Some contributions to filtering theory with applications in financial modelling

A thesis submitted for the degree of Doctor of Philosophy

by

Luka Jalen

Department of Mathematical Sciences
School of Information Systems,
Computing & Mathematics
Brunel University

July 2009

Abstract

Two main groups of filtering algorithms are characterised and developed. Their applicability is demonstrated using actuarial and financial time series data. The first group of algorithms involved hidden Markov models (HMM), where the parameters of an asset price model switch between regimes in accordance with the dynamics of a Markov chain. We start with the known HMM filtering set-up and extend the framework to the case where the drift and volatility have independent probabilistic behaviour. In addition, a non-normal noise term is considered and recursive formulae in the online re-estimation of model parameters are derived for the case of students' t-distributed noise. Change of reference probability is employed in the construction of the filters. Both extensions are then tested on financial and actuarial data. The second group of filtering algorithms deals with sigma point filtering techniques. We propose a method to generate sigma points from symmetric multivariate distributions. The algorithm matches the first three moments exactly and the fourth moment approximately; this minimises the worst case mismatch using a semidefinite programming approach. The sigma point generation procedure is in turn applied to construct algorithms in the latent state estimation of nonlinear time series models; a numerical demonstration of the procedure's effectiveness is given. Finally, we propose a partially linearised sigma point filter, which is an alternative technique for the optimal state estimation of a wide class of nonlinear time series models. In particular, sigma points are employed for generating samples of possible state values and then a linear programming-based procedure is utilised in the update step of the state simulation. The performance of the filtering technique is then assessed on simulated, highly non-linear multivariate interest rate process and is shown to perform significantly better than the extended Kalman filter in terms of computational time.

Acknowledgements

Several people made it possible for me to complete this thesis. First and foremost, I would like to thank my supervisor, Dr Rogemar Mamon, for his guidance and research support. His insights and knowledge had a profound effect on this work and have helped me overcome various challenges and hurdles during my research. This thesis has greatly benefited from his excellent comments and suggestions. In addition, I would like to thank him and the Department of Statistical and Actuarial Sciences at the University of Western Ontario, London, Canada for their hospitality during my research visit in 2007.

I would also like to thank Dr Paresh Date for many stimulating discussions and his valuable advice. The collaboration with him extended my research interests and added a new aspect reflected in this thesis. I would like to acknowledge Dr Gautam Mitra who was very supportive of my research and provided opportunities to broaden my knowledge through industrial placements.

I gratefully acknowledge the financial support provided by Ad-futura, Slovene human resources development and scholarship fund.

All my colleagues and friends at Brunel University deserve special thanks, especially Christian Valente, Gareth Clews, Carola Kruse, Christina Erlwein, Katharina Schwaiger, Leela Mitra and Jasmina Lazič who all contributed to a very enjoyable and stimulating working environment.

Finally, I would like to express my gratitude to my family and friends, especially to my parents Mira and Janez Jalen and my sister Spela Jalen, for their unlimited support and encouragement throughout these years. I am also deeply grateful to Ana Kneževič for her patience limitless support, that made the completion of this thesis possible.

Comments on collaboration with other researchers

This thesis is the result of research work carried out during my PhD studies. It is partially based on publications in several refereed journals and technical reports arising from collaborations with other researchers. These are detailed below.

Chapter 3 is an extended version of a publication in [93]. This research was mostly carried out during a 3-month visit at the University of Western Ontario in 2006. My supervisor, Dr Rogemar Mamon proposed and conceptualised the research plan whilst I derived the filters and recursive expressions for parameter updates. In addition, I implemented the filters and ran the numerical experiments.

Publication [73] is a modified version of chapter 6 and was developed together with Dr Rogemar Mamon. He proposed the research direction and supervised the derivation of the pricing equations as well as the implementation. The actual development of the valuation formulae and their implementation were my contributions.

Chapter 7 is a modified version of an article published in [37]. It is a joint work with Drs Paresh Date and Rogemar Mamon who proposed the idea and developed the sampling algorithm. My contributions in the article are the verification and implementation of the algorithm as well as generating the results of the numerical experiments.

Major parts of chapter 8 constitute a paper published in the journal Applied Mathematics and Computation [35]. This emerged from a joint venture with Drs Paresh Date and Rogemar Mamon who proposed the research topic and developed majority of the algorithms and numerical procedures. I was involved in the specification of the algorithm as well as implementation of it and all numerical experiments.

Chapter 9 is based on the joint work with Drs Paresh Date and Rogemar Mamon who outlined the research problem together with the preliminaries of the filtering algorithm. I contributed to the final version of the algorithm as well as the implementation and numerical experiments. A modified version of this chapter has been submitted for publication [36].

Contents

List of Figures

List of Tables

Nomenclature

The notation used throughout the discussion in the succeeding chapters is introduced here.

$A^\top$	transpose of a vector or matrix A
A_{ij}	entry in the i^{th} row and j^{th} column of a matrix A
e_i	i-th basis vector in $\mathbb{R}^N$
X_k	Markov chain
V_{k+1}	a martingale increment
S_X	state-space of the Markov chain
$\Pi = \{\pi_{ij}\}$	transition probability matrix
$(\Omega, \mathcal{F}, P)$	probability space
$\mathcal{F}$	filtration generated by the Markov chain X
$\mathcal{Y}$	filtration generated by the observation process
$\mathcal{H}$	global filtration, $\mathcal{H} = \mathcal{F} \vee \mathcal{Y}$
$\mathcal{M}$	filtration generated by the evolution of mortality rate process
$\{y_k\}$	a sequence of observations
$\{z_k\}$	a sequence of IID standard normal variables
$\mathcal{J}_k^{rs}$	number of jumps from state r to state s of a Markov chain X
$\mathcal{O}_k^r$	occupation time in state r of a Markov chain X
$\mathcal{T}_k^r(h)$	auxiliary process of an observation process for a function h
$\gamma(Y_k)$	unnormalised conditional expectation of Y_k given $\mathcal{H}_k$ under a reference probability measure, $\mathbb{E}[\Lambda_k Y_k \mid \mathcal{H}_k]$
$\gamma(X_k)$	estimator of the state of the Markov chain
α, $\underline{\alpha}$	drift vector
β, $\underline{\beta}$	volatility vector
ν	degrees of freedom for the student's t-distribution
W	standard Brownian motion
J	pure jump process
$\mu(x,t)$	force of mortality for an individual aged x at time t
$p(t,T,x)$	survival probability from t to T for an individual aged $x+t$
$S(t,x)$	survival function of a life aged x
m	number of random variables
s	number of scenarios
$\mathcal{X}$	discrete n-dimensional random variable
Φ	target mean vector for $\mathcal{X}$

R	target covariance matrix for $\mathcal{X}$
κ_i	target marginal 4^{th} central moment for the i^{th} random variable $\mathcal{X}_i$
$\mathcal{W}, \mathcal{V}$	symmetric vector-valued random variables with bounded mean, variance and marginal kurtosis
$\mathbf{f}$, $\mathbf{g}$, $\mathbf{h}$	given nonlinear, vector-valued functions
P_{xx}, P_{xy}, P_{vv}	convariance matrices under single factor, single measurement system
Σ_{XZ}, Σ_{YY}	covariance matrices
$\mathcal{G}$	multivariate, discrete distribution
$\mathbb{P}(A)$	probability of an event A
$\mathbb{E}[\mathcal{Z}]$	expected value of a random variable $\mathcal{Z}$

Chapter 1

Introduction

The problem of estimating the state of a dynamic system is practicably very important in many areas such as speech recognition, radar tracking, computer vision and control theory to name just a few. Filtering algorithms, developed primarily for engineering applications, have a broad range of usage and are increasingly popular in the analysis of economic time series. The main idea behind the filtering problems in the context of finance and economics is that the latent state of the system and other unobservable information about the system's or models's parameters can be estimated optimally from the observation process corrupted by noise. The observation process itself could be a univariate or a multivariate time series.

One type of filtering algorithms considered in this thesis is driven by a hidden Markov model (HMM). An HMM is a mathematical model where the system being modelled is assumed to be governed by a hidden Markov process. As is generally the case with filtering problems, the parameters of the model are unknown and must be determined from a set of observable data. The HMM modelling has its roots in speech recognition and signal processing, however it is becoming increasingly popular in mathematical finance. Pioneering works in the applications of HMM to financial time series were put forward by Hamilton in [64] and [65]. In the context of HMM, the hidden information is in the form of a finite state Markov chain in either discrete or continuous time which modulates the observation process.

In the applications of HMMs to financial or actuarial time series the states of the underlying Markov chain can be interpreted as the "states of the world". In the regime switching framework, the parameters of the model can take different values depending on the underlying state and are therefore capable of adapting to different stages of the business cycle, the current states of supply and demand, amongst other economic factors. In the context of mortality modelling investigated in the succeeding chapters, the states can correspond to different stages of medical advances, for example, a development of new drugs or dietary trends or habits of the target population, and in extreme cases, occurrences of natural disasters or terrorist attacks.

Whilst HMM filtering applications in finance have been explored by various researchers, we are not aware of papers examining its applications in mortality modelling. In addition, this thesis proposes and develops further extensions of the HMM framework based on the work of Elliot *et al* [43]. In particular, we relax the requirements on the distribution of the noise term as well as examine the case when the drift and volatility components of an HMM are independent. An associated challenge that arises from the estimation and implementation of HMMs is the calculation of optimal parameters. In this thesis, we contribute further to the literature by refining the change of probability measure techniques from Elliot *et al* [43] for the generalised case of noise distribution and the situation when dependence between drift and volatility is relaxed. Using the Expectation-Maximisation algorithm, the formulae for updating parameters of the models are derived for these extensions. Proposed models in the succeeding chapters feature an online updating of parameters, that is, parameters are updated as new information arrives, thus making these models self-tuning.

Moving away from the specialised group of HMM based models, the state estimation in general nonlinear models is considered and this can be numerically very challenging as well. Specifically, the optimal recursive solution to the state estimation problem requires the propagation of a full probability density. In the specialised case of linear Gaussian state-space models, a closed-form expressions exist for the

conditional state density and these are given by the linear Kalman filter. The hidden Markov model and Kalman filter have a certain degree of parallelism as there is a strong duality between the equations of the Kalman filter and those of the hidden Markov model; see for example [1] and [103]. Kalman filter provides a closed-form expressions in the filtering of linear processes, however there are no closed-form expressions for filtering the general nonlinear models. The current practical approaches to address this type of filtering problems are the Extended Kalman filter (EKF), Sequential Monte Carlo filtering and unscented filter. Under EKF the underlying model is locally linearised resulting in a linear state space system on which Kalman filter is utilised to obtain the conditional state density. Detailed discussion of EKF as well as its implementation can be found in standard textbooks such as [4]. However, EKF will only perform well when the system is indeed approximately linear, an assumption which is often very hard to validate. Under sequential Monte Carlo filtering the required density functions are obtained using a set of random samples. These density functions and the corresponding probability weights are in turn used to compute the needed conditional moment estimates. Monte Carlo filtering can perform significantly better than EKF for highly nonlinear systems and its result approaches the optimal Bayesian estimate as the number of samples becomes large enough as shown in [85] and [81]. On the downside, due to the large number of samples that need to be generated at each time step, this approach can be computationally very expensive. A compromise between EKF and Monte Carlo filtering is the unscented filter, which uses the closed form recursive expressions based on a linear Kalman filter to propagate the moments of the state vector. A small set of sample points or sigma points are propagated through the nonlinear transformation in order to compute the conditional moment estimates. In contrast to the Monte Carlo filter, the unscented filter uses a small number of sample points chosen in a way that they match some of the moment properties exactly. These filters have become popular especially in engineering, however they do suffer from several shortcomings. Amongst the drawbacks are: (i) the probability weights corresponding to the sample points are not guaranteed to be positive; (ii) there is no

randomness in the filtering procedure; and (iii) the square root of the covariance matrix has to be computed at each time step.

The main objective of this thesis is to elaborate and extend the filtering techniques briefly described above as well as to apply them to the field of financial and actuarial modelling. More specifically, we wish to (i) extend the existing HMM filtering framework capable of handling non-normal noise term and the assumption of independence between the dynamics of the drift and that of the volatility; (ii) calculate optimal parameter estimates for the extended framework using the change of probability measure techniques; (iii) investigate the performance of the extended HMM framework on observable market data as well as develop and implement an HMM-based mortality model; (iv) address the shortcomings of the existing sigma point generation algorithms and propose a new technique that could match the first four moments whilst guaranteeing the generated sample points always form a valid distribution; and (v) test the sigma point generation algorithm along with the development of a new unscented filter that is able to cope with highly nonlinear models without the use of computationally intensive Monte Carlo methods.

In order to attain the above-mentioned objectives this thesis is organised as follows. The first chapter gives an overview of hidden Markov models and an introduction to the change of measure technique. A brief description of the expectation-maximisation procedure, both of which are used in the succeeding chapters, namely chapters 3, 4, 5 and partially in chapter 6, is also presented. We assume the observation process has a drift and volatility as in the usual HMM set-up developed by Elliot *et al* [43]; that is, both the drift and volatility are functions of the underlying Markov chain. Unlike the standard setting, the drift and volatility have different number of states in chapter 3. The recursive filters for the parameter updates are developed for both the univariate and multivariate observation process. This modelling approach is further tested on the DOW JONES and NASDAQ indices, where we model logarithmic returns as a function of a Markov chain. The predictability of the indices within the HMM framework is assessed for several number of states for both drift and volatility and the errors are reported in terms of the root mean

square error (RMSE). In chapter 4, the discussion of HMM filtering is continued and we examine the case where the noise distribution is not normal. Assuming a general distribution for the error term complicates the algebra considerably. Needless to say, not all formulae for the parameter estimates can be developed in the general case. We outline the procedure to derive the expressions for the parameter updates when the noise term follows a Student's t-distribution. It is not possible to obtain recursive updating expressions for all the parameters as in the normally distributed noise case. Thus, one needs to resort to numerical methods during the update stage. In this generalised case, we first consider a numerical implementation on simulated data; the excellent performance of the algorithm is demonstrated on a small data set. In addition, we examine the performance and computational time requirements of the generalised algorithm on the same data set as in chapter 3, namely on the observed values of the DOW JONES and NASDAQ indices.

Chapter 5 utilises HMM filtering methods and explores an application in mortality modelling. Unlike the known mortality models, we introduce randomness in the whole mortality surface instead of only investigating how cohort mortality develops. In fact, we use the known and tested models for cohort mortality and let their parameters vary through time as a function of the Markov chain. We demonstrate through numerical examples that such a set-up can indeed capture mortality developments. Motivated by chapter 5, we investigate the pricing of common mortality-linked contracts in chapter 6. In particular, we examine the case where the dynamics of economic and demographic variables are not assumed independent. As noted by other researchers (see for example [24]), it is practical to assume the independence between demographic and economic factors, however they are not completely separate. It can be argued indeed that extreme events such as natural disasters can have influence on the economic variables. We therefore drop the assumption of independence and employ the change of measure technique and Bayes rule to derive closed form expressions for pricing common claims contingent on the survival or death of an individual.

We return to the main topic of filtering in chapter 7.1. Here, we briefly describe the

Kalman filter and its known extensions as well as highlight the similarities between Kalman filter and HMM filtering. As mentioned above, the two are quite analogues of each other. In chapter 7, we present a novel technique for sigma point generation which addresses the shortcomings of standard sample point generation procedures. In particular, the sigma points and corresponding weights are guaranteed to form a valid distribution as well as match its first four moments. Moreover, using the approach presented in chapter 7 it is possible (although not necessary) to introduce randomness to the filtering algorithm via sample point generation. In chapter 8, we present the unscented filter based on the sample point generation from chapter 7 and assess its performance on a simulated data from a nonlinear model. Finally, we give an alternative technique for the optimal state estimation of a nonlinear time series models. The sigma points for generating the possible state values are employed at the prediction step and then a linear programing-based procedure is used during the update stage of the state estimation. The performance of the algorithm is tested on simulated data generated from a multivariate and highly nonlinear interest rate process.

The thesis concludes in chapter 10 with a summary of research contributions as well as an outline of possible future research directions, which are motivated by the analysis carried out in this work.

Part I

Contributions to filtering of hidden Markov models

Chapter 2

Review of hidden Markov models

In this chapter, hidden Markov models are introduced together with a description of their basic properties and features. For completeness Markov chains are defined in discrete-time setting followed by the description of a general framework for HMMs. In the suceeding sections the characteristics of HMMs are underlined followed by a short overview of HMM filtering. We focus on the methodology used in deriving recursive formulae for optimal parameter estimates developed by Elliot *et al* (see for example, [41], [42], [43] and [45]). This methodology is in turn used and extended in the next chapters.

2.1 Markov chains in discrete time

A Markov chain, named after Andrey Markov, is a stochastic process with Markov property. Having a Markov property means that the process is without memory. This means that the future states depend only on its current state and is therefore conditionally independent of the past. Following the discussion in [97] we shall assume that the Markov chain has a countable number of states.

Define a probability space $(\Omega, \mathcal{F}, P)$ and let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of random numbers in the finite state space $S_X = \{s_1, \ldots, s_N\}$.

Definition 2.1

A Markov chain is a sequence of random variables $X_1, X_2, \ldots$ with the Markov property

$$\mathbb{P}(X_{k+1} \mid X_k = x_k, \ldots, X_0 = x_0) = \mathbb{P}(X_{k+1} \mid X_k = x_k) \tag{2.1}$$

$\forall k \geq 1$ and $x_0, \ldots, x_k \in S_X$.

In addition, the Markov chain is characterised by its transition matrix Π. In particular, a specific element π_{ij} of a transition matrix Π denotes the probability of Markov chain switching from state j to state i.

$$\pi_{ij} = \mathbb{P}(X_{k+1} = i \mid X_k = j) \quad i, j \in S_X \tag{2.2}$$

The Markov chain is said to be time homogenous when the transition probabilities $\pi_{ij} = \mathbb{P}(X_{k+1} = i \mid X_k = j)$ and $i, j \in S_X$ do not depend on time k. Furthermore the l-step ahead transition probabilities can therefore be calculated by multiplying the transition matrix Π by itself l times. That is,

$$\mathbb{P}(X_k = i \mid X_{k-l} = j) = \pi_{ij}^{(l)}, \tag{2.3}$$

where $\pi_{ij}^{(l)} = (\Pi^l)_{ij}$ is the (i, j) entry in the l-step transition probability matrix.

Any real function $f(X)$ can be expressed as a linear functional $f(X) = \langle s, X \rangle$ where $s = (s_1, \ldots, s_N)$ and $\langle s, e_i \rangle = s_i$. Here, $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean scalar product in $\mathbb{R}^N$. Therefore, remembering that the space S_X is finite, it is possible to represent the Markov chain by the canonical basis $\{e_1, \ldots, e_N\}$ of $\mathbb{R}^N$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ and $\top$ denotes the transpose of a vector. With the original state space S_X, when $s_k = i$ the Markov chain is represented by e_i, the

unit vector with the element 1 in the i-th row and zero, otherwise. The conditional expectation of X_k is then given by the i-th column of the transition matrix Π, i.e.,

$$\mathbb{E}\big[X_k \mid X_{k-1} = e_i\big] = \begin{bmatrix} \\ \pi_{i1} \\ \\ \vdots \\ \\ \pi_{iN} \\ \\ \end{bmatrix}. \tag{2.4}$$

Therefore, we have $\mathbb{E}\big[X_k \mid X_{k-1}\big] = \Pi X_{k-1}$.

