$

If we assign state one as the up state of the economy, we know from economic behaviour the economy in the long term follows an upward drift, we would expect it is more likely the HMM remains in state one rather than goto state two, given it is already in state one. Hence we assign probability 0.6 and this also gives 0.4 because we have 1-0.6=0.4 by the total law of probability.

Similarly, we know from economic behaviour and data that an economy is more likely to return to the up state given it is already in the down state. Therefore if we assign state two as the down state of the economy then we would expect HMM to return to state one rather than state two. This also captures the cyclical behaviour of economies and the tendency to follow a long term upward trend. Hence we assign probabilities 0.7, which in turn gives 0.3 from the total law of probability.

$\overline{\pi}$ *Initialisation*

Similar to $\bar{\mathbf{A}}$ we can initialise $\overline{\pi}$ based on our economic expectations we expect the HMM to possess. However, since we know the first observation we can more accurately estimate its hidden state. Since positive returns are associated with the up state of the economy (which we assign as state one), then we assign a probability greater than 0.5 to state one if the first observed return is positive. Similarly, if the first observed return is negative we assign state two with a probability greater than 0.5.

GM Initialisation ($\overline{\mathbf{B}}$)
The GM initialisation strongly affects the GM-BW algorithm optimisation. However it is well known that GM distribution fitting in general (without any regime switching) is a non-trivial problem. This is because:

- there are a large set of parameters to estimate.
- there exists the issue of uniqueness, that is for a given non-parametric distribution there does not always exist a unique set of GM parameter values.
- the flexibility of GM distributions to model virtually any unimodal or bimodal distribution means that it incorporates rather than rejects any noise in the data into the distribution. Therefore GM fitting is highly sensitive to noise.
- parameter estimation is further complicated with regime switching and the fact we cannot identify with certainty the (hidden) state associated with each observation.

Rather than randomly initialise the GM parameters (as is done in Murphy's program) we initialised the GM parameters in the following way. Firstly, we divided the S & P 500 data into two sets: one containing positive returns and one set containing negative returns data. This gave the approximate distribution for each regime, since we would expect the majority of positive returns to belong to the up state (state one) and negative returns to the down state (state two). Next, we applied a GM fitting program to each "regime's" data from Lund University (stixbox [A.H00]). This provided initial GM estimates for each regime, which in turn were inputted into GM-BW for initialisation.

Once the GM-BW had been initialised the GM-HMM parameters could be obtained. The initial parameter estimates could be adjusted to determine if minor adjustments improved the calibration. However, it was found from experiments that minor adjustments still resulted in GM-BW converging to the same set of parameters.

The Hamilton filter has been implemented by the Society of Actuaries (SOA) by Dr.M.Hardy and has also been used in academic research on Hamilton filters e.g. [Har01]. As mentioned in section 3.4, the Hamilton filter does not provide an optimisation (nor an initialisation) method, however the SOA implementation provides a self-contained initialisation and optimisation method.

To obtain the state sequences for some observation data, the GM-BW calibrated model state sequences were obtained using the the Viterbi algorithm. This was provided within K.Murphy's GM-BW basic package as it is a popular tool in engineering. As mentioned previously, the Hamilton filter does not provide state sequence estimation nor does it estimate the all parameters to enable state sequence estimation to some data. However Hamilton provides a method of “state inference” for his calibrated models and this is also provided within the SOA implementation.

4.2. Results

We present the results of the calibration of our model by the GM-BW and Hamilton filter methods for each market (S&P 500 and Nikkei 225).

4.2.1. Parameter Calibration Results

S&P 500 Index

Table 1: GM-BW Calibration: Initial State Probabilities (π_i) for S&P 500

State (i)	Probability
1	1×10^{-6}
2	$1 - 1 \times 10^{-6}$

GM-BW Calibration: State Transition Matrix for S&P 500

$$\overline{\mathbf{A}} = \begin{pmatrix} 0.78 & 0.22 \\ 0.82 & 0.18 \end{pmatrix}$$

Table 2: GM-BW Calibration: Mixture Means u_{jk} (%/year) for S&P 500

Gaussian Component (k)	State (j)	
	1	2
$\mathcal{N}_1$	13.0	-4.8
$\mathcal{N}_2$	28.0	1.4
Overall	14.8	-4.7

Table 3: GM-BW Calibration: Mixture Standard Deviations $\sqrt{\varphi_{jk}}$ (%/year) for S&P 500

Gaussian Component (k)	State (j)	
	1	2
$\mathcal{N}_1$	4.5	5.6
$\mathcal{N}_2$	28.0	110.0
Overall	10.7	12.3