Considering the above, note that the Markov chain represented as unit vectors can be expressed in the form

$$X_{k+1} = \Pi X_k + V_{k+1}, \tag{2.5}$$

where V_k is a martingale increment; see [43]. It is not possible to forecast V_k based on the previous states. Since V_k is a martingale increment, it follows from (2.5) that

$$\mathbb{E}\big[X_k \mid X_{k-l}\big] = \Pi^l X_{k-l}. \tag{2.6}$$

2.2 Hidden Markov models

Under the hidden Markov model setting, a Markov chain is assumed to be embedded in a stochastic process. In other words, we consider a stochastic process assumed to be (partially) driven by a Markov chain. The Markov chain itself is not observable directly, rather it is hidden in some observation process. The aim of the filtering in the HMM framework is to estimate the underlying Markov chain, that is, approximate in the best possible way the sequence $\{X_k\}$ from the series of observations. Under the real world measure as well as under the reference probability measure used in the succeeding chapters, the Markov chain follows the dynamics $X_{k+1} = \Pi X_k + V_{k+1}$, where Π is the transition probability matrix and V_k is a

martingale increment. Additionally, we assume the Markov chain itself is homogeneous with finite state-space in discrete time. The observation process, denoted by $\{y_k\}$, can follow various types of dynamics. In this thesis, however, we focus on the observation process of the form $y_{k+1} = \langle \alpha, X_k \rangle + \langle \beta, X_k \rangle z_{k+1}$, where α and β are real vectors of appropriate dimensions and $\{z_k\}$ is a sequence of independent, identically distributed (IID) random variables.

Hidden Markov models were first introduced by Baum and Petrie [9] in 1966 and were further developed by Baum and others in the following decade (see for example, [7], [8] and [11]). Baum *et al* [10] also introduced the Baum-Welch algorithm for parameter estimation within the framework of HMMs. Further details of the development and applications of HMMs are given in Ephraim and Merhav [48] and the references therein. The Baum-Welch algorithm is a particular instance of the EM algorithm specialised to HMM. The Baum-Welch algorithm is also called a forward-backward algorithm as it requires two passes through the data. Elliott *et al*'s [43] approach, in particular, requires only one, forward, pass through the data to achieve parameter updates.

More recently, Hamilton introduced the idea of regime switching in economics and finance in [63], [64] and [65]. Due to the flexibility of HMM set-up as well as apparent term structure in a range of observeable financial data, the HMMs have become increasingly popular in the modelling of financial time series. More recent developments and applications can be found for example in [45], [92] and [50], amongst others.

2.3 Change of reference probability technique

In this section, we give a summary of the change of probability measure technique employed in HMM filtering problem. The change of measure technique was introduced to stochastic filtering by Zakai [114] and has since become widely used in filtering applications. Such change of measure, based on a discrete time version

of Girsanov's theorem, is utilised in Elliott *et al* [43] to derive recursive optimal filters. This technique enables us to perform calculations under a mathematically more appropriate measure where the calculations are significantly simplified. Following Elliot *et al* [43], we shall name this new measure as a reference probability measure. The reference probability measure is equivalent to the real world measure; and under the reference probability measure the observation process $\{y_k\}_{k\in\mathbb{N}}$ are IID random variables. The dynamics of the underlying Markov chain do not change under the new measure which brings a considerable simplification to the calculations. This fact therefore enables us to employ Fubini-type results as opposed to direct calculations which would require hard semi-martingale methods.

Let $(\Omega, \mathcal{F})$ be a measurable space. For the purpose of completeness, we start with the definition of equivalence of two measures. The equivalence of two measures is an important concept and a useful tool in the succeeding discussion together with Cameron-Martin-Girsanov and Bayes' theorems.

Definition 2.2

A probability measure P is absolutely continuous with respect to the probability measure Q, written $P \ll Q$, if for each $A \in \mathcal{F}$, $P(A) = 0$ implies $Q(A) = 0$. The two measures are said to be equivalent (denoted by $P \equiv Q$) if $P \ll Q$ and also $Q \ll P$.

The theory underlying the change of measure relies on the equivalence of two probability measures defined on $(\Omega, \mathcal{F})$ linked via a Radon-Nikodŷm derivative. We suppose P is a probability measure on $\mathcal{F}$. To construct an equivalent measure Q on $\mathcal{F}$ we invoke the following theorem.

Theorem 2.3

If a measure P is absolutely continuous with respect to a positive measure Q then there exist a unique, nonegative function f such that for every $E \in \mathcal{F}$

$$P(E) = \int_E f \, dQ.$$

The function f is called the Radon-Nikodým derivative of P with respect to Q and is usually denoted by $\frac{dP}{dQ}$.

Proof

See pages 121–123 of Rudin [104].

□

The probability measure Q on $(\Omega, \mathcal{F})$ is thus defined via this Radon-Nikodým derivative (for further discussion, see Elliot *et al* [43]). Another important tool in the succeeding discussion is the well-known Cameron-Martin-Girsanov theorem.

Theorem 2.4 (Cameron-Martin-Girsanov theorem)

If W_t is a P-Brownian motion and γ_t is an $\mathcal{F}$-adapted process satisfying the boundedness condition $\mathbb{E}_P\left[\exp(\frac{1}{2}\int_0^T \gamma_t^2 dt)\right] < \infty$, then there exists a measure Q such that

1. *Q is equivalent to P*
2. $\frac{dQ}{dP} = \exp\left(-\int_0^T \gamma_t dWt - \frac{1}{2}\int_0^T \gamma_t^2 dt\right)$
3. $\tilde{W}_t = W_t + \int_0^t \gamma_s ds$ *is a Q-Brownian motion.*

In other words, W_t is a drifting Q-Brownian motion with drift $-\gamma_t$ at time t.

Proof

See Girsanov [58].

□

Write

$$\frac{dQ}{dP}\Big|_{\mathcal{F}} := \Lambda.$$

It then follows that

$$Q(E) = \int_E \Lambda dP, \quad \forall E \in \mathcal{F}.$$

In order to derive the filtering equations for the Markov chain process, it is imperative that we consider the conditional expectations that relate two equivalent measures, see Elliott *et al* [43]). The conditional Bayes' theorem stated below is a fundamental tool in obtaining many results related to hidden Markov models.

Theorem 2.5 (Conditional Bayes' theorem)

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{G} \subset \mathcal{F}$ a sub-σ-algebra and suppose H is any $\mathcal{G}$-measurable function. Assume further that Q is a probability measure equivalent to P and defined via the Radon-Nikodým derivative

$$\frac{dQ}{dP} = \Lambda.$$

Then

$$\mathbb{E}^Q[H \mid \mathcal{G}] = \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]}.$$

Proof

To prove that the above equation holds, we need to show that

$$\int_A \mathbb{E}^Q[H \mid \mathcal{G}]\, dQ = \int_A \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]}\, dQ, \quad \forall A \in \mathcal{G}\,.$$

We proceed by defining a measurable function ψ

$$\psi = \begin{cases} \mathbb{E}^P[\Lambda H \mid \mathcal{G}] & \text{if} \quad \mathbb{E}^P[\Lambda \mid \mathcal{G}] > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

Now we have to distinguish between two subsets of $\mathcal{G}$. Suppose $G = \{\omega : \mathbb{E}^P[\Lambda \mid \mathcal{G}] = 0\}$ and $G^c = \{\omega : \mathbb{E}^P[\Lambda \mid \mathcal{G}] > 0\}$. Therefore, $\Lambda = 0$ a.s. on G. For any set $A \in G$ we can write $B = A \cap G^c$ and $C = A \cap G$. Since G and G^c are disjoint, i.e., $G \cap G^c = \emptyset$, it follows that $A = B \cup C$.

With this distinction we have

$$\begin{aligned}\int_A \mathbb{E}^Q[H \mid \mathcal{G}]\, dQ &= \int_A H\, dQ \\ &= \int_A H\Lambda\, dP = \int_B H\Lambda\, dP + \underbrace{\int_C H\Lambda\, dP}_{=0}\,. \end{aligned} \tag{2.8}$$

From the first integral in (2.8) *we get*

$$\int_B \Lambda H\, dP = \mathbb{E}^P[I_B H\Lambda]\,. \tag{2.9}$$

Now using (2.7)

$$\begin{aligned}\int_B \psi\, dQ = &\quad \int_B \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]}\, dQ \\ = &\quad \mathbb{E}^Q\left[I_B \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]}\right] = \mathbb{E}^P\left[I_B \Lambda \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]}\right]\,. \end{aligned} \tag{2.10}$$

We apply the tower property to (2.10) *and obtain*

$$\begin{aligned}\mathbb{E}^P\left[I_B \Lambda \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]}\right] &= \mathbb{E}^P\Big[\mathbb{E}^P\Big[I_B \Lambda \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]} \mid \mathcal{G}\Big]\Big] \\ &= \mathbb{E}^P\left[I_B \mathbb{E}^P[\Lambda \mid \mathcal{G}] \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]}\right] \\ &= \mathbb{E}^P[I_B \Lambda H]\,. \end{aligned} \tag{2.11}$$

Hence from equation (2.9) *and* (2.11), *we have* $\int_B \Lambda H\, dP = \int_B \psi\, dQ$. *Using* (2.8), *we therefore see that*

$$\begin{aligned}\int_A \Lambda H\, dP &= \int_C \Lambda H\, dP + \int_B \Lambda H\, dP \\ &= \int_A \mathbb{E}^Q[H \mid \mathcal{G}]\, dQ = \int_A \psi\, dQ\,. \end{aligned}$$

So, we finally get

$$\mathbb{E}^Q[H \mid \mathcal{G}] = \frac{\mathbb{E}^P[\Lambda H \mid \mathcal{G}]}{\mathbb{E}^P[\Lambda \mid \mathcal{G}]} \ .$$

□

2.4 The EM algorithm

In this section, we shall briefly describe the Expectation-Maximisation (EM) algorithm. It is an important technique in estimating the parameters of a probabilistic model, where the model depends on the unobserved latent variables. EM is an iterative method which first performs the expectation step (E); this step computes an expectation of the log-likelihood with respect to the current estimate of the distribution of latent variables. The expectation step is followed by the maximisation step (M), which computes the parameters that maximise the expected log-likelihood found in the E-step. These are in turn used to determine the distribution of the latent variable in the next E-step.

The EM algorithm was first developed by Dempster, Laird and Rubin [38] and has been widely used in engineering, computing and economics. It is an alternative procedure to finding maximum likelihood estimates (MLEs) in incomplete data problems, where it is difficult to compute the MLEs due to missing values or where maximisation of the likelihood function is analytically untractable (see for example, McLachlan [94]).

Let θ be a set of parameters in the parameter space Θ and let $\{P^\theta, \theta \in \Theta\}$ be a family of absolutely continuous measures with respect to a fixed probability measure P^0 on a measurable space $(\Omega, \mathcal{F})$. In addition, assume that there is a filtration $\mathcal{Y} \subset \mathcal{F}$.

Our goal is to calculate an optimal estimate of the set of parameters θ. The likelihood function for the calculation of θ based on the information contained in

$\mathcal{Y}$ is therefore given by

$$F(\theta) = \mathbb{E}^0 \left[\frac{dP^\theta}{dP^0} \;\middle|\; \mathcal{Y} \right],$$

whilst the maximum likelihood estimate for θ is defined by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} F(\theta).$$

The MLE above is, however, not straightforward to compute; therefore, we rely on the EM algorithm which tackles the problem indirectly with an iterative approximation method as discussed below. For a detailed review, see Elliott and Krishnamurthy [46].

First set $n = 0$ and choose $\hat{\theta}_0$. Each iteration of the EM algorithm consists of two steps as mentioned above: the expectation step (E) and the maximisation step (M).

1. Expectation step: Determine the function $M_l(\theta, \hat{\theta}_n)$

$$M_l(\theta, \hat{\theta}_n) = \mathbb{E}^{\hat{\theta}_n} \left[\frac{dP^\theta}{dP^{\hat{\theta}_n}} \;\middle|\; \mathcal{Y} \right].$$

2. Maximisation step: Find a value of $\theta \in \Theta$ such that it maximises $M_l(\theta, \hat{\theta}_n)$, that is

$$\hat{\theta}_{n+1} = \arg\max_{\theta \in \Theta} M_l(\theta, \hat{\theta}_n).$$

Now replace n by $n + 1$ and repeat steps 1 (E-step) and 2 (M-step) until some stopping criteria is met, such as $|\hat{\theta}_{n+1} - \hat{\theta}_n| < \epsilon$, for some specified ϵ.

It was shown in Wu [113] that the sequence $\{\hat{\theta}_n\}$ yields a non-decreasing values of the likelihood function. It was shown as well that the sequence converges to the local maximum of the afore-mentioned likelihood function. The particular instance of the EM algorithm specialised for HMMs is the well-known Baum-Welch algorithm

[10]. It is a forward-backward algorithm that calculates the forward and backward probabilities for each state of the HMM and in turn uses these to compute the MLEs of the parameters. For further details of the Baum-Welch algorithm, see [26].

In the succeeding chapters, the EM algorithm will play a crucial role in the estimation of the parameters of the HMM models. Parameter estimation is optimised via an application of the EM algorithm to the $\log \frac{dP^{\theta}}{dP^{\hat{\theta}}}$ with the previously calculated $\hat{\theta}$ as explained above.

Chapter 3

Parameter estimation in a regime-switching model when the drift and volatility are independent

3.1 Preliminaries

In the last few years, there has been a surge in the use of regime-switching models in capturing the dynamics of variables in the financial markets. Such modelling of financial or economic variables is necessary when pricing and hedging derivatives, which are becoming increasingly sophisticated nowadays. Under a regime-switching modelling framework, the shift from one regime to another is usually modulated by a Markov chain either in discrete or continuous time. In order to be more realistic, we assume that the Markov chain is unobservable and thus we suppose that the switching evolves in accordance with an HMM. In finance, the observables are financial time series of asset prices, interest rates, exchange rates, etc. We aim to "filter" the noise out of these observations in the best possible way in order to come up with the best estimates of the Markov chain and parameters of the model.

Hamilton [64] popularised the use of regime-switching by mixing normal distributions in an attempt to describe the state of an economy at any given time; see a more detailed and accessible presentation in [66]. From the perspective of economic modelling, regime-switching models are inspired by structural changes brought about by some uncertain events, institutional policies or intervention of monetary authories; see, for example the case of Mexico in [102]. Evidence that abounds in empirical finance and economics provide support that regime-switching models are capable of capturing the dynamics of financial primitives that are being modelled. Regime shifts occur in many types of financial markets and hence regime-switching models have been employed by various authors in order to accomplish more impact in certain financial modelling endeavours. See, for instance, Elliott, Hunter & Jamieson [45], and Bansal and Zhou [6], amongst others, in modelling the term structure of interest rates; Bollen, Gray and Whaley [15] in pricing derivatives linked to foreign-exchange; Boyle and Draviam [20] for valuation of exotic options; Chang and Tsai [29] in initiating tax reforms; Chu, Santoni and Liu [30] in the analysis of the stock market; and Haldrup and Nielsen [62] in modelling electricity prices. More recent developments in the use of regime-switching models in finance and economics driven by HMM can be found in Mamon and Elliott [91].

We note, however, that in these previous papers and many prior applications the switching of regimes for both the drift and volatility is governed by only one Markov chain. In this chapter, we investigate if given a dataset there is evidence to suggest that the number of regimes for the drift is different from that for the volatility. To attain this objective, we proceed in the following manner. In section 3.2, we describe the modelling framework of the HMM in discrete time. Then, we briefly revisit the Bayes rule and the change of probability measure which are central in the parameter estimation. The filters will be derived for the optimal state of the Markov chain and other related quantities. The EM algorithm will be employed in linking the adaptive filters to the optimal estimates of the model parameters. In section 3.3, we shall develop an extension that handles the case when the drift's and volatility's probabilistic behaviour are independent. Furthermore, we extend the

idea of independent drift and volatility to the case where the observation process is multivariate in section 3.4. The dynamics of the NASDAQ and DOW JONES indices are then examined using the filtering techniques under the proposed extended set-up. We generate one-step ahead forecasts and assess the quality of these forecasts via their root-mean-square errors (RMSEs). A summary and some concluding comments are given in section 3.6.

3.2 Optimal parameter estimates using the change of measure technique

Suppose $(\Omega, \mathcal{F}, P)$ denotes a probability space under which X_k is a discrete-time $(k = 1, 2, \ldots)$ homogenous Markov chain with finite state space. Without loss of generality, we can assume the state space of X_k is associated with the canonical basis $\{e_1, \ldots, e_N\} \in \mathbb{R}^N$ and $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ where $\top$ denotes the transpose of a vector (see section 2.1). In addition, the initial distribution of X_0 is known and $\Pi = (\pi_{ij})$ is the transition probability matrix with $\pi_{ij} = P(X_{k+1} = e_i \mid X_k = e_j)$. It is an established result that X_k has a semimartingale representation: $X_{k+1} = \Pi X_k + V_{k+1}$. We may view X_k as representing the "conceivable" states of an economy and as argued in section 3.1 may not be directly observable.

Now, let S_k be a series of asset prices. Then, we could observe the logarithmic increments

$$y_{k+1} = \ln S_{k+1} - \ln S_k = \ln \frac{S_{k+1}}{S_k} \quad \text{or} \quad S_{k+1} = S_k \exp\{y_{k+1}\}. \tag{3.1}$$

We propose that

$$y_{k+1} = \ln \frac{S_{k+1}}{S_k} = f(X_k, z_{k+1}) \tag{3.2}$$

for some function f and z_k's are IID random variables and independent of X. More

specifically,

$$y_{k+1} = f(X_k, z_{k+1}) = \alpha(X_k) + \beta(X_k)z_{k+1} = \langle \alpha, X_k \rangle + \langle \beta, X_k \rangle z_{k+1}. \tag{3.3}$$

To formalise the setting of the model further, we consider histories or filtrations $\mathcal{F}_k$ as the complete filtration generated by $X_0, X_1, \ldots, X_k$; $\mathcal{Y}_k$ as the complete filtration generated by $y_0, y_1, \ldots, y_k$ and $\mathcal{H}_k := \mathcal{F}_k \vee \mathcal{Y}_k$.

3.2.1 Method of reference probability

It should be clear that the signal model with real-valued y-process on $(\Omega, \mathcal{F}, P)$ has the form

$$y_{k+1} = \alpha(X_k) + \beta(X_k)z_{k+1},$$

where

$$X_{k+1} = \Pi X_k + V_{k+1}$$

and z_k's are IID standard normals. Define a new probability measure Q via the Radon-Nikodŷm derivative Λ_k by

$$\left.\frac{dQ}{dP}\right|_{\mathcal{H}_k} = \Lambda_k = \prod_{l=1}^{k} \lambda_l$$

with

$$\Lambda_0 = 1 \quad \text{and} \quad \lambda_l = \frac{\langle \beta, X_{l-1} \rangle \phi(y_l)}{\phi(z_l)},$$

where $\phi(z)$ is the pdf of a standard normal.

Lemma 3.1

Under the probability measure Q, the y_k's are N(0,1) IID random variables.

Proof

See [43].

□

We wish as well to construct the measure P from Q. That is, conversely, suppose we start with Q on $(\Omega, \mathcal{F})$ such that

(i) $X_{k+1} = \Pi X_k + V_{k+1}$, $\mathbb{E}^Q[V_{k+1} \mid \mathcal{F}_k] = 0$

(ii) y_k is a sequence of N(0,1) IID.