Table 4: GM-BW Calibration: Mixture Weighting Matrix (c_{jk}) for S&P 500

Gaussian Component (k)	State (j)	
	1	2
$\mathcal{N}_1$	0.88	0.99
$\mathcal{N}_2$	0.12	0.01

Hamilton Filter Calibration: State Transition Matrix for S&P 500

$$\overline{\mathbf{A}} = \begin{pmatrix} 0.60 & 0.40 \\ 0.03 & 0.97 \end{pmatrix}$$

Table 5: Hamilton Filter Calibration: Distribution Parameters for S&P 500

Distribution Parameter	State (j)	
	1	2
μ_j	-2.2	1.25
$\sqrt{\varphi_j}$	7.73	3.56

Nikkei 225 Index

Table 6: GM-BW Calibration: Initial State Probabilities (π_i) for Nikkei 225

State (i)	Probability
1	1×10^{-6}
2	$1 - 1 \times 10^{-6}$

GM-BW Calibration: State Transition Matrix for Nikkei 225

$$\overline{\mathbf{A}} = \begin{pmatrix} 0.87 & 0.13 \\ 0.41 & 0.59 \end{pmatrix}$$

Table 7: GM-BW Calibration: Mixture Means u_{jk} (%/year) for Nikkei 225

Gaussian Component (k)	State (j)	
	1	2
$\mathcal{N}_1$	14.8	-2.9
$\mathcal{N}_2$	18.6	-34.4
Overall	14.9	-18.6

Table 8: GM-BW Calibration: Mixture Standard Deviations $\sqrt{\varphi_{jk}}$ (%/year) for Nikkei 225

Gaussian Component (k)	State (j)	
	1	2
$\mathcal{N}_1$	10.2	0.5
$\mathcal{N}_2$	11.7	10.8
Overall	10.2	7.6

Table 9: GM-BW Calibration: Weighting Matrix (c_{jk}) for Nikkei 225

Gaussian Component (k)	State (j)	
	1	2
$\mathcal{N}_1$	0.98	0.5
$\mathcal{N}_2$	0.02	0.5

Table 10: Hamilton Filter Calibration: Distribution Parameters for Nikkei 225

Distribution Parameters	State (j)	
	1	2
μ_j	15.7	-10.7
$\sqrt{\varphi_j}$	10.7	20.0

Hamilton Filter Calibration: State Transition Matrix for Nikkei 225

$$\overline{\mathbf{A}} = \begin{pmatrix} 0.93 & 0.07 \\ 0.08 & 0.92 \end{pmatrix}$$

4.2.2. Regime Sequence Results

We present the state sequences generated by the GM-BW (using the Viterbi algorithm) and the Hamilton filter for the in-sample (1976-96) and out of sample (1997-07) periods and for each market. We also give the tables for each graph for additional reference.

S&P 500 Index

Figure 1: Graphs of the Empirical Annual Returns of the S&P 500 Index (1976-96) and the Associated Regimes Under GM-BW and Hamilton Filter Calibration