We reiterate that the aim here is to construct a probability measure P from Q such that under P, z_k is a sequence of standard normals. That is, under P we have the required model $y_{k+1} = \alpha(X_k) + \beta(X_k)z_{k+1}$ with z_k, N(0,1) and IID. Introduce the inverses of λ_l and Λ_k. Write

$$\lambda_l^{-1} := \tilde{\lambda} = \frac{\phi(z_l)}{\langle \sigma, X_{l-1} \rangle \phi(y_l)},$$

with

$$\widetilde{\Lambda}_0 = 1 \quad \text{and} \quad \widetilde{\Lambda}_k = \prod_{l=1}^{k} \widetilde{\lambda}_l \quad \text{for} \quad k \geq 1.$$

Define P by putting $\frac{dP}{dQ}\Big|_{\mathcal{H}_k} = \widetilde{\Lambda}_k$.

Lemma 3.2

Under P, $\{z_k\}$ is a sequence of standard normal IID random variables.

Proof

See [43].

□

Remark 3.3

We shall work under measure Q. However, it is under P that $y_{k+1} = \alpha(X_k) + \beta(X_k)z_{k+1}$, with z_k being N(0,1) and IID.

3.2.2 Recursive filters for the state of the Markov chain and other related quantities

For any $\mathcal{H}_k$-adapted process $\{Y_k\}$, define

$$\gamma(Y_k) := E^Q[\widetilde{\Lambda}_k Y_k | \mathcal{H}_k].$$

From Bayes' rule, it follows that

$$\mathbb{E}[Y_k|\mathcal{H}_k] = \frac{E^Q[\widetilde{\Lambda}_k Y_k|\mathcal{H}_k]}{E^Q[\widetilde{\Lambda}_k|\mathcal{H}_k]} = \frac{\gamma(Y_k)}{\gamma(1)}. \tag{3.4}$$

Write

$$\Gamma^i(y_k) := \frac{\phi\left(\frac{y_k-\alpha_i}{\beta_i}\right)}{\beta_i \phi(y_k)}. \tag{3.5}$$

Furthermore, define the following:

1. $\mathcal{J}_k^{rs} = \sum_{l=1}^k \langle X_{l-1}, e_r\rangle\langle X_l, e_s\rangle$ as the number of jumps from e_r to e_s in time k,
2. $\mathcal{O}_k^r = \sum_{l=1}^k \langle X_{l-1}, e_r\rangle$ as the occupation time in e_r and
3. $\mathcal{T}_k^r(h) = \sum_{l=1}^k \langle X_{l-1}, e_r\rangle h(y_l)$ as an auxiliary, where $h(y) = y$ or y^2.

Theorem 3.4

If $\Gamma^i(y_k)$ is defined as in (3.5) *then the recursive relations for $\gamma(J_k^{rs}X_k)$, $\gamma(O_k^r X_k)$ and $\gamma(T_k^r(h)X_k)$ are given by*

$$\gamma(\mathcal{J}_k^{rs}X_k) = \sum_{l=1}^k \langle \gamma(\mathcal{J}_{l-1}^{rs}X_{l-1}), e_i\rangle \Gamma^i(y_k)\Pi e_i + \langle \gamma(X_{k-1}), e_r\rangle \pi_{sr}\Gamma^r(y_k)e_s$$

$$\gamma(\mathcal{O}_k^r X_k) = \sum_{i=1}^N \langle \gamma(\mathcal{O}_{k-1}^r X_{k-1}), e_i\rangle \Gamma^i(y_k)\Pi e_i + \Gamma^r(y_k)\langle \gamma(y_{k-1}), e_r\rangle \Pi e_r$$

$$\gamma(\mathcal{T}_k^r(h)X_k) = \sum_{i=1}^N \langle \gamma(\mathcal{T}_{k-1}^r(h)X_{k-1}), e_i\rangle \Gamma^i(y_k)\Pi e_i + \Gamma^r(y_k)\langle \gamma(X_{k-1}), e_r\rangle h(y_l)\Pi e_r.$$

Proof

See Elliot *et al* [43].

□

3.2.3 Relating recursive filters for vector processes to scalar quantities

Consider again equation (3.4), whose numerator and denominator are both scalars. So far, the recursions given in Theorem 3.4 only hold for vector processes. However, one can observe that for any process Y_k,

$$\sigma(Y_k) = \gamma(Y_k\langle X_k, \mathbf{1}\rangle) = \langle\gamma(Y_k X_k), \mathbf{1}\rangle \quad \text{where} \quad \mathbf{1} = (1, 1, \ldots, 1)^\top. \tag{3.6}$$

Here, $Y = \mathcal{J}, \mathcal{O}$ and $\mathcal{T}$. So, we have estimates for the quantities required in equation (3.4). In addition,

$$\gamma(1) = \mathbb{E}^Q[\widetilde{\Lambda}_k \mid H_k] = \langle\gamma(X_k), \mathbf{1}\rangle. \tag{3.7}$$

Finally, in order to obtain optimal estimates of model parameters, the EM-algorithm adopted from Dempster, Laird & Rubin [38] is applied. In our case, we have the set $\widehat{\theta}$ for the optimal parameter estimates given by $\widehat{\theta} = \left\{(\widehat{\pi}_{ji}), \widehat{\alpha}_i, \widehat{\beta}_i, 1 \leq i, j \leq N\right\}$, which determines the proposed model. We suppose that we are given a family of probability measures $\left\{P^\theta, \theta \in \Theta\right\}$ on some measurable space $(\Omega, \mathcal{G})$ and $\mathcal{Y} \in \mathcal{G}$. Our aim is to calculate the parameter θ.

1. Set $n = 0$ and choose $\widehat{\theta}_0$.
2. Set $\theta^* = \widehat{\theta_n}$ and determine $L(\theta, \theta^*) = \mathbb{E}^{\theta^*}\left[\log\dfrac{dP^\theta}{dP^{\theta^*}}\middle|\mathcal{Y}\right]$; this is called the E-step.
3. Find $\hat{\theta}_{n+1} \in \arg\max\limits_{\theta\in\Theta} L(\theta, \theta^*)$; this is called the M-step.

4. Replace n by $n+1$ and repeat procedures (E and M steps) until some stopping criterion is satisfied.

As mentioned earlier, it was shown in Wu [113] that the sequence $\left\{\widehat{\theta}_n\right\}$ produces non-decreasing likelihood values which converge to a local maximum of the likelihood function.

The recursive expressions above can be employed to estimate the parameters of the model. Application of the EM algorithm gives the following result:

Theorem 3.5

If a sequence of observations $y_1, \ldots, y_k$ are available at time k and the set of parameters $\{\hat{\pi}_{rs}, \hat{\alpha}_r, \hat{\beta}_r\}$ determines the model then the EM filter estimates for these parameters are given by

$$\hat{\pi}_{rs}(k) = \frac{\gamma(\mathcal{J}_k^{rs})}{\gamma(\mathcal{O}_k^r)} \tag{3.8}$$

$$\hat{\alpha}_r(k) = \frac{\gamma(\mathcal{T}_k^r(y))}{\gamma(\mathcal{O}_k^r)} \tag{3.9}$$

$$\hat{\beta}_r(k) = \sqrt{\frac{\gamma(\mathcal{T}_k^r(y^2)) - 2\hat{\alpha}_r\gamma(\mathcal{T}_k^r(y)) + \hat{\alpha}_r^2\gamma(\mathcal{O}_k^r)}{\gamma(\mathcal{O}_k^r)}}. \tag{3.10}$$

Proof

Formula for the re-estimation of the transition probabilities (3.8) follows as a special case of the result presented in Theorem 4.5. Alternative proof for (3.8) as well as proofs for (3.9) and (3.10) can be found in Elliott *et al* [43].

□

3.3 Extension to the case when the drift and volatility have independent probabilistic behaviour

Suppose α_k takes values in a finite set $C_\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ and β_k takes values in $C_\beta = \{\beta_1, \beta_2, \ldots, \beta_m\}$. Then, there are bijections c_α and c_β of C_α and C_β,

respectively with the set of finite vectors $D_\alpha = \{e_1, e_2, \ldots, e_n\} \in \mathbb{R}^n$ and $D_\beta = \{e_1, e_2, \ldots, e_m\} \in \mathbb{R}^m$. That is,

$$
\begin{aligned}
&c_\alpha : C_\alpha \longrightarrow D_\alpha \quad \text{and} \\
&c_\beta : C_\beta \longrightarrow D_\beta.
\end{aligned}
$$

Write $X_k^\alpha = c_\alpha(\alpha_k) \in \mathbb{R}^n$ and $X_k^\beta = c_\beta(\beta_k) \in \mathbb{R}^m$. Suppose X_k^l is a Markov chain on its state space C_l with transition matrix Π_l where $l \in \{\alpha, \beta\}$. We essentially assume our observation process is driven by two independent Markov chains, possibly with unequal number of states for the drift and volatility. The observation process y_k follows the form

$$
y_{k+1} = \langle \alpha, X_k^\alpha \rangle + \langle \beta, X_k^\beta \rangle z_{k+1}, \tag{3.11}
$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^\top \in \mathbb{R}^n$, $\beta = (\beta_1, \beta_2, \ldots, \beta_m) \in \mathbb{R}^m$, z_k are standard normal IID's. Both X_k^α and X_k^β are Markov chains of appropriate dimensions with respective dynamics

$$
\begin{aligned}
X_{k+1}^\alpha &= \Pi_\alpha X_k^\alpha + V_{k+1}^\alpha \\
X_{k+1}^\beta &= \Pi_\beta X_k^\beta + V_{k+1}^\beta.
\end{aligned} \tag{3.12}
$$

Here, Π_α and Π_β are the corresponding transition matrices, and V_k^α and V_k^β are the corresponding martingale increments.

Remark 3.6

Recall that if A is an $m \times n$ matrix and B is a $p \times q$ matrix, then the Kronecker product is the $mp \times nq$ block matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{22}B & \dots & a_{mn}B \end{bmatrix}. \tag{3.13}$$

Define $X_k = X_k^\alpha \otimes X_k^\beta$, where $\otimes$ denotes the tensor or Kronecker product. Then, we can identify X_k with a unit vector in $\mathbb{R}^{nm}$. From (3.12),

$$X_{k+1} = \Pi X_k + V_{k+1}, \tag{3.14}$$

where $\Pi = \Pi_\alpha \otimes \Pi_\beta$ and

$$V_{k+1} = \Pi_\alpha X_k^\alpha \otimes V_k^\beta + V_k^\alpha \otimes \Pi_\beta X_k^\beta + V_k^\alpha \otimes V_k^\beta. \tag{3.15}$$

Therefore $\mathbb{E}\big[V_k \mid \mathcal{G}_{k-1}\big] = 0$.

Example 3.7

Assume the situation where we have a 3-state drift and 2-state volatility. That is, $\alpha = (\alpha_1, \alpha_2, \alpha_3)^\top$ *and* $\beta = (\beta_1, \beta_2)^\top$. *In this case, re-formulate* α *and* β *as* $\alpha = (\alpha_1, \alpha_1, \alpha_2, \alpha_2, \alpha_3, \alpha_3)^\top$ *and* $\beta = (\beta_1, \beta_2, \beta_1, \beta_2, \beta_1, \beta_2)^\top$, *respectively. Then,* e_1 *picks up* α_1 *&* β_1*;* e_2 *picks up* α_1 *&* β_2*;* e_3 *picks up* α_2 *&* β_1*;* e_4 *picks up* α_2 *&* β_2*;* e_5 *picks up* α_3 *&* β_1*; and* e_6 *picks up* α_3 *&* β_2.

Although this idea was mentioned in [44], no further development, examples and details of numerical implementation to datasets were given by the authors.

By encapsulating the two Markov chains with unequal states driving the drift and volatility into one Markov chain, we should in principle be able to use the results given in section 3.2. However, by identifying both X_k^α and X_k^β as $X_k = X_k^\alpha \otimes X_k^\beta$,

which is a unit vector in $\mathbb{R}^{nm}$, we lose part of the information about the different states for the drift and volatility of the original model (3.11)-(3.12). In the general re-estimation of parameter values using Theorem 3.5, there would be nm distinct values for both α and β.

We note that some modification to the algebra, however, would allow the results presented in section 3.2 to be applied for the case where the drift and volatility have independent probabilistic behaviour. That will lead us to the recursive filters for X_k, Π, α and β. In order to do so, we shall first define the quantities used and introduce the notation. Write

$$\underline{\alpha} = \alpha \otimes 1(n) = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_1 \\ \vdots \\ \alpha_n \\ \vdots \\ \alpha_n \end{bmatrix}, \qquad \underline{\beta} = 1(m) \otimes \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \vdots \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \tag{3.16}$$

where $1(k)$ is a vector of ones of length k. If the probabilistic behaviour of the drift α and volatility β are independent, (3.11) can be re-written as

$$y_{k+1} = \langle \underline{\alpha}, X_k \rangle + \langle \underline{\beta}, X_k \rangle z_{k+1}. \tag{3.17}$$

Filtering algorithms for the process in (3.17) are known, however as argued above they do not take into account the additional information derived from the construc-

tion. In other words, there is only n different elements in vector $\underline{\alpha}$, each repeated m times and the same is true for vector $\underline{\beta}$. As in section 3.2, define

$$\begin{aligned}
\mathcal{O}_{k+1}^{r} &= \sum_{i=1}^{k+1} \langle X_i, e_r \rangle \\
\mathcal{J}_{k+1}^{rs} &= \sum_{i=1}^{k+1} \langle X_i, e_r \rangle \langle X_i, e_s \rangle \\
\mathcal{T}_{k+1}^{r}(f) &= \sum_{i=1}^{k+1} \langle X_i, e_r \rangle h(y_i).
\end{aligned} \tag{3.18}$$

In addition to (3.18), we need the occupation time and number of jumps for the processes X_k^{α} and X_k^{β} as well. Write

$$\begin{aligned}
{}_{\alpha}\mathcal{O}_k^{r} &:= \sum_{i=1}^{k} \langle X_i^{\alpha}, e_r \rangle = \sum_{i=1}^{k} \sum_{j=1}^{m} \langle X_i, e_{j+(r-1)m} \rangle \\
&= \sum_{j=1}^{m} \mathcal{O}_k^{j+(r-1)m}, \quad \forall r \in \{1, \ldots, n\}
\end{aligned} \tag{3.19}$$

and

$$\begin{aligned}
{}_{\alpha}\mathcal{J}_k^{rs} &:= \sum_{i=1}^{k} \langle X_i^{\alpha}, e_r \rangle \langle X_i^{\alpha}, e_s \rangle = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{l=1}^{m} \langle X_i, e_{(r-1)m+j} \rangle \langle X_i, e_{(s-1)m+l} \rangle \\
&= \sum_{j=1}^{m} \sum_{l=1}^{m} \mathcal{J}_k^{(r-1)m+j,(s-1)m+l}, \quad \forall r, s \in \{1, \ldots, n\}.
\end{aligned} \tag{3.20}$$

In order to derive the expressions in equations (3.19) and (3.20) we use only the definition of occupation times coupled with the structure of the model in (3.11). In the same way, we can derive expressions for occupation time and number of jumps for the Markov chain X_k^{β} by considering

$$\begin{aligned}
{}_{\beta}\mathcal{O}_k^{r} &= \sum_{j=1}^{n} \mathcal{O}_k^{r+(j-1)m}, \quad \forall r \in \{1, \ldots, m\} \\
{}_{\beta}\mathcal{J}_k^{rs} &= \sum_{j=1}^{n} \sum_{l=1}^{n} \mathcal{J}_k^{r+(j-1)m,s+(l-1)m}, \quad \forall r, s \in \{1, \ldots, m\}.
\end{aligned} \tag{3.21}$$

We can now derive the recursive filters for the parameters of the model, i.e., equations to re-estimate the transition matrix Π and vectors α and β.

Theorem 3.8

If a sequence of observations $y_1, \ldots, y_k$ are available at time k and the set of parameters $\{\hat{\Pi}, \hat{\alpha}_r, \hat{\beta}_r\}$ determines the model (3.11) *then the EM filter estimates for these parameters are given by*

$$\begin{aligned}
&\hat{\Pi} = \hat{\Pi}_\alpha \otimes \hat{\Pi}_\beta \qquad \text{where} \\
&\qquad\qquad {}_\alpha\hat{\pi}_{rs} = \frac{\sum_{j=1}^m \sum_{l=1}^m \gamma\big(\mathcal{J}_k^{(r-1)m+j,(s-1)m+l}\big)}{\sum_{j=1}^m \gamma\big(\mathcal{O}_k^{j+(r-1)m}\big)} \\
&\qquad \text{and} \qquad {}_\beta\hat{\pi}_{rs} = \frac{\sum_{j=1}^n \sum_{l=1}^n \gamma\big(\mathcal{J}_k^{r+(j-1)m,s+(l-1)m}\big)}{\sum_{j=1}^n \gamma\big(\mathcal{O}_k^{r+(j-1)m}\big)} \qquad (3.22) \\
&\hat{\alpha}_r = \frac{\sum_{j=1}^m \gamma\big(\mathcal{T}_k^{j+(r-1)m}(y)\big)}{\sum_{j=1}^m \gamma\big(\mathcal{O}_k^{j+(r-1)m}\big)} \\
&\hat{\beta}_r = \frac{\sum_{l=1}^n \Big(\gamma\big(\mathcal{T}_k^{r+(l-1)m}\big)(y^2) - 2\gamma\big(\mathcal{T}_k^{r+(l-1)m}\big)(y)\alpha_l + \gamma\big(\mathcal{O}_k^{r+(l-1)m}\big)\alpha_l^2\Big)}{\gamma\big({}_\beta\mathcal{O}_k^r\big)}
\end{aligned}$$

Proof

By Theorem 3.5, we have

$${}_\alpha\hat{\pi}_{rs} = \frac{\gamma\big({}_\alpha\mathcal{J}_k^{rs}\big)}{\gamma\big({}_\alpha\mathcal{O}_k^r\big)} = \frac{\sum_{j=1}^m \gamma\big(\mathcal{J}_k^{j+(r-1)m}\big)}{\sum_{j=1}^m \gamma\big(\mathcal{O}_k^{j+(r-1)m}\big)} \qquad (3.23)$$

and

$${}_\beta\hat{\pi}_{rs} = \frac{\gamma\big({}_\beta\mathcal{J}_k^{rs}\big)}{\gamma\big({}_\beta\mathcal{O}_k^r\big)} = \frac{\sum_{j=1}^n \sum_{l=1}^n \gamma\big(\mathcal{J}_k^{r+(j-1)m,s+(l-1)m}\big)}{\sum_{j=1}^n \gamma\big(\mathcal{O}_k^{r+(j-1)m}\big)}. \qquad (3.24)$$

From (3.23) and (3.24), the transition matrix Π can be derived readily as $\Pi = \Pi_\alpha \otimes \Pi_\beta$.