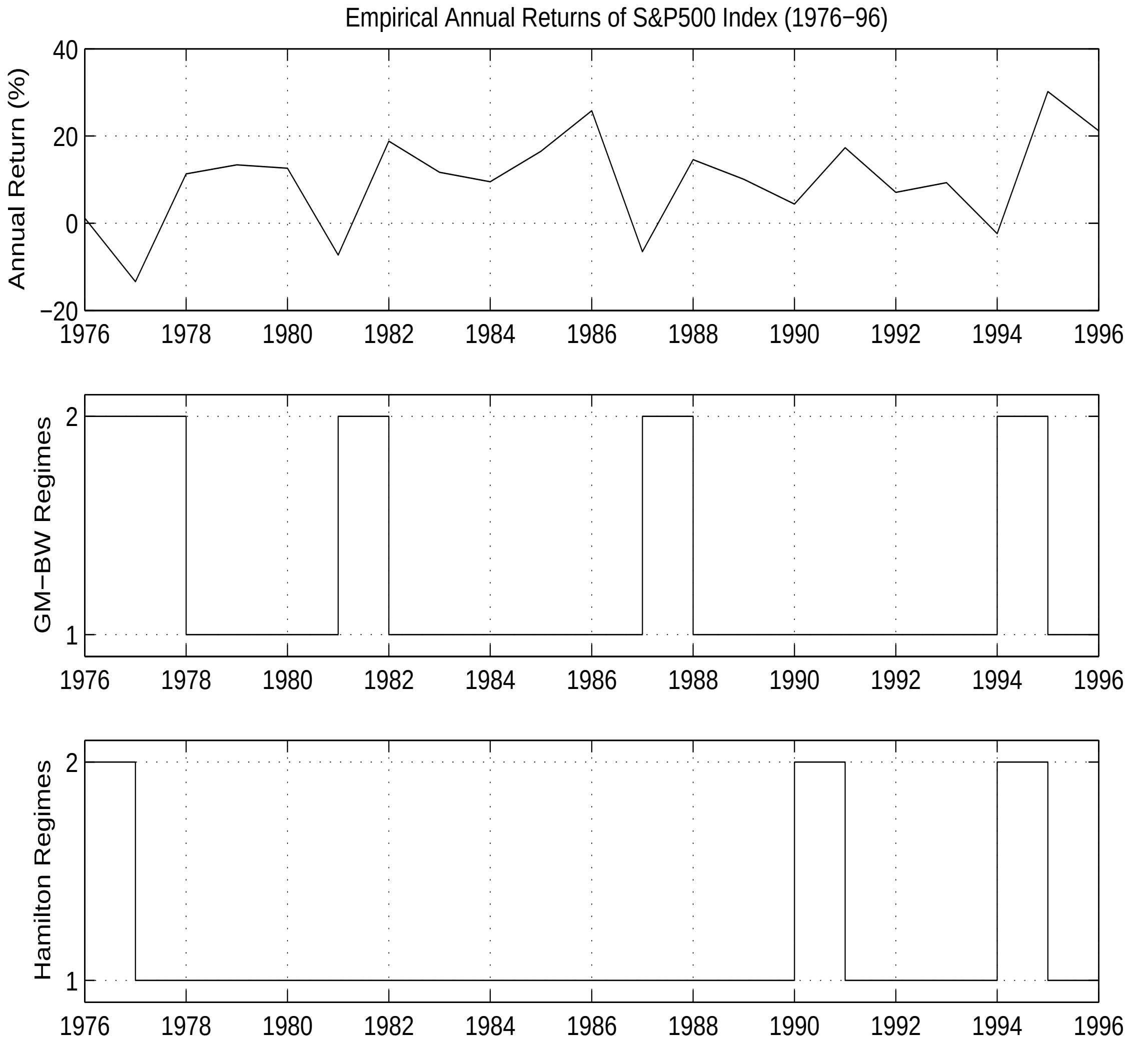

Figure 2: Graphs of the Empirical Annual Returns of the S&P 500 Index (1997-2007) and the Associated Regimes Under GM-BW and Hamilton Filter Calibration

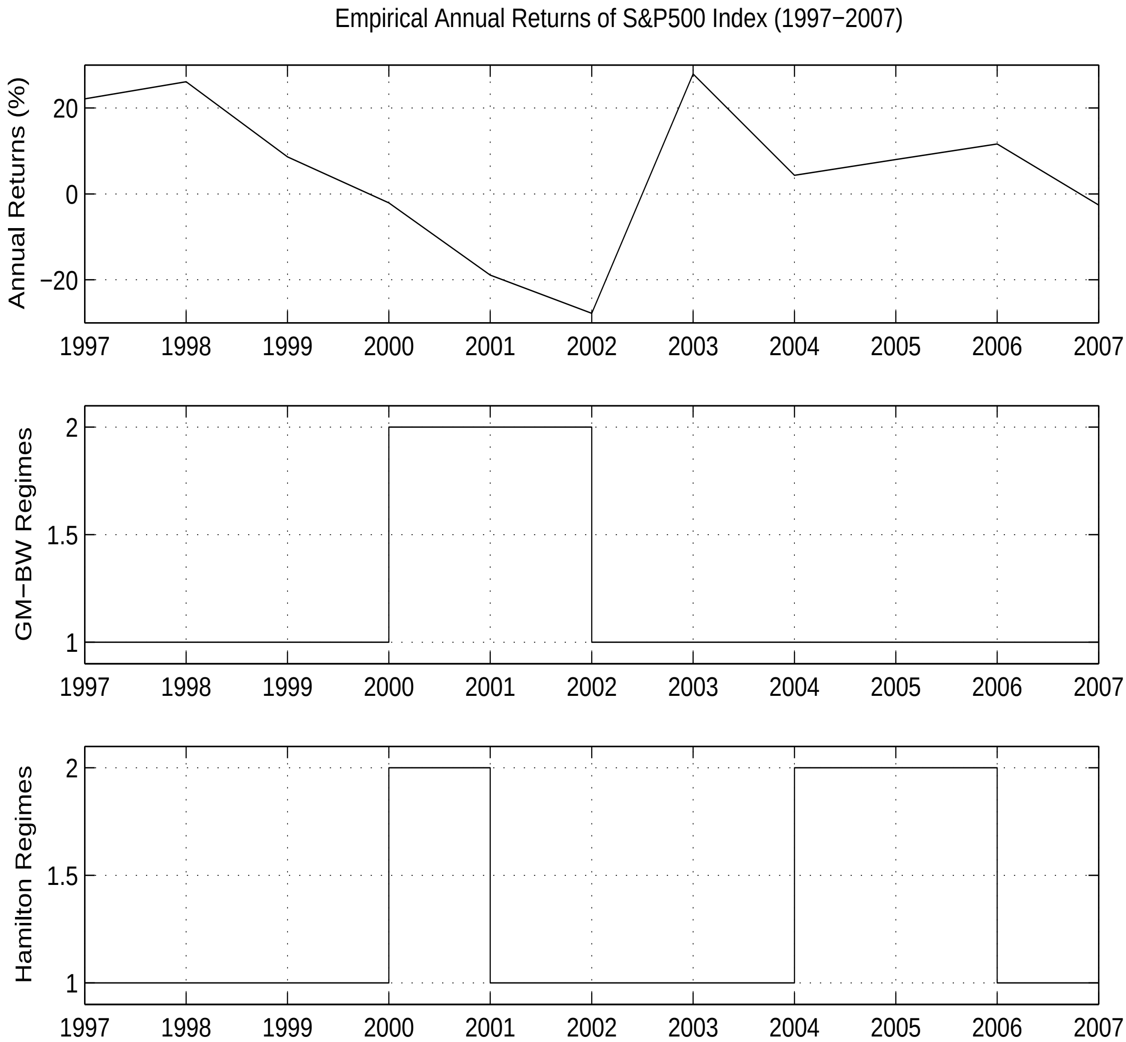

Table 11: S&P 500 Index: Regime Sequence Results for 1976-96 (In Sample)

Year	GM-BW Regime	Hamilton Regime	Empirical Annual Return (%)
1976	two	two	1.2
1977	two	one	-13.4
1978	one	one	11.3
1979	one	one	13.4
1980	one	one	12.6
1981	two	one	-7.3
1982	one	one	18.8
1983	one	one	11.7
1984	one	one	9.5
1985	one	one	16.5
1986	one	one	25.8
1987	two	one	-6.5
1988	one	one	14.6
1989	one	one	10.1
1990	one	two	4.4
1991	one	one	17.3
1992	one	one	7.1
1993	one	one	9.3
1994	two	two	-2.4
1995	one	one	30.2
1996	one	one	21.2

Table 12: S&P 500 Index: Regime Sequence Results for 1997-2007 (Out of Sample)

Year	GM-BW Regime	Hamilton Regime	Empirical Annual Return (%)
1997	one	one	22.1
1998	one	one	26.6
1999	one	one	8.6
2000	two	two	-2.1
2001	two	one	-18.9
2002	one	one	-27.8
2003	one	one	27.9
2004	one	two	4.3
2005	one	two	8.0
2006	one	one	11.6
2007	two	two	-2.6

Nikkei 225 Index

Figure 3: Graphs of the Empirical Annual Returns of the Nikkei 225 Index (1976-96) and the Associated Regimes Under GM-BW and Hamilton Filter Calibration