To derive expressions in estimating elements of vector $\underline{\alpha}$, we need to consider the

expression

$$\begin{aligned}\log \Lambda_k &= \sum_{i=1}^{k}\sum_{r=1}^{nm} \frac{\langle X_{i-1}, e_r\rangle \big(2y_i \underline{\alpha}_r - \underline{\alpha}_r^2\big)}{2\underline{\beta}_r} \\ &= \sum_{i=1}^{k}\sum_{r=1}^{n} \frac{2\alpha_r \sum_{j=1}^{m}\langle X_{i-1}, e_{(r-1)m+j}\rangle y_i - \alpha_r^2 \sum_{j=1}^{m}\langle X_{i-1}, e_{(r-1)m+j}\rangle}{2\sum_{j=1}^{m}\beta_j} \\ &= \sum_{r=1}^{n} \frac{2_\alpha \mathcal{T}_k^r(y)\alpha_r - {}_\alpha\mathcal{O}_k^r \alpha_r^2}{2\sum_{j=1}^{m}\beta_j}.\end{aligned} \tag{3.25}$$

Hence,

$$\mathbb{E}\big[\log \Lambda_k \mid \mathcal{H}_k\big] = \sum_{r=1}^{n} \frac{2\gamma\big({}_\alpha\mathcal{T}_k^r(y)\big)\alpha_r - \gamma\big({}_\alpha\mathcal{O}_k^r\big)\alpha_r^2}{2\sum_{j=1}^{m}\beta_j}. \tag{3.26}$$

Differentiating (3.26) with respect to α_r and equating the derivative to zero, we find the optimal choice for α_r given the observation up to time k and this is

$$\alpha_r = \frac{\gamma\big({}_\alpha\mathcal{T}_k^r(y)\big)}{\gamma\big({}_\alpha\mathcal{O}_k^r\big)} = \frac{\sum_{j=1}^{m}\gamma\big(\mathcal{T}_k^{j+(r-1)m}\big)}{\sum_{j=1}^{m}\gamma\big(\mathcal{O}_k^{j+(r-1)m}\big)}. \tag{3.27}$$

Similarly, in order to derive the recursive expressions for the elements of vector $\underline{\beta}$, one needs to consider

$$\begin{aligned}\log \Lambda_k &= -\frac{1}{2}\left(\sum_{i=1}^{k}\sum_{r=1}^{nm}\langle X_{i-1}, e_r\rangle \log \underline{\beta}_r + \frac{\langle X_{i-1}, e_r\rangle}{\underline{\beta}_r}\big(y_i^2 + 2\underline{\alpha}_r y_i + \underline{\alpha}_r^2\big)\right) \\ &= -\frac{1}{2}\Bigg(\sum_{i=1}^{k}\sum_{j=1}^{m}\log\beta_j \sum_{l=1}^{n}\langle X_{i-1}, e_{j+(l-1)m}\rangle \\ &\quad + \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{1}{\beta_j}\sum_{l=1}^{n}\langle X_{i-1}, e_{j+(l-1)m}\rangle\big(y_i^2 - 2\alpha_l y_i + \alpha_l^2\big)\Bigg) \\ &= -\frac{1}{2}\sum_{j=1}^{m}\Bigg(\log\beta_j \; {}_\beta\mathcal{O}_k^j \\ &\quad + \frac{1}{\beta_j}\sum_{l=1}^{n}\big(\mathcal{T}_k^{j+(l-1)m}(y^2) - 2\mathcal{T}_k^{j+(l-1)m}(y)\alpha_l + \mathcal{O}_k^{j+(l-1)m}\alpha_l^2\big)\Bigg).\end{aligned} \tag{3.28}$$

Taking the expected value of (3.28) conditioned on the available observations, differentiating it with respect to β_j and equating the derivative to zero, we see that

the optimal choice for β_j given the observations, is

$$\beta_j = \frac{\sum_{l=1}^{n} \Big(\gamma\big(\mathcal{T}_k^{j+(l-1)m}(y^2)\big) - 2\gamma\big(\mathcal{T}_k^{j+(l-1)m}(y)\big)\alpha_l + \gamma\big(\mathcal{O}_k^{j+(l-1)m}\big)\alpha_l^2 \Big)}{\gamma\big(\sum_{l=1}^{n} \mathcal{O}_k^{r+(l-1)m}\big)}. \quad (3.29)$$

□

Theorem 3.8 provides recursive expressions for the re-estimation of model parameters when the drift and volatility are driven by two independent Markov chains, ${}_{\alpha}X_k$ and ${}_{\beta}X_k$. The recursive expressions as given in Theorem 3.8 do not differ much from the known case when the drift and volatility are both driven by the same Markov chain in a sense that they are functions of the "supplementary" processes defined in (3.18). The natural question that arises is whether we could recover the known formulae (as specified in Theorem 3.5) if ${}_{\alpha}X_k$ and ${}_{\beta}X_k$ are the same.

In order to confirm that start with a model as in (3.17)

$$y_{k+1} = \langle \underline{\alpha}, X_k \rangle + \langle \underline{\beta}, X_k \rangle z_{k+1},$$

where X_k is a Markov chain with a state space $S_X = \{e_1, e_2, \ldots, e_{nn}\}$ and transition probability matrix $\Pi \in \mathbb{R}^{nn}$, whilst z_k is a sequence of IID standard normals. In addition, assume that there exists a Markov chain X_k^{α} with transition probability matrix ${}_{\alpha}\Pi \in \mathbb{R}^n$ such that

$$X_k = X_k^{\alpha} \otimes X_k^{\alpha},$$

as well as vectors $\alpha, \beta \in \mathbb{R}^n$ such that $\underline{\alpha} = \alpha \otimes 1(n)$ and $\underline{\beta} = 1(n) \otimes \beta$ as in (3.16). Consequently, our model (3.17) collapses to

$$y_{k+1} = \langle \alpha, X_k^{\alpha} \rangle + \langle \beta, X_k^{\alpha} \rangle z_{k+1}.$$

Using the relationship between the supplementary processes (number of jumps $\mathcal{J}$, occupation time $\mathcal{O}$ and auxiliary process $\mathcal{T}$) for X_k and X_k^{α} as specified in equations (3.19) – (3.21) and plugging them in Theorem 3.8, one recovers the

recursive expressions given in Theorem 3.5.

We can therefore consider the case where the drift and volatility are driven by the same Markov chain as a special case of the general model where the drift and volatility are allowed to be independent. The extension to the independent drift and volatility is further investigated in the next section where we explore how to estimate the parameters when the observation process is not just a scalar process but assumed to be multi-dimensional.

3.4 Vector observations

As in the previous section, suppose we have a Markov chain with state space $S_X = \{e_1, e_2, \ldots, e_{nm}\}$ and $X_{k+1} = \Pi X_k + V_{k+1}$. In addition, suppose there exist two independent Markov chains, X_k^α with state space $S_{X^\alpha} = \{e_1, e_2, \ldots, e_n\}$ and X_k^β with state space $S_{X^\beta} = \{e_1, e_2, \ldots, e_m\}$ with dynamics

$$\begin{aligned} X_{k+1}^\alpha &= \Pi_\alpha X_k^\alpha + V_{k+1}^\alpha \\ X_{k+1}^\beta &= \Pi_\beta X_k^\beta + V_{k+1}^\beta, \end{aligned}$$

such that $X_k = X_k^\alpha \otimes X_k^\beta$ and $\Pi = \Pi_\alpha \otimes \Pi_\beta$ hold. Suppose now that the observation process is a d-dimenstional vector process with components

$$\begin{aligned} y_{k+1}^1 &= \langle \alpha^1, X_k^\alpha \rangle + \langle \beta^1, X_k^\beta \rangle z_{k+1}^1 \\ y_{k+1}^2 &= \langle \alpha^2, X_k^\alpha \rangle + \langle \beta^2, X_k^\beta \rangle z_{k+1}^2 \\ &\vdots \\ y_{k+1}^d &= \langle \alpha^d, X_k^\alpha \rangle + \langle \beta^d, X_k^\beta \rangle z_{k+1}^d. \end{aligned} \tag{3.30}$$

Here $\alpha^i = (\alpha_1^i, \alpha_2^i, \ldots, \alpha_n^i)^\top \in \mathbb{R}^n$, $\beta^i = (\beta_1^i, \beta_2^i, \ldots, \beta_m^i)^\top \in \mathbb{R}^m$ and z_k^i are IID standard normal random variables for $i \in \{1, 2, \ldots, d\}$.

For $i \in \{1, 2, \ldots, nm\}$ write

$$\Gamma^i(\underline{y}_{k+1}) = \Gamma^i(y^1_{k+1}, y^2_{k+1}, \ldots, y^d_{k+1}) = \prod_{j=1}^{d} \frac{\phi\left(\frac{y^j_{k+1} - \underline{\alpha}^j_i}{\underline{\beta}^j_i}\right)}{\underline{\beta}^j_i \phi\left(y^j_{k+1}\right)} \tag{3.31}$$

where $\underline{\alpha}$ and $\underline{\beta}$ are defined in (3.16). We can then prove an equivalent result to that in Theorem 3.8 for vector observations.

Lemma 3.9

If a sequence of observations $\underline{y}_1, \ldots, \underline{y}_k$ are available at time k and the set of parameters $\{\hat{\Pi}, \hat{\alpha}_r, \hat{\beta}_r\}$ determines the model (3.30) *then the EM filter estimates for these parameters are given by*

$$\begin{aligned}
&\hat{\Pi} = \hat{\Pi}_\alpha \otimes \hat{\Pi}_\beta, \qquad \text{where} \\
&\qquad {}_\alpha\hat{a}_{rs} = \frac{\sum_{j=1}^m \sum_{l=1}^m \gamma\big(\mathcal{J}_k^{(r-1)m+j,(s-1)m+l}\big)}{\sum_{j=1}^m \gamma\big(\mathcal{O}_k^{j+(r-1)m}\big)} \\
&\qquad \text{and} \qquad {}_\beta\hat{a}_{rs} = \frac{\sum_{j=1}^n \sum_{l=1}^n \gamma\big(\mathcal{J}_k^{r+(j-1)m,s+(l-1)m}\big)}{\sum_{j=1}^n \gamma\big(\mathcal{O}_k^{r+(j-1)m}\big)} \\
&\hat{\alpha}^j_r = \frac{\sum_{i=1}^m \gamma\big(\mathcal{T}_k^{i+(r-1)m}(y^j)\big)}{\sum_{i=1}^m \gamma\big(\mathcal{O}_k^{i+(r-1)m}\big)} \\
&\hat{\beta}^j_r = \frac{\sum_{l=1}^n \Big(\gamma\big(\mathcal{T}_k^{r+(l-1)m}\big)((y^j)^2) - 2\gamma\big(\mathcal{T}_k^{r+(l-1)m}\big)(y^j)\alpha^j_l + \gamma\big(\mathcal{O}_k^{r+(l-1)m}\big)(\alpha^j_l)^2\Big)}{\gamma\big({}_\beta\mathcal{O}^r_k\big)}.
\end{aligned} \tag{3.32}$$

Proof

We note that the recursive relations for $\gamma(J^{rs}_k X_k)$, $\gamma(O^r_k X_k)$ and $\gamma(T^r_k(h)X_k)$ stated in Theorem 3.4 hold for the vector observation case provided one defines $\Gamma^i(\underline{y}_{k+1})$ as in (3.31). The proof then reduces to the proof of Theorem 3.8 and therefore, it will not be repeated.

□

Hence, we have shown that the recursive filters for the parameters of the Markov

chain can be derived for the case where the drift and volatilty have independent probabilistic behaviour. Indeed, even for the vector observation case the recursive equations for estimation of model parameters are readily available as shown in Lemma 3.9. In the next section, we shall apply these filters to observed market data and demonstrate how these filters perform in practice.

3.5 Numerical application of the filters

The recursive filters from the previous section are implemented to the observed data for the two well-known indices, namely the NASDAQ and DOW JONES indices. The recursive filters ere applied to both the NASDAQ and DOW JONES datasets covering an approximately 4-year period from 28 February 2003 to 16 February 2007. The summary statistics for the returns of these datasets are given in Table 3.1.

Statistic	**NASDAQ data**	**DOW JONES data**
Mean	3.998×10^{-4}	4.819×10^{-4}
Median	-4.834×10^{-4}	4.585×10^{-4}
Standard deviation	0.008	0.007
Skewness	0.446	0.124
Kurtosis	4.532	4.881
Range	0.063	0.072
Minimum	-0.026	-0.036
Maximum	0.036	0.035
Count	1000	1000

Table 3.1: Summary statistics for the NASDAQ and DOW JONES logarithmic returns for the period 28/02/2003–16/02/2007

Suppose we choose n and m as the dimension of state spaces of the Markov chains X_k^{α} and X_k^{β} respectively. For any price process S_k, $k \in Z^+$, the steps to undertake in implementing the filtering under the extended framework are as follows:

1. Calculate $y_{k+1} = \ln \frac{S_{k+1}}{S_k}$.

2. Initialise the set of values $\{(\alpha_i, \beta_j), i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, m\}$.

3. Initialise the elements of the matrix $\Pi = (\pi_{ij}),\ 1 \leq i, j \leq N, \quad \sum_{i=1}^{N} \pi_{ij} = 1,\ \pi_{ij} \geq 0$.

4. After k values of y have been observed, compute new estimates for $\pi_{ij},\ \alpha$ and β using the recursive filters for $X, \mathcal{J}, \mathcal{O}$ and $\mathcal{T}$.

5. Use the values after further observations to re-estimate (π_{ij}), α and β. This yields a self-tuning model.

We also obtain one-step ahead forecasts with the aid of the formulae as in [92]. These are

$$\mathbb{E}[y_{k+1} \mid y_1, y_2, \ldots, y_k] = \langle \widehat{\alpha}, \widehat{\Pi}\widehat{X}_k \rangle \quad \text{and}$$
$$Var[y_{k+1} \mid y_1, y_2, \ldots, y_k] = \widehat{\alpha}^{\top} \text{diag}\widehat{\Pi}\widehat{X}_k\alpha + \beta^{\top}\text{diag}\widehat{\Pi}\widehat{X}_k\beta - \langle \widehat{\alpha}, \widehat{\Pi}\widehat{X}_k \rangle^2.$$

The data were processed in batches of ten data points. A batch of data being processed constitutes one pass in the algorithm and each pass produces a different set of new parameter estimates.

The initial parameter values for Π, α and β are displayed in Table 3.2 (page 38). After 3 passes, we have the re-estimated values of the parameters exhibited in Table 3.3 (page 38).

Table 3.4 (page 39) depicts the values of the parameters in the final pass. In Table 3.5 (page 39), we show the RMSEs and the computational time (in seconds) it takes to complete all the necessary calculations under different drift and volatility settings. The calculations were performed on a 1.83GHz dual core procesor using Matlab.

It can be seen from Table 3.5 (page 39) that the error is decreasing with increasing number of states driving both drift and volatility. In addition, a naive, no change model $E[y_{k+1} \mid y_k] = y_k$ has a RMSE value of 7.2682×10^{-3} for the DOW JONES

$$\Pi = \begin{bmatrix} 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 & 0.1111 \end{bmatrix}$$

$$\begin{aligned} X_k &= [0.8100 \;\; 0.0450 \;\; 0.0450 \;\; 0.0450 \;\; 0.0025 \;\; 0.0025 \;\; 0.0450 \;\; 0.0025 \;\; 0.0025]^\top \\ \alpha &= [-0.0100 \;\; -0.0100 \;\; -0.0100 \;\; 0.0000 \;\; 0.0000 \;\; 0.0000 \;\; 0.0110 \;\; 0.0110 \;\; 0.0110]^\top \\ \beta &= [0.0050 \;\; 0.0600 \;\; 0.0100 \;\; 0.0050 \;\; 0.0600 \;\; 0.0100 \;\; 0.0050 \;\; 0.0600 \;\; 0.0100]^\top \end{aligned}$$

Table 3.2: Initial parameter values for Π, X_k, α and β

$$\Pi = \begin{bmatrix} 0.2709 & 0.1229 & 0.1229 & 0.1267 & 0.0575 & 0.0575 & 0.1267 & 0.0575 & 0.0575 \\ 0.1641 & 0.1884 & 0.1641 & 0.0768 & 0.0882 & 0.0768 & 0.0768 & 0.0882 & 0.0768 \\ 0.1649 & 0.1649 & 0.1868 & 0.0771 & 0.0771 & 0.0874 & 0.0771 & 0.0771 & 0.0874 \\ 0.1595 & 0.0723 & 0.0723 & 0.2054 & 0.0932 & 0.0932 & 0.1595 & 0.0723 & 0.0723 \\ 0.0966 & 0.1109 & 0.0966 & 0.1244 & 0.1429 & 0.1244 & 0.0966 & 0.1109 & 0.0966 \\ 0.0971 & 0.0971 & 0.1100 & 0.1250 & 0.1250 & 0.1417 & 0.0971 & 0.0971 & 0.1100 \\ 0.1713 & 0.0777 & 0.0777 & 0.1713 & 0.0777 & 0.0777 & 0.1817 & 0.0824 & 0.0824 \\ 0.1038 & 0.1192 & 0.1038 & 0.1038 & 0.1192 & 0.1038 & 0.1101 & 0.1264 & 0.1101 \\ 0.1043 & 0.1043 & 0.1182 & 0.1043 & 0.1043 & 0.1182 & 0.1106 & 0.1106 & 0.1253 \end{bmatrix}$$

$$\begin{aligned} \hat{X}_k &= [0.1142 \;\; 0.1067 \;\; 0.1068 \;\; 0.1191 \;\; 0.1104 \;\; 0.1105 \;\; 0.1161 \;\; 0.1081 \;\; 0.1082]^\top \\ \hat{\alpha} &= [-0.0099 \;\; -0.0099 \;\; -0.0099 \;\; 0.0000 \;\; 0.0000 \;\; 0.0000 \;\; 0.0110 \;\; 0.0110 \;\; 0.0110]^\top \\ \hat{\beta} &= [0.0105 \;\; 0.0180 \;\; 0.0182 \;\; 0.0105 \;\; 0.0181 \;\; 0.0182 \;\; 0.0105 \;\; 0.0181 \;\; 0.0182]^\top \end{aligned}$$

Table 3.3: Parameter values for Π, X_k, α and β after the 3rd pass

data set. The RMSEs reported in Table 3.5 are considerably lower as soon as the number of states driving the drift is two or more.

The plots of the actual values versus the one-step ahead forecasts are displayed in

$$\Pi = \begin{bmatrix} 0.2761 & 0.1225 & 0.1225 & 0.1269 & 0.0563 & 0.0563 & 0.1269 & 0.0563 & 0.0563 \\ 0.1642 & 0.1927 & 0.1642 & 0.0754 & 0.0886 & 0.0754 & 0.0754 & 0.0886 & 0.0754 \\ 0.1650 & 0.1650 & 0.1911 & 0.0758 & 0.0758 & 0.0878 & 0.0758 & 0.0758 & 0.0878 \\ 0.1593 & 0.0707 & 0.0707 & 0.2112 & 0.0937 & 0.0937 & 0.1593 & 0.0707 & 0.0707 \\ 0.0947 & 0.1112 & 0.0947 & 0.1256 & 0.1475 & 0.1256 & 0.0947 & 0.1112 & 0.0947 \\ 0.0952 & 0.0952 & 0.1103 & 0.1262 & 0.1262 & 0.1462 & 0.0952 & 0.0952 & 0.1103 \\ 0.1718 & 0.0762 & 0.0762 & 0.1718 & 0.0762 & 0.0762 & 0.1863 & 0.0827 & 0.0827 \\ 0.1021 & 0.1199 & 0.1021 & 0.1021 & 0.1199 & 0.1021 & 0.1108 & 0.1301 & 0.1108 \\ 0.1026 & 0.1026 & 0.1189 & 0.1026 & 0.1026 & 0.1189 & 0.1113 & 0.1113 & 0.1290 \end{bmatrix}$$

$$\hat{X}_k = [0.1139 \quad 0.1065 \quad 0.1066 \quad 0.1192 \quad 0.1105 \quad 0.1106 \quad 0.1162 \quad 0.1083 \quad 0.1083]^\top$$
$$\hat{\alpha} = [-0.0099 \quad -0.0099 \quad -0.0099 \quad 0.0000 \quad 0.0000 \quad 0.0000 \quad 0.0110 \quad 0.0110 \quad 0.0110]^\top$$
$$\hat{\beta} = [0.0114 \quad 0.0184 \quad 0.0185 \quad 0.0114 \quad 0.0184 \quad 0.0185 \quad 0.0114 \quad 0.0184 \quad 0.0185]^\top$$

Table 3.4: Parameter values for Π, X_k, α and β after the final pass

Drift	**Volatility**	**RMSE**	**Computational time**
1-state	1-state	1.0414×10^{-2}	0.421
1-state	2-state	1.6499×10^{-2}	0.686
2-state	1-state	4.2686×10^{-3}	0.671
2-state	2-state	2.5714×10^{-3}	1.763
2-state	3-state	1.7258×10^{-3}	3.931
3-state	1-state	2.9231×10^{-3}	1.139
3-state	2-state	1.6745×10^{-3}	3.947
3-state	3-state	1.6140×10^{-3}	9.578

Table 3.5: Comparison of RMSEs and computational time (in secs) for the DOW JONES returns data.

Figures 3.1 and 3.2 for the NASDAQ and DOW JONES, respectively.

3.6 Some concluding remarks

In this chapter, a model for the evolution of a risky asset or a financial variable, in which a derivative contract may depend upon, is considered. The increment of

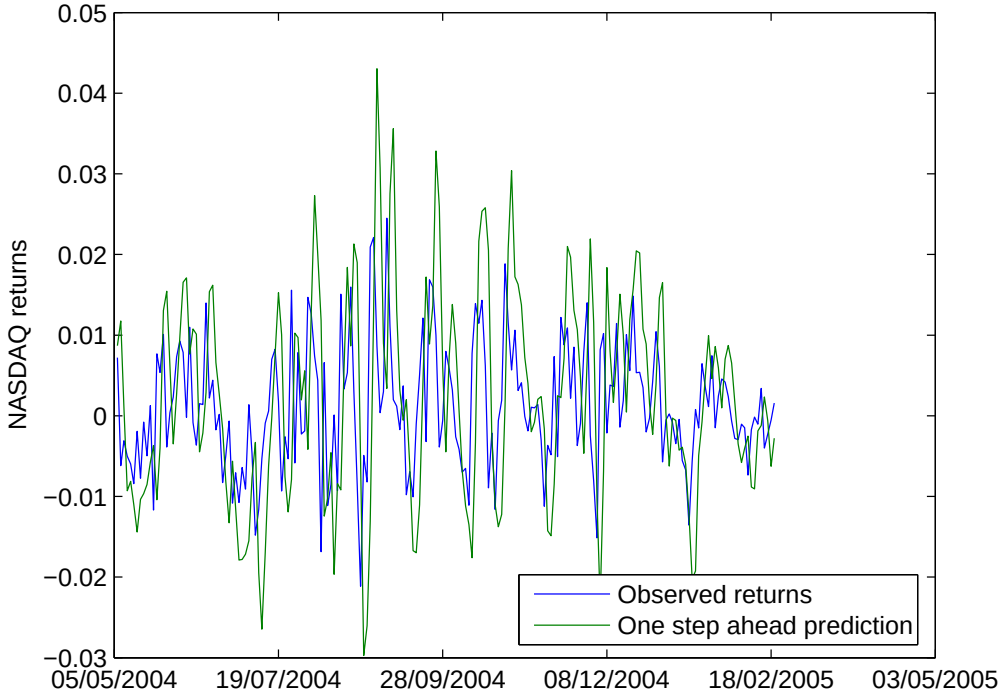

Figure 3.1: NASDAQ actual returns series and one-step ahead predictions: 3-state drift and 3-state volatility

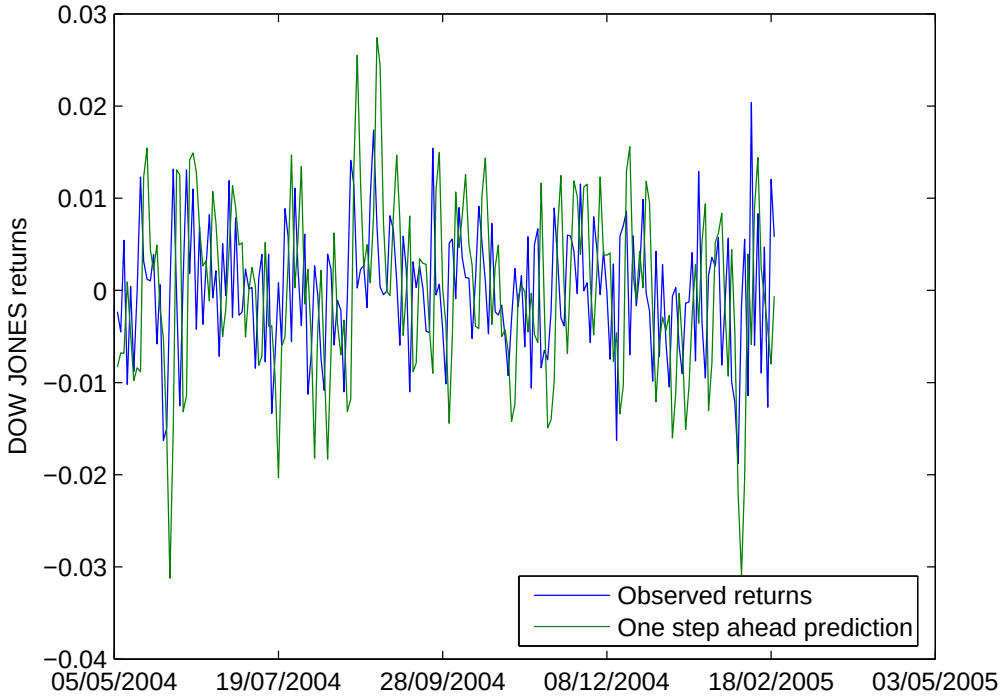

Figure 3.2: DOW JONES actual return series and one-step ahead predictions: 3-state drift and 3-state volatility

the logarithm of the price involves a Gaussian noise and parameters are governed by a finite state Markov chain. We revisited the estimation techniques from HMM theory and applied these not only to obtain the best estimate of the Markov chain and transition probability matrix but also to re-estimate all model parameters. The EM-based estimation procedure ensures that the model parameter estimates improve with each iteration. An extension of the estimation was formulated to handle the case when the drift and volatility are driven by independent Markov chains. The extended estimation procedures were tested on NASDAQ and DOW JONES data sets. Several combinations were considered, wherein the number of states for the drift and volatility parameters are unequal. On the basis of the RMSE calculated by comparing the actual values and the one-step ahead predictions, we found that there is evidence to support that the drift and volatility have different number of states at a given period. In addition, we have derived the recursive formulae for estimation of the parameters in the case of multivariate observations. In the next chapter we take the idea further and explore how HMM filtering performs when the noise term is not normally distributed. We study the filtering of vector observations as well as the case of independent drift and volatility considered in this chapter, however the noise term has a non-normal distribution.

Chapter 4

Parameter estimation in a regime-switching model with non-normal noise terms

4.1 Introduction

In the previous chapter, we considered a discrete-time, finite state Markov chain which is observed through a real-valued function whose values are corrupted by noise. Furthermore, we assumed the noise are IID normals.

It is however an accepted fact that many of the observable processes do not follow the normality assumption. In financial time series modelling assumptions of normal IID noise term results in tails thinner than observed on the market. There have been many attempts to move away from normality assumption in the literature, however non-normal noise complicates the algebra considerably.

In this chapter, we relax the assumption on the noise term and allow it to have a general distribution, potentially dependent on the state of the underlying Markov chain. With the relaxation of this assumption, it is in general not possible to come up with all the recursive formulae for the parameter re-estimation as in chapter 3. Without regard to the distribution of the noise, it is, nonetheless, possible to derive

the recursive formulae for estimating the transition probabilities. However, it is necessary to rely on numerical methods to approximate the remaining parameters, namely, the drift (vector α) and volatility (vector β) specified in equation (4.1).

All processes will be defined on a complete probability space $(\Omega, \mathcal{F}, P)$. Suppose X_k is a homogenous Markov chain with a finite state space on Ω. Without loss of generality, we can assume that the state space of X_k is associated with the canonical basis $\{e_1, \ldots, e_N\} \in \mathbb{R}^N$ and $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$, where $\top$ denotes the transpose of a vector as in the previous chapter. Further, assume X_0 is given, or its distribution is known and $\Pi = (\pi_{ij})$ is the transition probability matrix with $\pi_{ij} = P(X_{k+1} = e_i \mid X_k = e_j)$. Additionally, $X_{k+1} = \Pi X_k + V_{k+1}$ is not observed directly, rather there is an observation process

$$y_{k+1} = \langle \alpha, X_k \rangle + \langle \beta, X_k \rangle z_{k+1}(X_k), \tag{4.1}$$

where $\{z_{k+1}(X_k)\}$ is a sequence of independent random variables with a distribution function $\phi_{X_k}(\cdot)$, possibly dependent on the state of the Markov chain.

Consider further the filtrations or histories $\mathcal{F}_k$ as the complete filtration generated by $X_0, X_1, \ldots, X_k$; $\mathcal{Y}_k$ as the complete filtration generated by $y_0, y_1, \ldots, y_k$ and $\mathcal{H}_k := \mathcal{F}_k \vee \mathcal{Y}_k$.

4.2 Reference probability measure

It is under a real world measure P that the signal model with the real valued y-process on $(\Omega, \mathcal{F}, P)$ has the dynamics

$$\begin{aligned} X_{k+1} &= \Pi X_k + V_{k+1} \\ y_{k+1} &= \langle \alpha, X_k \rangle + \langle \beta, X_k \rangle z_{k+1}(X_k). \end{aligned} \tag{4.2}$$

We wish to introduce a new probability measure Q via a Radon-Nikodŷm derivative $\frac{dP}{dQ}$, such that under Q the random variable y_{k+1} has density $\phi_{X_k}(\cdot)$. Write

$$\begin{aligned} \lambda_l &= \sum_{i=1}^{N} \langle X_{l-1}, e_i \rangle \frac{\phi_{X_i}\left(\frac{y_l - \alpha_i}{\beta_i}\right)}{\beta_i \phi_{X_i}(y_l)}, \\ \Lambda_k &= \prod_{l=1}^{k} \lambda_l, \qquad \Lambda_0 = 1. \end{aligned} \tag{4.3}$$

Consider Λ_k in (4.3) as the Radon-Nikodŷm derivative of P with respect to Q. That is,

$$\frac{dP}{dQ}\Big|_{\mathcal{H}_k} = \Lambda_k, \tag{4.4}$$

where, as noted above, $\mathcal{H}_k$ is the joint filtration generated by both X and y.

Lemma 4.1

Write $\Phi_X(\cdot) = \sum_{i=1}^{N} \langle X_k, e_i \rangle \phi_{X_i}(\cdot)$. Under Q, the y_k are IID random variables distributed as a mixture of distributions $\phi_{X_i}(\cdot)$, $\Phi_X(\cdot)$.

Proof

From Bayes' Theorem,

$$\begin{aligned} Q(y_{k+1} \le t \mid \mathcal{H}_k) &= \mathbb{E}^Q\big[I(y_{k+1} \le t) \mid \mathcal{H}_k\big] \\ &= \frac{\Lambda_k}{\Lambda_k} \cdot \frac{\mathbb{E}\big[\lambda_{k+1}^{-1} I(y_{k+1} \le t) \mid \mathcal{H}_k\big]}{\mathbb{E}\big[\lambda_{k+1}^{-1} \mid \mathcal{H}_k\big]}. \end{aligned} \tag{4.5}$$

Now

$$\mathbb{E}\big[\lambda_{k+1} \mid \mathcal{H}_k\big] = \int_{-\infty}^{\infty} \frac{\langle \beta, X_k \rangle \Phi_X(y_{k+1})}{\Phi_X(z_{k+1})} \Phi_X(z_{k+1}) dz_{k+1} = 1. \tag{4.6}$$

Therefore we can write

$$\begin{aligned} Q(y_{k+1} \leq t \mid \mathcal{H}_k) &= \mathbb{E}^Q\big[I(y_{k+1} \leq t) \mid \mathcal{H}_k\big] \\ &= \int_{-\infty}^{\infty} \frac{\langle \beta, X_k \rangle \Phi_X(y_{k+1})}{\Phi_X(z_{k+1})} \Phi_X(z_{k+1}) I(y_{k+1} \leq t) dz_{k+1} \\ &= \int_{-\infty}^{t} \Phi_X(y_{k+1}) dy_{k+1} = Q(y_{k+1} \leq t). \end{aligned} \tag{4.7}$$

The result follows.

□

Remark 4.2

The inverse of Lemma 4.1 can be proven similarly. That is, if we start with a probability measure Q on $(\Omega, \mathcal{F})$, such that under Q X_k is a Markov chain with dynamics $X_{k+1} = \Pi X_k + V_{k+1}$, where V_k is the martingale increment, $\{y_k\}$ is a sequence of random IID variables with a distribution function $\Phi_X(\cdot)$ then under P z_k is a sequence of IID random variables following distribution $\Phi_X(\cdot)$ on $(\Omega, \mathcal{F})$.

Our aim is to estimate X given the observation under P, the real world probability. The calculations however will be performed under Q due to the convenience brought about by the introduction of the measure Q. For example, under the real world probability P even the recursive formulae for X_k are not linear in X_k, which leads to complications of the algebra (see [43] for more details).

Write $\xi_k := \mathbb{E}^Q\big[\Lambda_k X_k \mid \mathcal{Y}_k\big]$ and observing that $\sum_{i=1}^N \langle X_k, e_i \rangle = 1$, we have

$$\sum_{i=1}^{N} \mathbb{E}^Q\big[\langle \Lambda_k X_k, e_i \rangle\big] = \sum_{i=1}^{N} \langle \mathbb{E}^Q[\Lambda_k X_k \mid \mathcal{Y}_k], e_i \rangle = \sum_{i=1}^{N} \langle \xi_k, e_i \rangle. \tag{4.8}$$

In addition, let

$$\tilde{p}_k^i = \mathbb{P}(X_k = e_i \mid \mathcal{Y}_k) \quad \text{and} \quad \tilde{p}_k = (\tilde{p}_k^1, \ldots, \tilde{p}_k^N) \tag{4.9}$$

so that we get an explicit form for the conditional distribution of X_k under P given

$\mathcal{Y}_k$ given by

$$\tilde{p}_k = \frac{\xi_k}{\sum_{i=1}^{N} \langle \xi_k, e_i \rangle}. \tag{4.10}$$

4.3 Recursive estimation

In this section, we shall briefly present a recursive filter for vector process ξ_k as well as derive the recursive relations and formulae which will in turn be used in model parameter estimation. Write D for a diagonal matrix whose i-th element on the diagonal is

$$\frac{\phi_{X_i}\big(\frac{y_l - \alpha_i}{\beta_i}\big)}{\beta_i \phi_{X_i}(y_l)}. \tag{4.11}$$

Lemma 4.3

If $\xi_k := \mathbb{E}^Q\big[\Lambda_k X_k \mid \mathcal{Y}_k\big]$, *D is a matrix with diagonal elements of the form* (4.11) *and* Π *a transition matrix corresponding to the Markov chain* X_k *then*

$$\xi_{k+1} = \Pi D \xi_k \tag{4.12}$$

Proof

From the definition of ξ_k, we have

$$\xi_{k+1} := \mathbb{E}^Q\big[\Lambda_{k+1} X_{k+1} \mid \mathcal{Y}_{k+1}\big] \tag{4.13}$$

and therefore

$$\begin{aligned}
\xi_{k+1} &= \mathbb{E}^Q\big[\Lambda_{k+1} X_{k+1} \mid \mathcal{Y}_{k+1}\big] = \mathbb{E}^Q\big[\Lambda_k \lambda_{k+1}\big(\Pi X_k + V_{k+1}\big) \mid \mathcal{Y}_{k+1}\big] \\
&= \mathbb{E}^Q\left[\Lambda_k\Big(\sum_{i=1}^{N} \langle X_k, e_i \rangle \frac{\phi_{X_i}\big(\frac{y_{k+1} - \alpha_i}{\beta_i}\big)}{\beta_i \phi_{X_i}(y_{k+1})}\Big)\Pi X_k \mid \mathcal{Y}_{k+1}\right] \\
&= \sum_{i=1}^{N} \mathbb{E}^Q\big[\Lambda_k \langle X_k, e_i \rangle \mid \mathcal{Y}_{k+1}\big] \frac{\phi_{X_i}\big(\frac{y_{k+1} - \alpha_i}{\beta_i}\big)}{\beta_i \phi_{X_i}(y_{k+1})} \Pi e_i = \Pi D \xi_k
\end{aligned} \tag{4.14}$$

□

Our objective is to estimate the parameters of the model given in (4.1), i.e., estimate the transition matrix Π and vectors α and β. To do this, we define (as in chapter 3) the processes representing the respective occupation time, number of jumps of the Markov chain, and the auxiliary process below.

$$\begin{aligned}\mathcal{O}^r_{k+1} &= \sum_{i=1}^{k+1}\langle X_i, e_r\rangle \\ \mathcal{J}^{rs}_{k+1} &= \sum_{i=1}^{k+1}\langle X_i, e_r\rangle\langle X_i, e_s\rangle \\ \mathcal{T}^r_{k+1}(g) &= \sum_{i=1}^{k+1}\langle X_i, e_r\rangle g(y_i).\end{aligned} \tag{4.15}$$

Notation: For any $\mathcal{Y}$-adapted process C we shall use the notation $\gamma(C)_k := \mathbb{E}^Q[\Lambda_k C_k \mid \mathcal{Y}_k]$.

Theorem 4.4

If D is the diagonal matrix as defined in (4.11)*, the recursive relations for $\gamma(\mathcal{J}^{sr}X)_k$, $\gamma(\mathcal{O}^r X)_k$ and $\gamma(\mathcal{T}^r X)_k$ are*

$$\begin{aligned}\gamma(\mathcal{J}^{sr}X)_k &= \Pi D(y_k)\gamma(\mathcal{J}^{sr}X)_{k-1} + \langle \xi_{k-1}, e_r\rangle \frac{\phi_{X_r}\left(\frac{y_k-\alpha_r}{\beta_r}\right)}{\beta_r \phi_{X_r}(y_k)}\pi_{sr}e_s \\ \gamma(\mathcal{O}^r X)_k &= \Pi D(y_k)\gamma(\mathcal{O}^r X)_{k-1} + \langle \xi_{k-1}, e_r\rangle \frac{\phi_{X_r}\left(\frac{y_k-\alpha_r}{\beta_r}\right)}{\beta_r \phi_{X_r}(y_k)}\Pi e_r \\ \gamma(\mathcal{T}^r(g) X)_k &= \Pi D(y_k)\gamma(\mathcal{T}^r(g) X)_{k-1} + \langle \xi_{k-1}, e_r\rangle \frac{\phi_{X_r}\left(\frac{y_k-\alpha_r}{\beta_r}\right)}{\beta_r \phi_{X_r}(y_k)} g(y_k)\Pi e_r.\end{aligned} \tag{4.16}$$

Proof

$$\begin{aligned}
\gamma(\mathcal{J}^{sr}X)_k &= \mathbb{E}^Q\big[\Lambda_k \mathcal{J}_k^{sr} X_k \mid \mathcal{Y}_k\big] \\
&= \mathbb{E}^Q\big[\Lambda_{k-1}\lambda_k(\mathcal{J}_{k-1}^{sr} + \langle X_{k-1}, e_r\rangle\langle X_k, e_s\rangle)X_k \mid \mathcal{Y}_k\big] \\
&= \sum_{i=1}^{N} \mathbb{E}^Q\big[\Lambda_{k-1}\langle X_{k-1}, e_i\rangle \mathcal{J}_{k-1}^{sr} \mid \mathcal{Y}_k\big]\Big[\sum_{i=1}^{N}\langle X_{k-1}, e_i\rangle \frac{\phi_{X_i}(y_k - \alpha_i)}{\phi_{X_i}(y_k)}\Big] \quad (4.17) \\
&\quad + \mathbb{E}^Q\big[\Lambda_{k-1}\langle X_{k-1}, e_r\rangle \mid \mathcal{Y}_k\big]\frac{\phi_{X_r}(y_k - \alpha_r)}{\phi_{X_r}(y_k)}\pi_{sr}e_s \\
&= \Pi D(y_k)\gamma(\mathcal{J}^{sr}X)_{k-1} + \langle \xi_k, e_r\rangle \frac{\phi_{X_r}(y_k - \alpha_r)}{\phi_{X_r}}\pi_{sr}e_s
\end{aligned}$$

The proofs of the other two recursive formulae, namely $\gamma(\mathcal{O}^r X)_k$ and $\gamma(\mathcal{T}^r X)_k$ follow almost exactly the same arguments and are thus omitted.