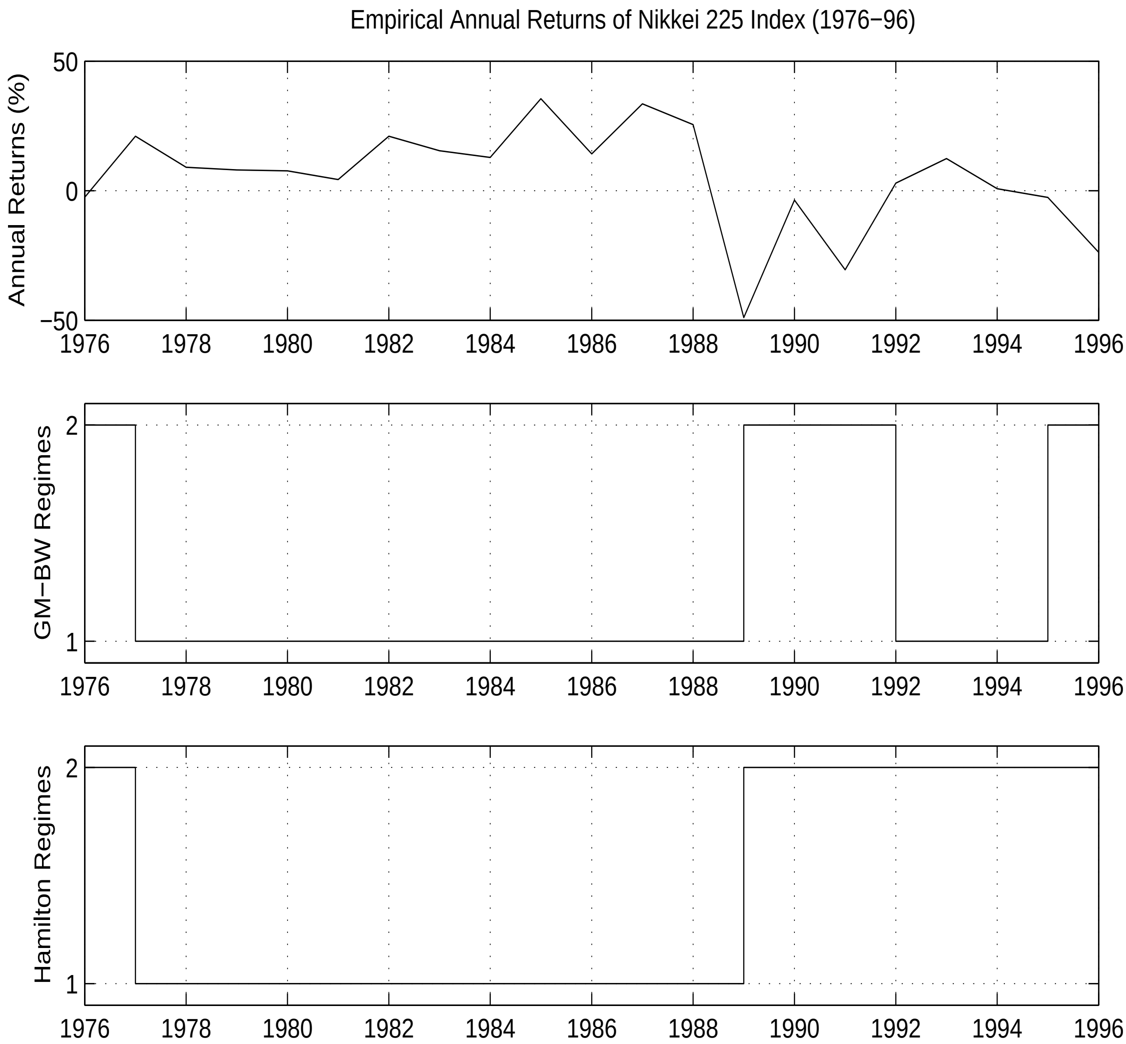

Figure 4: Graphs of the Empirical Annual Returns of the Nikkei 225 Index (1997-2007) and the Associated Regimes Under GM-BW and Hamilton Filter Calibration