□

4.4 Parameter estimation

In order to estimate the parameters of the model the EM algorithm is employed to determine the optimal approximation for each parameter in the set θ. Initial values for the EM algorithm are assumed to be given. Starting from the initial values, the updated parameter approximations are carried out based on the maximisation of the conditional expected log-likelihoods. Filters for the occupation time process, jump process and the auxiliary process given in Theorem 4.4 are required to calculate the optimal parameter estimates.

The EM algorithm requires the change of measure from P^θ to $P^{\tilde{\theta}}$. Under P^θ, X is a Markov chain with transition matrix Π. Under $P^{\tilde{\theta}}$, X is still a Markov chain with transition matrix $\tilde{\Pi}$ and as explained in section 2.4, the set of optimal parameters is obtained by maximising $\mathbb{E}^\theta\left[\frac{dP^{\tilde{\theta}}}{dP^\theta} \,\middle|\, \mathcal{Y}\right]$ with respect to the set of parameters $\tilde{\theta}$.

Theorem 4.5

If at time k a sequence $y_1, \ldots, y_k$ is available and the set of parameters $\{\pi_{sr}, \alpha_r, \beta_r\}$

determines the model then the EM estimates for transition probabilities are

$$\pi_{sr} = \frac{\gamma(\mathcal{J}^{sr}X)_k}{\gamma(\mathcal{O}^r X)_k}. \tag{4.18}$$

Proof

Using the Radon-Nikodým derivative of $P^{\tilde{\theta}}$ with respect to P^{θ} we have

$$\begin{aligned}
\log \frac{dP^{\tilde{\theta}}}{dP^{\theta}} &= \sum_{l=1}^{k} \log \left(\sum_{s,r=1}^{N} \left(\frac{\tilde{\pi}_{sr}}{\pi_{sr}} \right)^{\langle X_l, e_s \rangle \langle X_{l-1}, e_r \rangle} \right) \\
&= \sum_{l=1}^{k} \sum_{s,r=1}^{N} (\log \tilde{\pi}_{sr} - \log \pi_{sr}) \langle X_l, e_s \rangle \langle X_{l-1}, e_r \rangle \\
&= \sum_{s,r=1}^{N} \mathcal{J}_k^{sr} \log \tilde{\pi}_{sr} + R(\pi_{sr})
\end{aligned} \tag{4.19}$$

where the $R(\pi_{sr})$ does not involve $\tilde{\pi}_{sr}$. Observe that $\sum_{s=1}^{N} \mathcal{J}_k^{sr} = \mathcal{O}_k^r$, hence

$$\sum_{s=1}^{N} \tilde{\mathcal{J}}_k^{sr} = \tilde{\mathcal{O}}_k^r. \tag{4.20}$$

The $\tilde{\pi}_{ji}$'s optimal estimate is the value that maximises the log-likelihood (4.19) subject to the constraint $\sum_{s=1}^{N} \tilde{\pi}_{sr} = 1$.

Introducing the Lagrange multiplier δ, we consider the function

$$L(\tilde{\pi}, \delta) = \sum_{r,s=1}^{N} \tilde{\mathcal{J}}_k^{sr} \log \tilde{\pi}_{sr} + \delta \left(\sum_{s=1}^{N} \tilde{\pi}_{sr} - 1 \right) + R(\pi_{sr}). \tag{4.21}$$

Differentiating (4.21) with respect to $\tilde{\pi}_{sr}$ and δ and equating the derivatives to 0, we have

$$\frac{1}{\tilde{\pi}_{sr}} \tilde{\mathcal{J}}_k^{sr} + \delta = 0. \tag{4.22}$$

Equation (4.22), however, can be re-written as

$$\tilde{\pi}_{sr} = \frac{\tilde{J}_k^{sr}}{-\delta}. \tag{4.23}$$

Therefore,

$$\sum_{j=1}^{N} \tilde{\pi}_{sr} = \frac{\sum_{s=1}^{N} \tilde{\mathcal{J}}_k^{sr}}{-\delta}. \tag{4.24}$$

Considering $\sum_{s=1}^{N} \tilde{\pi}_{sr} = 1$ together with (4.20), equation (4.24) simplifies to

$$\delta = -\tilde{\mathcal{O}}_k^r.$$

Hence, from equation (4.23), the optimal estimate for $\tilde{\pi}_{sr}$ is

$$\tilde{\pi}_{sr} = \frac{\tilde{\mathcal{J}}_k^{sr}}{\tilde{\mathcal{O}}_k^r} = \frac{\gamma_k(\mathcal{J}_k^{sr})}{\gamma_k(\mathcal{O}_k^r)},$$

which concludes the proof.

□

In the completely general case, i.e., making no assumptions concerning the distribution of the noise term, it is not possible to derive the formulae for the re-estimation of the model parameters. Indeed, in order to use the EM algorithm one needs to change the measure from P^{θ} to $P^{\tilde{\theta}}$ which depends on the specific distribution function of the noise.

Furthermore, it is not possible to get "neater" recursive formulae for the parameters apart from the transition probabilities in the general case. One needs to resort to numerical methods to find the maximum likelihood. Nevertheless, as will be shown in the next section for the case of noise distributed as student t-distribution, it is possible to reduce the estimation problem to finding a zero of a function. In conjunction with the result in Theorem 4.5, we illustrate that the estimation problem becomes a relatively simple numerical problem which can be solved quickly using modern computers.

4.4.1 Student's t-distributed noise term

In this section, we shall analyse the model as in (4.2), where the noise term follows a student's t-distribution with ν degrees of freedom. Observe that the degrees of freedom in the noise term can in general depend on the state of the underlying Markov chain. Write $\nu = (\nu_1, \ldots, \nu_n)^\top$ for a vector representing the degrees of freedom and

$$y_{k+1} = \langle \alpha, X_k \rangle + \langle \beta, X_k \rangle z_{k+1}(\langle \nu, X_k \rangle), \tag{4.25}$$

where $\{z_k(\nu_i)\}$ is a sequence of independent random variables distributed as a student's t-distribution with ν_i degrees of freedom.

Theorem 4.6

Let $y_1, \ldots, y_k$ be sequence of observations available at time k and let the set of parameters $\{\pi_{sr}, \alpha_r, \beta_r\}$ determine the model. The EM estimates for vectors α and β, i.e., $\{\hat{\alpha}_r, \hat{\beta}_r\}$ then solve the equations

$$(\nu_i + 1)\gamma(\mathcal{T}^{(i)}(g_{\hat{\alpha}_i})X)_k = 0,$$

$$\text{where} \quad g_{\hat{\alpha}_i} : x \longmapsto \frac{\hat{\alpha}_i - x}{\beta_i \nu_i + (\hat{\alpha}_i - x)^2}, \quad \text{and}$$

$$(\nu_i + 1)\left(\gamma(\mathcal{T}^{(i)}(h_{\hat{\beta}_i})X)_k - \frac{\gamma(\mathcal{O}^{(i)}X)_k}{\hat{\beta}_i}\right) = 0,$$

$$\text{where} \quad h_{\hat{\beta}_i} : x \longmapsto \frac{\hat{\beta}_i \nu_i}{\hat{\beta}_i \nu_i + (x - \alpha_i)^2}.$$

Proof

Consider the parameters α_i, $i \in \{1, 2, \ldots, N\}$. To get the updates of parameters

$\hat{\alpha}_i$ from α_i we need to look at the factors

$$\begin{aligned}
&\lambda_{k+1}\big(X_k, y_{k+1}\big) \\
&\quad = \frac{\phi_{\langle X_k,\nu\rangle}\big(\frac{y_{k+1}-\langle X_k,\hat{\alpha}\rangle}{\langle X_k,\beta\rangle}\big)}{\phi_{\langle X_k,\nu\rangle}\big(\frac{y_{k+1}-\langle X_k,\alpha\rangle}{\langle X_k,\beta\rangle}\big)} \\
&\quad = \frac{\big(1+\frac{(y_{k+1}-\langle X_k,\hat{\alpha}\rangle)^2}{\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle}\big)^{-\frac{\langle X_k,\nu\rangle+1}{2}}}{\big(1+\frac{(y_{k+1}-\langle X_k,\alpha\rangle)^2}{\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle}\big)^{-\frac{\langle X_k,\nu\rangle+1}{2}}} \\
&\quad = \left(\frac{\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle + y_{k+1}^2 - 2y_{k+1}\langle X_k,\hat{\alpha}\rangle + \langle X_k,\hat{\alpha}\rangle^2}{\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle + y_{k+1}^2 - 2y_{k+1}\langle X_k,\alpha\rangle + \langle X_k,\alpha\rangle^2}\right)^{-\frac{\langle X_k,\nu\rangle+1}{2}}.
\end{aligned} \tag{4.26}$$

Write $\Lambda_{k+1}\big(X_k, y_{k+1}\big) = \prod_{l=1}^{k} \lambda_{l+1}\big(X_l, y_{l+1}\big)$ and consider a new measure P^* defined by

$$\frac{dP^*}{dP}\Big|_{\mathcal{G}_k} = \Lambda_{k+1}\big(X_k, y_{k+1}\big). \tag{4.27}$$

Now

$$\begin{aligned}
\log\Lambda_{k+1} &= -\sum_{l=1}^{k}\frac{\langle X_l,\nu\rangle+1}{2}\log\left(\frac{\langle X_k,\beta\rangle^2\langle X_l,\nu\rangle + y_{l+1}^2 - 2y_{l+1}\langle X_l,\hat{\alpha}\rangle + \langle X_l,\hat{\alpha}\rangle^2}{\langle X_k,\beta\rangle^2\langle X_l,\nu\rangle + y_{k+1}^2 - 2y_{l+1}\langle X_l,\alpha\rangle + \langle X_l,\alpha\rangle^2}\right) \\
&= -\sum_{l=1}^{k}\sum_{i=1}^{n}\langle X_l, e_i\rangle\frac{\nu_i+1}{2}\log\big(\beta_i\nu_i + y_{l+1}^2 - 2y_{l+1}\hat{\alpha}_i + \hat{\alpha}_i^2\big) + R(\alpha).
\end{aligned} \tag{4.28}$$

We wish to find the maximum of

$$\mathbb{E}\left[\frac{dP^*}{dP}\;\Big|\;\mathcal{Y}_k\right] \tag{4.29}$$

at $\hat{\alpha}$. To do this, we differentiate the expected value in (4.29) with respect to $\hat{\alpha}_i$ and equate the resulting derivative to 0. We have

$$\begin{aligned}
\frac{\partial}{\partial\hat{\alpha}_i}\mathbb{E}\big[\log(\Lambda_k)\mid\mathcal{Y}_k\big] &= \mathbb{E}\Big[\frac{\partial}{\partial\hat{\alpha}_i}\log(\Lambda_k)\;\Big|\;\mathcal{Y}_k\Big] \\
&= \mathbb{E}\Big[(\nu_i+1)\sum_{l=1}^{k}\langle X_{l-1}, e_i\rangle\frac{\hat{\alpha}_i - y_l}{\beta_i\nu_i + (\hat{\alpha}_i - y_l)^2}\;\Big|\;\mathcal{Y}_k\Big].
\end{aligned} \tag{4.30}$$

It can be seen from (4.30) that in order to maximise (4.29) one needs to find $\hat{\alpha}_i$ which makes the weighted sum of the differences $y_l - \hat{\alpha}_i$ (with weights roughly the square distance between y_l and $\hat{\alpha}_i$) equal to 0. There is no useful recursive expression for $\hat{\alpha}_i$, however. We find

$$\begin{aligned}\frac{\partial}{\partial\hat{\alpha}_i}\mathbb{E}\big[\log(\Lambda_k) \mid \mathcal{Y}_k\big] &= (\nu_i+1)\mathbb{E}\Big[\sum_{l=1}^{k}\langle X_{l-1}, e_i\rangle \frac{\hat{\alpha}_i - y_l}{\beta_i\nu_i + (\hat{\alpha}_i - y_l)^2} \;\Big|\; \mathcal{Y}_k\Big] \\ &= (\nu_i+1)\gamma(\mathcal{T}^{(i)}(g_{\hat{\alpha}_i})X)_k \\ \text{where} \quad g_{\hat{\alpha}_i} &: x \longmapsto \frac{\hat{\alpha}_i - x}{\beta_i\nu_i + (\hat{\alpha}_i - x)^2},\end{aligned} \tag{4.31}$$

which finishes the proof of the first part.

Remark 4.7

Suppose we re-write (4.30) *as*

$$\begin{aligned}&\frac{\partial}{\partial\hat{\alpha}_i}\mathbb{E}\big[\log(\Lambda_k) \mid \mathcal{Y}_k\big] \\ &\quad= \mathbb{E}\left[(\nu_i+1)\frac{\sum_{l=1}^{k}\langle X_{l-1}, e_i\rangle(\hat{\alpha}_i - y_l)\prod_{j=1,j\neq l}^{k}\left(\beta_i\nu_i + (\hat{\alpha}_i - y_j)^2\right)}{\prod_{l=1}^{k}\left(\beta_i\nu_i + (\hat{\alpha}_i - y_l)^2\right)} \;\middle|\; \mathcal{Y}_k\right] \\ &\quad= \mathbb{E}\left[\frac{\sum_{l=1}^{k}\langle X_{l-1}, e_i\rangle(\hat{\alpha}_i - y_l)(\nu_i+1)\prod_{j=1,j\neq l}^{k}\left(\beta_i\nu_i + (\hat{\alpha}_i - y_j)^2\right)}{\prod_{l=1}^{k}\left(\beta_i\nu_i + (\hat{\alpha}_i - y_l)^2\right)} \;\middle|\; \mathcal{Y}_k\right] \\ &\quad= \mathbb{E}\left[\frac{\sum_{l=1}^{k}\langle X_{l-1}, e_i\rangle(\hat{\alpha}_i - y_l)\big(\beta_i\nu_i^k + \mathcal{O}(\nu_i^{k-1})\big)}{\beta_i\nu_i^k + \mathcal{O}(\nu_i^{k-1})} \;\middle|\; \mathcal{Y}_k\right].\end{aligned} \tag{4.32}$$

It can quickly be seen that we recover known recursive formulae for the case of normally distributed noise terms for α_i as $\nu_i \to \infty$; see Theorem 3.5. In addition, considering that $\hat{\alpha}_i - y_l$ has a student's t-distribution, one should be able to recover the same formulae for α_i when $k \to \infty$.

In order to get the updated parameter $\tilde{\beta}_i$ from β_i, we need to consider the factors

$$\begin{aligned}
&\lambda_{k+1}\big(X_k, y_{k+1}\big) \\
&\quad= \frac{\phi_{\langle X_k,\nu\rangle}\big(\frac{y_{k+1}-\langle X_k,\alpha\rangle}{\langle X_k,\hat{\beta}\rangle}\big)}{\phi_{\langle X_k,\nu\rangle}\big(\frac{y_{k+1}-\langle X_k,\alpha\rangle}{\langle X_k,\beta\rangle}\big)} \\
&\quad= \frac{\big(1+\frac{(y_{k+1}-\langle X_k,\alpha\rangle)^2}{\langle X_k,\hat{\beta}\rangle^2\langle X_k,\nu\rangle}\big)^{-\frac{\langle X_k,\nu\rangle+1}{2}}}{\big(1+\frac{(y_{k+1}-\langle X_k,\alpha\rangle)^2}{\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle}\big)^{-\frac{\langle X_k,\nu\rangle+1}{2}}} \\
&\quad= \left(\frac{\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle\langle X_k,\hat{\beta}\rangle^2+\langle X_k,\beta\rangle^2\big(y_{k+1}-\langle X_k,\alpha\rangle\big)^2}{\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle\langle X_k,\hat{\beta}\rangle^2+\langle X_k,\hat{\beta}\rangle^2\big(y_{k+1}-\langle X_k,\alpha\rangle\big)^2}\right)^{-\frac{\langle X_k,\nu\rangle+1}{2}}.
\end{aligned} \tag{4.33}$$

Write $\Lambda_{k+1}\big(X_k, y_{k+1}\big) = \prod_{l=1}^{k} \lambda_{l+1}\big(X_l, y_{l+1}\big)$ and consider a new measure P^* defined via

$$\frac{dP^*}{dP}\Big|_{\mathcal{G}_k} = \Lambda_{k+1}\big(X_k, y_{k+1}\big). \tag{4.34}$$

Now,

$$\begin{aligned}
\log\Lambda_{k+1} = -\sum_{l=1}^{k}\frac{\langle X_l,\nu\rangle+1}{2}\Bigg(&\log\Big(\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle\langle X_k,\hat{\beta}\rangle^2 \\
&+\langle X_k,\beta\rangle^2\big(y_{k+1}-\langle X_k,\alpha\rangle\big)^2\Big) \\
&-\log\Big(\langle X_k,\beta\rangle^2\langle X_k,\nu\rangle\langle X_k,\hat{\beta}\rangle^2+\langle X_k,\hat{\beta}\rangle^2\big(y_{k+1}-\langle X_k,\alpha\rangle\big)^2\Big)\Bigg).
\end{aligned} \tag{4.35}$$

Similar to the situation of calculating the EM estimates for $\hat{\alpha}_i$, it is not possible to derive a recursive formula for $\hat{\beta}_i$ unless a standard normal noise term is assumed. However, it is still possible to re-estimate $\hat{\beta}_i$ numerically.

Following similar arguments as in equation (4.30), we need to differentiate (4.35) with respect to $\hat{\beta}_i$ and equate the resulting derivative to zero. Doing this, we have

$$
\begin{aligned}
\frac{\partial}{\partial\hat{\beta}_i}\mathbb{E}\big[\log(\Lambda_k)\mid\mathcal{Y}_k\big] &= \mathbb{E}\Big[\frac{\partial}{\partial\hat{\beta}_i}\log(\Lambda_k)\;\Big|\;\mathcal{Y}_k\Big]\\
&= \frac{\nu_i+1}{2}\mathbb{E}\Big[\sum_{l=1}^{k}\langle X_{l-1},e_i\rangle\Big(\frac{2\hat{\beta}_i\nu_i\beta_i^2}{\hat{\beta}_i^2\nu_i\beta_i^2+\beta_i^2(y_l-\alpha_i)^2}\\
&\quad+\frac{2\hat{\beta}_i\nu_i\beta_i^2+2\hat{\beta}_i(y_l-\alpha_i)^2}{\hat{\beta}_i^2\nu_i\beta_i^2+\hat{\beta}_i^2(y_l-\alpha_i)^2}\Big)\;\Big|\;\mathcal{Y}_k\Big]\\
&= \frac{\nu_i+1}{2}\mathbb{E}\Big[\sum_{l=1}^{k}\langle X_{l-1},e_i\rangle\Big(\frac{2\hat{\beta}_i\nu_i}{\hat{\beta}_i\nu_i+(y_l-\alpha_i)^2}\\
&\quad-\frac{1}{\hat{\beta}_i}\frac{2\beta_i^2\nu_i+2(y_l-\alpha_i)^2}{\beta_i^2\nu_i+(y_l-\alpha_i)^2}\Big)\;\Big|\;\mathcal{Y}_k\Big]\\
&= \big(\nu_i+1\big)\mathbb{E}\Big[\sum_{l=1}^{k}\langle X_{l-1},e_i\rangle\Big(\frac{\hat{\beta}_i\nu_i}{\hat{\beta}_i\nu_i+(y_l-\alpha_i)^2}-\frac{1}{\hat{\beta}_i}\Big)\;\Big|\;\mathcal{Y}_k\Big]\\
&= (\nu_i+1)\left(\gamma(\mathcal{T}^{(i)}(h_{\hat{\beta}_i})X)_k-\frac{\gamma(\mathcal{O}^{(i)}X)_k}{\hat{\beta}_i}\right)\\
&\qquad\text{where}\quad h_{\hat{\beta}_i}:x\longmapsto\frac{\hat{\beta}_i\nu_i}{\hat{\beta}_i\nu_i+(x-\alpha_i)^2}.
\end{aligned}
\tag{4.36}
$$

This completes the proof.

□

4.4.2 Extension to vector observations and independent drift and volatility

Similar to the previous chapter, it is possible to extend the result presented in Theorem 4.6 for vector observations as well as the case where the drift and volatility are independent. For brevity, we shall only present the general case here as all the specific cases follow immediately.

Formally, assume we have two independent Markov chains X_k^α and X_k^β with dynamics

$$
\begin{aligned}
X_{k+1}^\alpha &= \Pi_\alpha X_k^\alpha + V_{k+1}^\alpha\\
X_{k+1}^\beta &= \Pi_\beta X_k^\beta + V_{k+1}^\beta,
\end{aligned}
\tag{4.37}
$$

and state space $S_{X^\alpha} = \{e_1, e_2, \ldots, e_n\}$ and $S_{X^\beta} = \{e_1, e_2, \ldots, e_m\}$, respectively. Again, define $X_k = X_k^\alpha \otimes X_k^\beta$ and $\Pi = \Pi_\alpha \otimes \Pi_\beta$.

Write $\nu^j = (\nu_1^j, \ldots, \nu_n^j)^\top$, $j \in \{1, 2, \ldots, d\}$ for a set of vectors representing the degrees of freedom and assume the observation process is a d-dimensional vector process with components

$$\begin{aligned} y_{k+1}^1 &= \langle \alpha^1, X_k^\alpha \rangle + \langle \beta^1, X_k^\beta \rangle z_{k+1}^1(\langle \nu^1, X_k \rangle) \\ y_{k+1}^2 &= \langle \alpha^2, X_k^\alpha \rangle + \langle \beta^2, X_k^\beta \rangle z_{k+1}^2(\langle \nu^2, X_k \rangle) \\ &\vdots \\ y_{k+1}^d &= \langle \alpha^d, X_k^\alpha \rangle + \langle \beta^d, X_k^\beta \rangle z_{k+1}^d(\langle \nu^d, X_k \rangle). \end{aligned} \tag{4.38}$$

Here, $z_{k+1}^j(\langle \nu, X_k \rangle)$ are independent random variables following a student's t-distribution with ν degrees of freedom for each $j \in \{1, 2, \ldots, d\}$. Moreover, define $\underline{y}_k := (y_k^1, y_k^2, \ldots, y_k^d)^\top$ and write D for a diagonal matrix whose i-th $(1 \leq i \leq nm)$ element on the diagonal is

$$\prod_{j=1}^{d} \frac{\phi_{X_i}\left(\frac{y_l^j - \alpha_i^j}{\beta_i^j}\right)}{\beta_i^j \phi_{X_i}(y_l^j)}. \tag{4.39}$$

Let ξ_k be the unnormalised estimate of the state X_k. As in Lemma 4.3, it can be shown that

$$\xi_{k+1} = \Pi D \xi_k.$$

We would like to estimate the parameters of the model in (4.38). In other words, we wish to find the formulae for calculating the transition probabilities of the Markov chain Π and the drift vectors α^j and β^j for $j \in \{1, 2, \ldots, d\}$. As we have seen in the preceding discussion, the recursive expressions for occupation time, number of jumps and the observation process play a crucial role in deriving re-estimating formulae. We have the following result.

Theorem 4.8

If D is the diagonal matrix as defined in (4.39)*, the recursive relations for $\gamma(\mathcal{J}^{sr}X)_k$, $\gamma(\mathcal{O}^{r}X)_k$ and $\gamma(\mathcal{T}^{r}X)_k$ are*

$$\begin{aligned}
\gamma(\mathcal{J}^{sr}X)_k &= \Pi D(\underline{y}_k)\gamma(\mathcal{J}^{sr}X)_{k-1} + \langle \xi_{k-1}, e_r\rangle \prod_{j=1}^{d} \frac{\phi_{X_r}\left(\frac{y_k^j-\alpha_r^j}{\beta_r^j}\right)}{\beta_r^j \phi_{X_r}(y_k^j)} \pi_{sr} e_s \\
\gamma(\mathcal{O}^{r}X)_k &= \Pi D(\underline{y}_k)\gamma(\mathcal{O}^{r}X)_{k-1} + \langle \xi_{k-1}, e_r\rangle \prod_{j=1}^{d} \frac{\phi_{X_r}\left(\frac{y_k^j-\alpha_r^j}{\beta_r^j}\right)}{\beta_r^j \phi_{X_r}(y_k^j)} \Pi e_r \\
\gamma(\mathcal{T}^{r}(g)X)_k &= \Pi D(\underline{y}_k)\gamma(\mathcal{T}^{r}(g)X)_{k-1} + \langle \xi_{k-1}, e_r\rangle \prod_{j=1}^{d} \frac{\phi_{X_r}\left(\frac{y_k^j-\alpha_r^j}{\beta_r^j}\right)}{\beta_r^j \phi_{X_r}(y_k^j)} g(y_k^j) \Pi e_r.
\end{aligned} \tag{4.40}$$

Proof

The proof is similar to that of Theorem 4.4 and is omitted.

□

As previously noted, the transition probabilities for the Markov chain do not depend on the specific distribution of the noise term (see Theorem 4.5 and its proof). Coupled with the result in Lemma 3.9, we state the following whose proof is straightforward.

Lemma 4.9

If a sequence of observations $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_k$ are available at time k then the EM filter estimates for the transition matrix are given by

$$\begin{aligned}
\hat{\Pi} = \hat{\Pi}_\alpha \otimes \hat{\Pi}_\beta, \qquad & where \\
{}_\alpha\hat{\pi}_{rs} &= \frac{\sum_{j=1}^{m}\sum_{l=1}^{m}\gamma\left(\mathcal{J}_k^{(r-1)m+j,(s-1)m+l}\right)}{\sum_{j=1}^{m}\gamma\left(\mathcal{O}_k^{j+(r-1)m}\right)} \\
and \qquad {}_\beta\hat{\pi}_{rs} &= \frac{\sum_{j=1}^{n}\sum_{l=1}^{n}\gamma\left(\mathcal{J}_k^{r+(j-1)m,s+(l-1)m}\right)}{\sum_{j=1}^{n}\gamma\left(\mathcal{O}_k^{r+(j-1)m}\right)}.
\end{aligned} \tag{4.41}$$

All that is left to calculate are the expressions to re-estimate the elements of the vectors α and β in the multivariate observation model in (4.38). These are given in the next theorem.

Theorem 4.10

Let $y_1, \ldots, y_k$ be sequence of observations available at time k and let the set of parameters $\{\pi_{sr}, \alpha_r, \beta_r\}$ determine the model (4.38). The EM estimates for vectors α^p and β^p, i.e., $\{\hat{\alpha}_r^p, \hat{\beta}_r^p\}$ solve the equations

$$\sum_{j=1}^{m}(\nu_j^p+1)\gamma({}_\alpha\mathcal{T}^{(r)}(g_{\hat{\alpha}_{r,j}^p}(y^p))X)_k = 0$$

$$\text{where} \quad g_{\hat{\alpha}_{r,j}^p} : x \longmapsto \frac{\hat{\alpha}_r^p - x}{\beta_j^p\nu_j^p + (\hat{\alpha}_r^p - x)^2}, \quad \text{and}$$

$$\sum_{i=1}^{n}(\nu_r^p+1)\left(\gamma({}_\beta\mathcal{T}^{(r)}(h_{\hat{\beta}_{r,i}^p}(y^p))X)_k - \frac{\gamma({}_\beta\mathcal{O}^{(r)}X)_k}{\hat{\beta}_r^p}\right) = 0$$

$$\text{where} \quad h_{\hat{\beta}_{r,i}^p} : x \longmapsto \frac{\hat{\beta}_r^p\nu_r^p}{\hat{\beta}_r^p\nu_r^p + (x - \alpha_i^p)^2}.$$

Proof

We start the proof with the univariate observation case. Assume the observation process follows the specification

$$y_{k+1} = \langle \alpha, X_k^\alpha \rangle + \langle \beta, X_k^\beta \rangle z_{k+1}(\langle \nu, X_k \rangle). \tag{4.42}$$

Note that the model in (4.42) can be represented with $\underline{\alpha} = \alpha \otimes 1(m)$ and $\underline{\beta} = \beta \otimes 1(n)$ as in (3.16). We keep in mind the structure of $\underline{\alpha}$ and $\underline{\beta}$ and note that for any function $f(\cdot,\cdot)$

$$\sum_{i=1}^{mn}\langle X_k, e_i\rangle f(\underline{\alpha}_i, \underline{\beta}_i) = \sum_{i=1}^{n}\sum_{j=1}^{m}\langle X_k^\alpha, e_i\rangle\langle X_k^\beta, e_j\rangle f(\alpha_i, \beta_j).$$

Consequently, the EM estimates for α and β in the univariate case can be determined by solving the equations

$$\sum_{j=1}^{m}(\nu_j+1)\gamma({}_\alpha\mathcal{T}^{(r)}(g_{\hat{\alpha}_{r,j}}(y))X)_k = 0 \quad \text{and}$$

$$\sum_{i=1}^{n}(\nu_r+1)\left(\gamma({}_\beta\mathcal{T}^{(r)}(h_{\hat{\beta}_{r,i}}(y))X)_k - \frac{\gamma({}_\beta\mathcal{O}^{(r)}X)_k}{\hat{\beta}_r}\right) = 0.$$

This follows from the proof of Theorem 4.6. Here, the functions g and h are defined

in the Theorem 4.6.

The formulae for the multivariate case, which we need to prove, now follow using the same arguments as in the proof of Lemma 3.9.

□

4.4.3 Numerical application of the filters

In this section, we provide a numerical implementation of the results presented in the previous section; this concerns the applications of Theorems 4.6 and 4.4. We intend to investigate the performance of the filtering on a simulated data.

The filtering algorithm is tested on three sets of simulated data generated from a Markov chain process with two, three and four states. In each of the three examles presented below, 200 data points were generated by simulation using the reported initial values in Tables 4.1, 4.5 and 4.9 using the model

$$y_{k+1} = \langle \alpha, X_k \rangle + \langle \beta, X_k \rangle z_{k+1}(\langle \nu, X_k \rangle), \tag{4.43}$$

where $\{z_k(\nu_i)\}$ is a sequence of independent random variables following a student's t-distribution with $\nu_i = 3$ degrees of freedom. The filtering procedure is applied to the simulated data for three different numbers of states of the underlying Markov chain, with the re-estimation period containing 50 data points. In other words, the parameters were re-estimated once 50 new data points were processed. All calculations were performed in Matlab, on a 1.83Ghz dual core processor.

The results of the filtering algorithm with the calculation times for the three examples are reported. We also present the graphs of the simulated data versus the estimated values. In all the graphs, the estimated value v_i is calculated as $\alpha^\top \cdot \xi_i$, where ξ_i is the filtered state vector after i-th data point is processed. In each example, we state the errors of the estimated parameters, calculated as the second norm of the difference between the filtered parameters and the ones used to simulate the

data.

$$||\alpha - \hat{\alpha}||_2 = \sqrt{(\alpha_1 - \hat{\alpha}_1)^2 + \ldots + (\alpha_N - \hat{\alpha}_N)^2},$$
$$||\beta - \hat{\beta}||_2 = \sqrt{(\beta_1 - \hat{\beta}_1)^2 + \ldots + (\beta_N - \hat{\beta}_N)^2},$$
$$||\Pi - \hat{\Pi}||_2 = \sqrt{\text{eig}_{max}\big((\Pi - \hat{\Pi})^\top (\Pi - \hat{\Pi})\big)},$$

where eig_{max} denotes the largest eigenvalue of a matrix.

The error is calculated after each re-estimation of the parameter values and decreasing errors depict the improvement in filtering as more data is processed. Due to the nature of the EM algorithm behind the parameter re-estimation procedure, a good guess of initial values in filtering is required. This is elaborated in [43]. In all three examples presented below however, we are able to set initial values for the transition matrix Π in a random manner with some structure. That is, the elements of the matrix are drawn from a uniform distribution on $(0, 1)$ and then normalised to ensure the columns of Π sum up to one.

Here, we are not interested in trying to calculate the one-step (or more) ahead predictions; we are simply examining the performance of the filtering itself. Applications of the filtering algorithm to observed data are presented in the next chapter.

Example 4.11

The values used to simulate the data for the case of a two-state Markov chain are reported in Table 4.1 and the initial values for α and β used in filtering are reported in Table 4.2. The transition matrix Π used as an initial guess for filtering was random, i.e., its elements were drawn from the uniformly distributed random numbers on interval $(0, 1)$. The calculated final values of the parameters are reported in Table 4.3 and the errors on the parameter estimates are displayed in Table 4.4. The graph of the simulated data (blue) together with the estimated value is shown in Figure 4.1 and the total calculation time is 24.8 seconds.

Example 4.12

For the three-state Markov chain, the initial values for the data simulation are

$$\Pi = \begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.09 \\ 0.1 \end{bmatrix}.$$

Table 4.1: Values of parameters (Π, α, β) used in the simulation for a two-state Markov chain.

$$\alpha = \begin{bmatrix} 0.9 \\ -0.9 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.1 \\ 0.09 \end{bmatrix}.$$

Table 4.2: Initial values of parameters (α, β) used in filtering for a two-state Markov chain.

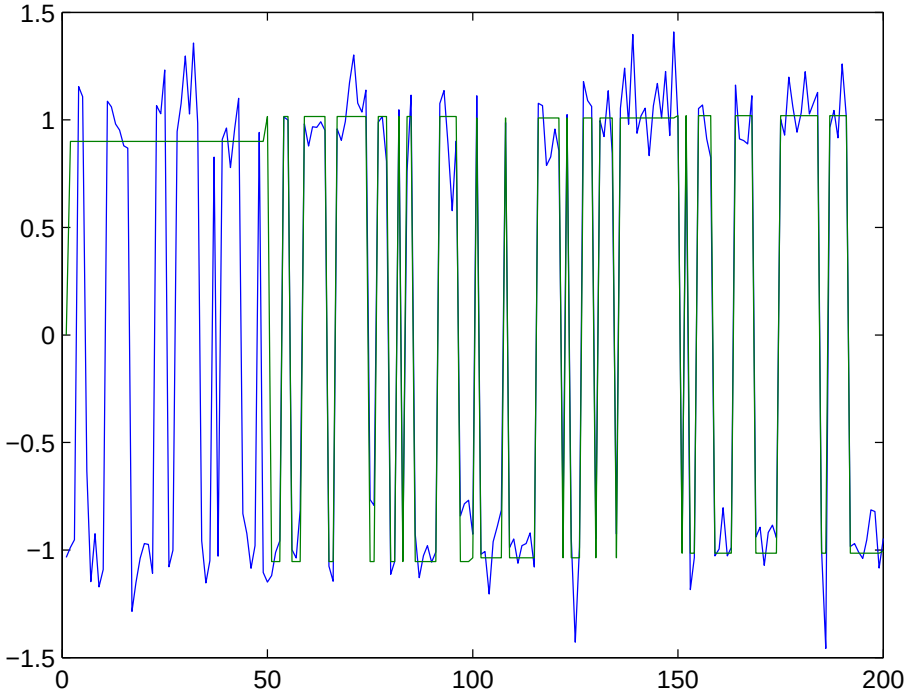

Figure 4.1: Simulated data (blue) with the estimated values (green).

$$\hat{\Pi} = \begin{bmatrix} 0.7590 & 0.2924 \\ 0.2421 & 0.7091 \end{bmatrix}, \quad \hat{\alpha} = \begin{bmatrix} 1.0205 \\ -1.0046 \end{bmatrix} \quad \text{and} \quad \hat{\beta} = \begin{bmatrix} 0.1073 \\ 0.1281 \end{bmatrix}.$$

.

Table 4.3: Final values of parameters (Π, α, β) calculated from the simulated data for a two-state Markov chain.

re-estimation number	errors		
	α	β	Π
1	0.0556	0.0396	0.1367
2	0.0375	0.0329	0.0763
3	0.0247	0.0330	0.0645
4	0.0210	0.0330	0.0599

Table 4.4: Errors of the estimated parameter values in the case of a two-state Markov chain.

reported in Table 4.5. The values for α and β used as initial guesses in the filtering process can be found in Table 4.6 whilst the initial guesses for the transition matrix are random, its elements were drawn from uniformly distributed random numbers on interval $(0, 1)$. The outputs of the filtering algorithm are reported in Table 4.7 with the errors reported in Table 4.8. The graph of the simulated data (blue) versus the estimated values is shown in Figure 4.2. Finally, the entire calculation took 56.3 seconds.

$$\Pi = \begin{bmatrix} 0.8 & 0.2 & 0.05 \\ 0.1 & 0.7 & 0.15 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.08 \\ 0.09 \\ 0.1 \end{bmatrix}.$$

Table 4.5: Values of parameters (Π, α, β) used to simulation for a three-state Markov chain.

Example 4.13

Finally, we report the results for the case of a four-state Markov chain driving the observation process. The values used for data simulation are reported in Table 4.9 and the initial guesses for the filtering algorithm in Table 4.10 with the transition matrix being random as in the previous two examples. The calculation time for the

$$\alpha = \begin{bmatrix} 0.01 \\ 0.9 \\ -0.9 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}.$$

Table 4.6: Initial values of parameters (α, β) used in filtering for a three-state Markov chain.

$$\hat{\Pi} = \begin{bmatrix} 0.8157 & 0.1641 & 0.0465 \\ 0.0982 & 0.7238 & 0.0618 \\ 0.0944 & 0.1168 & 0.8937 \end{bmatrix}, \quad \hat{\alpha} = \begin{bmatrix} 0.01335 \\ 1.0156 \\ -1.0063 \end{bmatrix} \quad \text{and} \quad \hat{\beta} = \begin{bmatrix} 0.1471 \\ 0.0855 \\ 0.0957 \end{bmatrix}.$$

Table 4.7: Final values of parameters (Π, α, β) calculated from the simulated data for a three-state Markov chain.