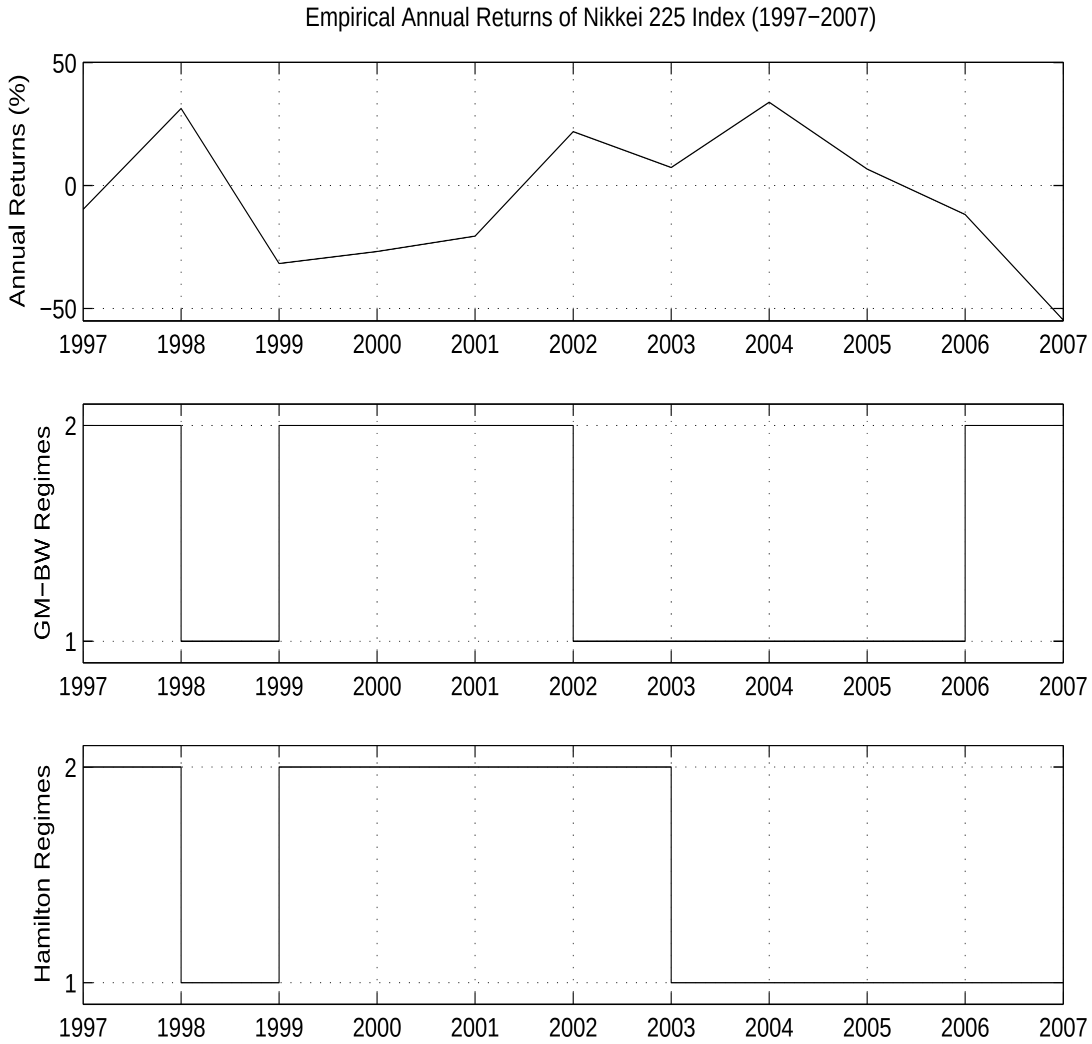

Table 13: Nikkei 225 Index: Regime Sequence Results for 1976-96 (In Sample)

Year	GM-BW Regime	Hamilton Regime	Empirical Annual Return (%)
1976	two	two	-2.5
1977	one	one	21.0
1978	one	one	9.0
1979	one	one	8.0
1980	one	one	7.6
1981	one	one	4.3
1982	one	one	21.0
1983	one	one	15.4
1984	one	one	12.8
1985	one	one	35.5
1986	one	one	14.2
1987	one	one	33.5
1988	one	one	25.5
1989	two	two	-49.0
1990	two	two	-3.7
1991	two	two	-30.6
1992	one	two	2.9
1993	one	two	12.4
1994	one	two	0.7
1995	two	two	-2.6
1996	two	two	-23.8

Table 14: Nikkei 225 Index: Regime Sequence Results for 1997-2007 (Out of Sample)

Year	GM-BW Regime	Hamilton Regime	Empirical Annual Return (%)
1997	two	two	-9.7
1998	one	one	31.3
1999	two	two	-31.7
2000	two	two	-26.8
2001	two	two	-20.6
2002	one	two	21.9
2003	one	one	7.3
2004	one	one	33.8
2005	one	one	6.7
2006	two	one	-11.8
2007	two	one	-54.7

4.3. Discussion

For the GM-BW method, from tables 2 and 7 we can infer that the method has attributed state two as the down state since their overall means are negative, unlike state one. This is also consistent with the state sequences generated in the S&P 500 and Nikkei 225 markets (see figures 1-4). Furthermore the initial state probabilities $\overline{\pi}$ for both markets suggest that we start 1976 in state 2. This is consistent with the empirical data where the 1976 returns are low: for the S&P 500 and Nikkei 225 it is 1.2% and -2.5% respectively (see tables 11 and 13).

The GM-BW transition matrices $\bar{\mathbf{A}}$ capture the differing time dynamics for each market, which one can observe from the empirical returns in figures 1 and 3. For the S&P 500 market 1976-96 we can see the annual returns exhibit an approximate cyclical relation with a cycle time of 5 years (average life time of an economic cycle). The GM-BW S&P 500 $\bar{\mathbf{A}}$ captures this reversionary dynamic through the probability 0.82 of returning to state 1, given we are in state 2. The Nikkei market from 1976-96 on the other hand tends not to exhibit cyclical behaviour, rather tends to remain in its current state (be it state 1 or 2). Hence the diagonal transition probabilities (which capture the memory effect of returning to the same state) are higher for the Nikkei than in the S&P 500. In conclusion, the GM-BW parameter estimates (distribution and HMM parameters) are consistent with each other and consistent with the empirical data in both markets.

For the Hamilton filter calibration, the calibrated models were less consistent than the GM-BW models. As one can see from tables 5 and 10 the filter has given state one a positive mean for the Nikkei 225 market but not for the S&P 500. Since state one is meant to represent the up state it should have a positive mean in both markets, hence the Hamilton calibration has not been as satisfactory. Note that we could have assigned state one as the down state for the Hamilton filter, however this would have produced regime sequences extremely inconsistent with the empirical data (see figures 1-4).

The Hamilton filter produces state transition matrices that are less consistent with the empirical data. Firstly, as explained previously the S&P 500 exhibits cyclical behaviour, which was successfully captured by the GM-BW's transition matrix, for example the probability of returning to state one given we are in state 2 is 0.82. However for the same transition probability the Hamilton filter only assigns a probability of 0.03, which implies the model should not cycle or quickly revert back to state 1. The Hamilton filter model also implies that the S&P 500 market has a probability of 0.97 remaining in state 2 given it is in state 2, yet the empirical returns clearly do not remain stuck in a down state. Secondly, the Hamilton filter for the Nikkei market assigns a probability of 0.92 to state 2, given we are in state 2, so that the model remains depressed in state 2 once it enters it. However the empirical returns of the Nikkei 1976-96 market (see figure 3) shows the market returns fluctuate.

To compare the quality of the models against the empirical data, regime sequences were produced from the GM-BW and Hamilton filter models for the in sample and out of sample periods. For the GM-BW model one can see from figures 1 and 3 the regime switches correspond well to the in sample empirical data: the model switches to state 2 during low returns and state 1 otherwise for both markets. The Hamilton filter on the other hand (for the in sample period) incorrectly identifies regimes for some years. For the S&P 500, the Hamilton filter identifies state one for years with significant negative returns rather than as state 2: 1977 (-13.4%), 1981 (-7.3%) and 1987 (-6.5%). Similarly for the Nikkei market the Hamilton filter identifies 1994 as state 2 (7.1%) rather than state 1. Overall, the Hamilton filter tends to incorrectly identify states more than the GM-BW method for the in sample period.

In the out of sample period (1997-07), regime sequencing performance of the GM-BW tends to outperform the Hamilton filter. For both markets the GM-BW model accurately identifies the states for most years (see figures 2 and 4) e.g. for the S&P 500 for 2000-2 was the period of the "internet bubble crash" and is identified as state 2 by GM-BW. However the Hamilton filter model does not identify 2001 as state 2 (see figure 2) despite the highly negative return (-18.9%), nor the following years as state 1 despite the positive returns: 2004 (4.3%) and 2005 (8%). Similarly for the Nikkei market the GM-BW accurately identifies the state of the market whereas the Hamilton identifies 2006 and 2007 as state 1 when they have highly negative returns -11.8% and -54.7% respectively. In conclusion we can say that the GM-BW calibration method outperforms the Hamilton filter for both markets, in and out of sample, in terms of parameter model estimation and regime sequence identification.

5. Conclusions

This paper has shown the advantages of Baum-Welch calibration over standard Hamilton filter method for calibration of regime switching volatility models. Not only does the Baum-Welch method offer a complete calibration procedure but also is able to estimate the full set of HMM parameters, unlike the Hamilton filter. We have also validated the usage of the Baum-Welch method through numerical experiments on S&P 500 and Nikkei 225 index data, in and out of sample, and compared its performance against the Hamilton filter.

References

[A.H00] A.Holtsberg. A statistics toolbox for matlab and octave. *http://www.maths.lth.se/matstat/stixbox/*, 2000.

[AK08] C. Alexander and A. Kaeck. Regime dependent determinants of credit default swap spreads. *Journal of Banking and Finance*, 32(6):637–648, 2008.

[BEDE04] P. Boufounos, S. El-Difrawy, and D. Ehrlich. Basecalling using hidden Markov models. *Journal of the Franklin Institute*, 341(1-2):23–36, 2004.

[BHL06] G. Bekaert, C.R. Harvey, and C. Lundblad. Growth volatility and financial liberalization. *Journal of International Money and Finance*, 25(3):370–403, 2006.

[BLR06] L. Bauwens, S. Laurent, and J.V.K. Rombouts. Multivariate GARCH Models: A Survey. *Journal of Applied Econometrics*, 21(1):79–109, 2006.

[BR02] C. Brooks and A.G. Rew. Testing for non-stationarity and cointegration allowing for the possibility of a structural break: an application to EuroSterling interest rates. *Economic Modelling*, 19(1):65–90, 2002.

[BSS07] I. Buckley, D. Saunders, and L. Seco. Portfolio optimization when asset returns have the Gaussian mixture distribution. *European Journal of Operational Research*, 185(3):1434–1461, 2007.

[DM94] J.M. Durland and T.H. McCurdy. Duration-Dependent Transitions in a Markov Model of US GNP Growth. *Journal of Business and Economic Statistics*, 12(3):279–288, 1994.

[EvdH97] R.J. Elliott and J. van der Hoek. An application of hidden Markov models to asset allocation problems. *Finance and Stochastics*, 1(3):229–238, 1997.

[Fam65] E.F. Fama. The behaviour of stock market prices. *Journal of Business*, 38(1):34–105, 1965.

[FJ73] G.D. Forney Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[Ham89] J.D. Hamilton. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–384, 1989.

[Ham91] J.D. Hamilton. A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions. *Journal of Business & Economic Statistics*, 9(1):27–39, 1991.

[Ham94] J.D. Hamilton. *Time series analysis*. Princeton, 1994.

[Har01] M.R. Hardy. A Regime-Switching Model of Long-Term Stock Returns. *North American Actuarial Journal*, 5(2):41–53, 2001.

[Hon03] T. Honda. Optimal portfolio choice for unobservable and regime-switching mean returns. *Journal of Economic Dynamics and Control*, 28(1):45–78, 2003.

[HS94] J.D. Hamilton and R. Susmel. Autoregressive Conditional Heteroskedasticity and Changes in Regime. *Journal of Econometrics*, 64(1-2):307–33, 1994.

[JR91] B.H. Juang and L.R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[KY95] M.J. Kim and J.S. Yoo. New index of coincident indicators: A multivariate Markov switching factor model approach. *Journal of Monetary Economics*, 36(3):607–630, 1995.

[LDLK04] N. Liu, R.I.A. Davis, B.C. Lovell, and P.J. Kootsookos. Effect of initial HMM choices in multiple sequence training for gesture recognition. *Information Technology: Coding and Computing*, 1(1):5–7, 2004.

[Lev05] S.E. Levinson. *Mathematical Models for Speech Technology*. Wiley England, 2005.

[MGPG06] C. Melodelima, L. Guéguen, D. Piau, and C. Gautier. A computational prediction of isochores based on hidden Markov models. *Gene*, 385(1):41–49, 2006.

[Mit09a] S. Mitra. A Review of Volatility and Option Pricing. *Arxiv*, 2009.

[Mit09b] S. Mitra. Regime Switching Stochastic Volatility with Perturbation Based Option Pricing. *Arxiv*, 2009.

[MT07] E. Messina and D. Toscani. Hidden Markov models for scenario generation. *IMA Journal of Management Mathematics*, 19(4):379–401, 2007.

[Mur08] K. Murphy. Hidden markov model (hmm) toolbox for matlab. *http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html*, 2008.

[PS98] Z. Psaradakis and M. Sola. Finite-sample properties of the maximum likelihood estimator in autoregressive models with Markov switching. *Journal of Econometrics*, 86(2):369–386, 1998.

[Rab89] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[Rab08] L.R. Rabiner. *Private Communication*, 2008.

[RI02] A. Ridolfi and J. Idier. Penalized Maximum Likelihood Estimation for Normal Mixture Distributions. *School of Computer and Information Sciences, Ecole Polytechnique Federale de Lausanne*, 200285, 2002.

[Sch89] G.W. Schwert. Why Does Stock Volatility Change Over Time? *Journal of Finance*, 44(5):1115–1153, 1989.

[SF06] T. Salih and K. Fidanboylu. Modeling and analysis of queuing handoff calls in single and two-tier cellular networks. *Computer Communications*, 29(17):3580–3590, 2006.

[SSS02] M. Sola, F. Spagnolo, and N. Spagnolo. A test for volatility spillovers. *Economics Letters*, 76(1):77–84, 2002.

[TG01] E. Trentin and M. Gori. A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126, 2001.

[Tim00] A. Timmermann. Moments of Markov switching models. *Journal of Econometrics*, 96(1):75–111, 2000.

[TK84] H.M. Taylor and S. Karlin. *An introduction to stochastic modeling.* Academic Press San Diego, 1984.

[VPHS04] P.L. Valls-Pereira, S. Hwang, and S.E. Satchell. How persistent is volatility? An answer with stochastic volatility models with Markov regime switching state equations. *Journal of Business Finance and Accounting*, 34(5-6):1002–1024, 2004.