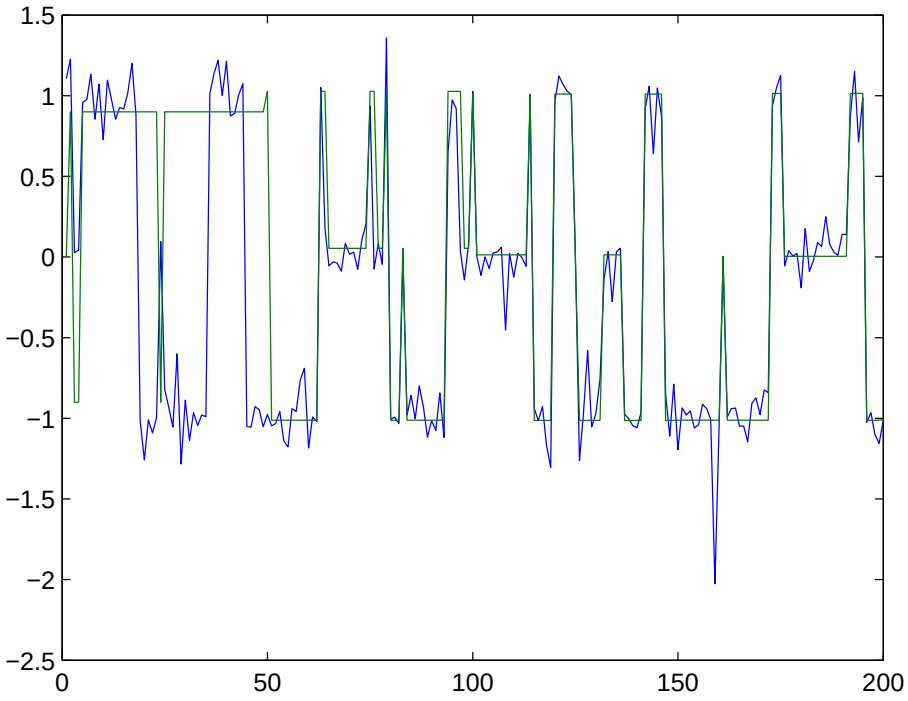

Figure 4.2: Simulated data (blue) with the estimated values (green).

re-estimation number	errors		
	α	β	Π
1	0.0607	0.0191	0.6159
2	0.0212	0.0574	0.2058
3	0.0192	0.0711	0.1066
4	0.0215	0.0674	0.1288

Table 4.8: Errors of the estimated parameter values in the case of a three-state Markov chain.

implementation under the four-state Markov chain is 120.1 seconds. The estimated parameter values for the vectors α and β can be found in Table 4.11 whilst the graph of the simulated data and the estimated values is displayed in Figure 4.3. The errors of the parameters for each re-estimation are given in Table 4.12.

$$\Pi = \begin{bmatrix} 0.8 & 0.15 & 0.05 & 0.05 \\ 0.1 & 0.7 & 0.05 & 0.05 \\ 0.05 & 0.1 & 0.8 & 0.1 \\ 0.05 & 0.05 & 0.1 & 0.8 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 0 \\ 0.5 \\ -0.5 \\ -1 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.06 \\ 0.07 \\ 0.08 \\ 0.09 \end{bmatrix}.$$

Table 4.9: Values of parameters (Π, α, β) used in the simulation for a four-state Markov chain.

$$\alpha = \begin{bmatrix} 0.01 \\ 0.4 \\ -0.4 \\ -1.2 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}.$$

Table 4.10: Initial values of parameters (α, β) used in filtering for a four-state Markov chain.

It is clear from the examples above that despite the departure from the normally distributed noise term, we can still filter the model parameters from the observed data. The calculations are, however, more demanding due to the fact there are no recursive formulae available and one needs to resort to numerical methods. As

$$\hat{\Pi} = \begin{bmatrix} 0.8184 & 0.1043 & 0.1299 & 0.0478 \\ 0.0995 & 0.7379 & 0.0898 & 0.1168 \\ 0.0441 & 0.0338 & 0.7745 & 0.0545 \\ 0.0442 & 0.1298 & 0.0151 & 0.7869 \end{bmatrix}, \quad \hat{\alpha} = \begin{bmatrix} -0.0115 \\ 0.5479 \\ -0.4837 \\ -1.0201 \end{bmatrix} \quad \text{and} \quad \hat{\beta} = \begin{bmatrix} 0.0935 \\ 0.0563 \\ 0.0762 \\ 0.0974 \end{bmatrix}.$$

Table 4.11: Final values of parameters (Π, α, β) calculated from the simulated data for a four-state Markov chain.

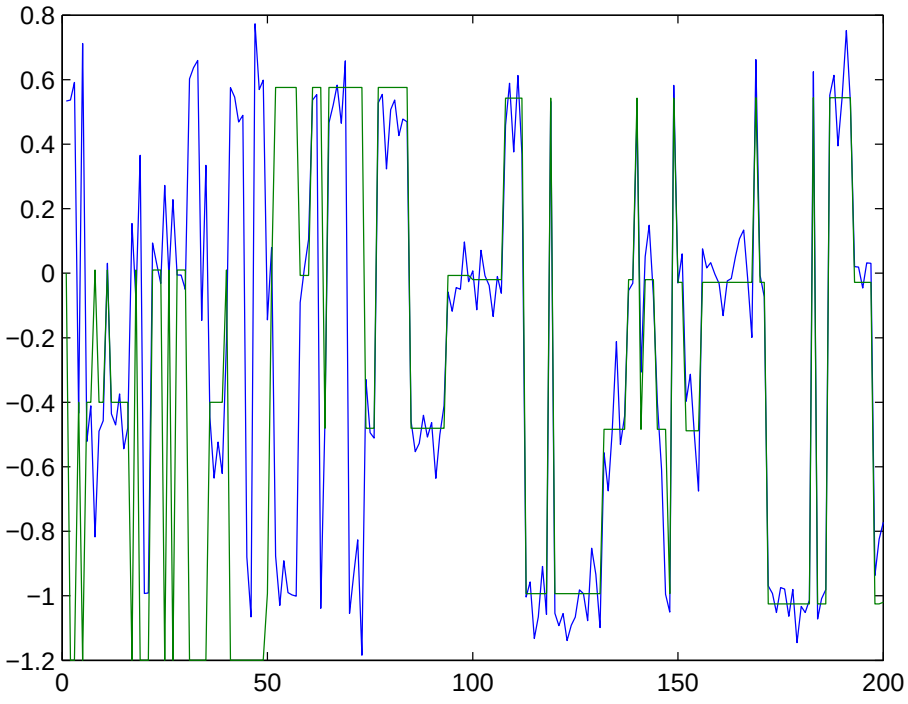

Figure 4.3: Simulated data (blue) with the estimated values (green).

re-estimation number	errors		
	α	β	Π
1	0.0901	0.0435	0.5193
2	0.0594	0.0473	0.1752
3	0.0462	0.0461	0.1575
4	0.0389	0.0371	0.1976

Table 4.12: Errors of the estimated parameter values in the case of a four-state Markov chain.

mentioned above, there is no maximisation involved, but it is necessary to find a zero of a function; refer to Theorem 4.6 for the calculation of the model parameters. Hence, the computation times are longer. Nevertheless, we feel that given the recursive formulae for re-estimating the transition probabilities, it is still worth going through the procedure of changing the measure to calculate the formulae to re-estimate the remaining model parameters.

It is also evident from Figures 4.1 to 4.3 that the estimated values follow very closely the state of the underlying Markov chain after the first parameter re-estimation. Parameters were re-estimated after processing 50 data points, however there is no notable difference in the goodness of fit after the second and third parameter update. Therefore, assuming the dynamics of the observation data does not change much, there is no need to increase the frequency of re-estimations, i.e., shorten the data length processed in each pass. A pass in this case comprises of 50 data points.

4.4.4 Application of the filters to observed market data

In the previous section, we have seen that the derived filters can be successfully used to estimate the parameters of the model on a simulated data set. In this section, we shall show that the filters can be sucessfully applied on the larger data set of observed market data. As in section 3.5, we use both the NASDAQ and DOW JONES datasets for the period 28 February 2003 – 16 February 2007.

Suppose S_k is a sequence of asset prices. Then we can observe the logarithmic

increments

$$y_k = \ln S_k - \ln S_{k-1} = \ln \frac{S_k}{S_{k-1}}$$

or

$$S_k = S_{k-1} \exp(y_{k-1}).$$

As in chapter 3, we assume the logarithmic increments are driven by a function f of the underlying Markov chain and some noise term, that is $y_k = f(X_k, z_{k+1})$. Here, $\{z_k\}$ is a sequence of IID random variables following a student's t-distribution with 3 degrees of freedom.

Data set	Number of MC states	RMSE	Computational time (s)
NASDAQ	2	3.2258×10^{-3}	86.3
	3	3.1522×10^{-3}	166.4
	4	3.1163×10^{-3}	303.3
DOW JONES	2	1.1243×10^{-3}	81.0
	3	1.0983×10^{-3}	153.9
	4	1.0859×10^{-3}	252.06

Table 4.13: Comparison of RMSEs and computational time in seconds for the DOW JONES and NASDAQ data.

The filters from the previous section are applied to both data sets whose summary statistics are given in Table 3.1. The data was processed in batches of 50 data points and after each batch was processed the parameters were re-estimated using the results of Theorem 4.6. Table 4.13 depicts the fitting errors (RMSEs) and the computational time in seconds needed to complete the calculations under the assumption of student's t-distributed noise term. Compared with the naive, no change model $E[y_{k+1} \mid y_k] = y_k$, which has the RMSEs of 7.2682×10^{-3} and 8.3338×10^{-3} for DOW JONES and NADAQ data sets respectively, the HMM based filters perform very well. In Figures 4.4 and 4.5, we present the plots of returns of the actual data in blue and the one step-ahead predictions in green for

a period 5 May 2004 to 18 February 2005. The plots of the returns as well as the actual observations together with one step ahead prediction for the whole period under consideration can be found in Appendix A. These figures are generated using a three-state Markov chain with the filtering procedure outlined earlier in the chapter. All computations were performed on a 1.83GHz dual core processor using Matlab.

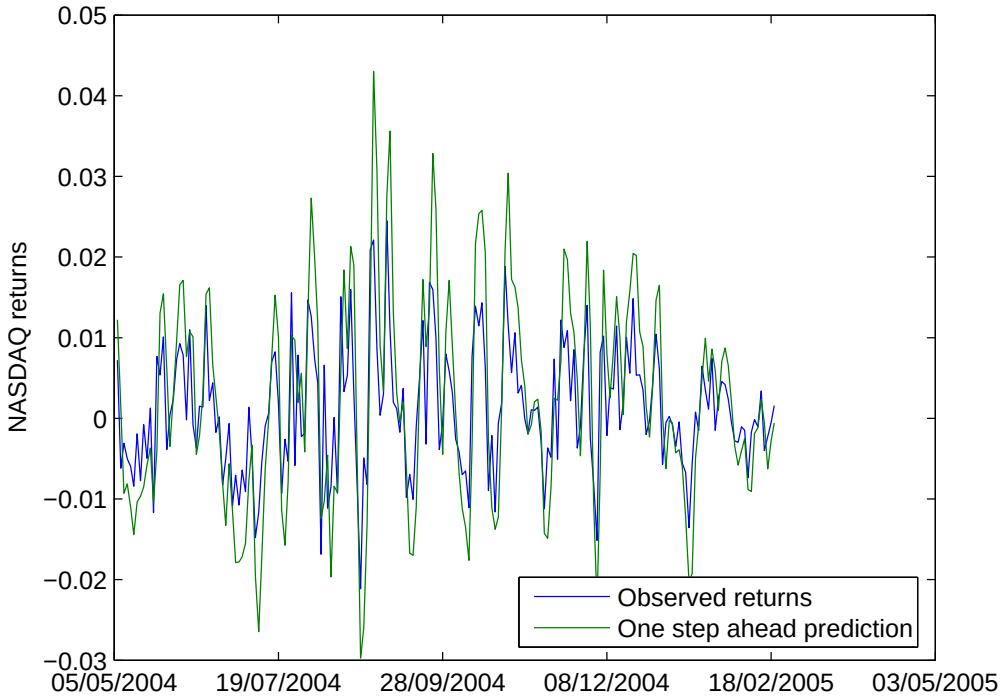

Figure 4.4: NASDAQ actual returns series (blue) and one-step ahead predictions (green).

4.5 Conclusions

In this chapter we revisited the estimation techniques from HMM filtering theory and extended it to include non-normal noise term distribution. Recursive formulae were obtained for re-estimation of transition probabilities of the underlying Markov chain for general noise term distribution. In addition we provided a general way to derive the re-estimation expressions for any noise term distribution; student's t-distribution is considered in particular for which all the re-estimation expressions

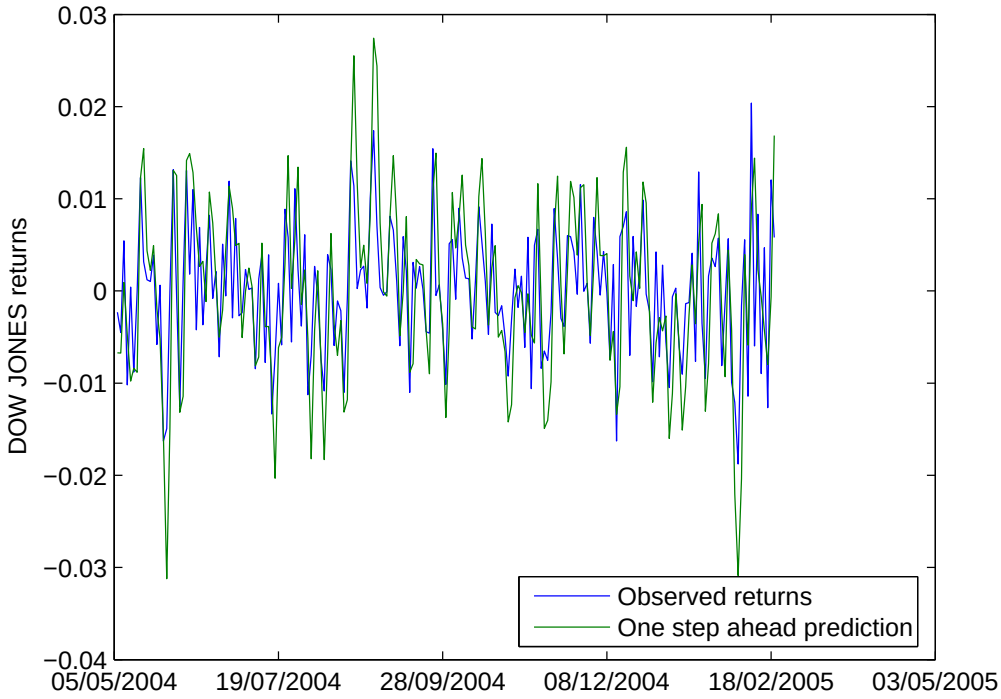

Figure 4.5: DOW JONES actual returns series (blue) and one-step ahead predictions (green).

are derived as well as implemented and tested on simulated and observed data sets. The estimation technique are not used only to obtain the best estimate of the Markov chain but also to re-estimate all the model parameters. We also explored the filtering of vector observations as well as the case of independent drift and volatility from previous chapter.

Chapter 5

A stochastic mortality model with HMM filtering

5.1 Introduction

Human mortality has improved significantly over the last few decades as documented in many actuarial, medical and scientific publications (see for example Macdonald *et al* [90], Currie, Durban and Eilers [33], and Renshaw and Haberman [101], [100]). Although this is a positive development it brought considerable stress in pension and health care support for the elderly. Furthermore, contrary to traditional and deterministic approaches to mortality modelling, mortality is now widely accepted to be evolving in a stochastic fashion. As mortality is by their very nature a primary source of risk for a large number of products in life insurance, pensions and some other recently issued financial instruments it is imperative to understand its dynamics better. In particular, recent mortality trends have proved particularly challenging for the pricing and reserving of long-term mortality-linked contracts, such as contracts providing living benefits.

Mortality will certainly continue to improve in the future with the advances made in the health sciences and medicine. This realisation, however, was not incorporated in mortality modelling even in the late 1970's. As noted in Bolton *et al* [16] and

Boyle and Hardy [21] amongst others, actuaries still then used out-of-date mortality tables without explicit allowance for future mortality improvements when pricing and reserving for mortality-based contracts.

Additionally, it is well-known that insurance companies are also exposed to financial risks, and since their investments are predominantly on fixed income investments it means they are also heavily exposed to interest rate risk. However, there is still a considerable gap in the tools available for modelling these two types of risk. As pointed out in Cairns, Blake and Dowd [24], stochastic modelling of interest rate is very well developed whilst the theory of stochastic mortality risk modelling is still at its infancy.

It can be observed that there are important similarities between mortality and interest rate modelling. Specifically, if we assume that the mortality process is driven by a force of mortality similar to the short rate in interest rate modelling, we can quickly deduce that they are both positive processes, have term structures and are fundamentally stochastic in nature. These similarities were exploited by Milevsky and Promislow [95], Cairns, Blake and Dowd [24], Biffis [13], Dahl [34] or Schrager [106]) to model force of mortality using tools and techniques developed for interest rate modelling.

Although we have been drawing parallels between the pricing of financial and mortality-linked instruments, there are some important differences and specific problems with mortality risk modelling. Eventhough it is an accepted fact that interest rates are mean-reverting, this is not the case with mortality rates. In particular, long-term stochastic improvements in mortality rates should not be mean reverting to some deterministic projection. Otherwise, the inclusion of mean reversion implies that if mortality improvements have been faster than what have been expected in the past then the potential for further mortality improvements will be significantly reduced in the future (see Cairns, Blake and Dowd [24]).

There are a number of recent studies that have sought to model mortality as a random process. The first milestone in stochastic mortality modelling was marked

by the work of Lee and Carter [84] that introduced a model for central mortality rates involving both age and time dependent terms. The model was applied to US population data where the time dependency was modelled using a univariate ARIMA time series. Their idea was later extended and improved by several authors including Renshaw and Haberman [101] and Brouhns, Denuit and Vermunt [23]. Another approach that also models mortality as a stochastic variable in discrete time was proposed by Lee [83]. Lee took a deterministic projection of spot mortality rates as given, and then apply an adjustment that evolves stochastically over time. Similar approach was later used in the work of Cairns, Blake and Dowd [25]. These models were developed in discrete time, but certain models were proposed to describe the dynamics of the force of mortality in continuous time. One of these, inspired by a previous work of Carriere [28], is contained in the study of Milevsky and Promislow [95] and assumed that the force of mortality $\mu(x,t)$ has a Gompertz form $\mu(x,t) = \zeta_0 \exp(\zeta_1 x + \sigma Y_t)$. In this Gompertz form, x refers to a life aged x and Y_t is an Ornstein-Uhlenbeck process satisfying the stochastic differential equation (SDE)

$$dY_t = -bY_t dt + dW_t. \tag{5.1}$$

In (5.1) W_t denotes a standard Wiener process. The process in (5.1) is expected to grow exponentially but exhibits a mean reversion. Dahl [34] further improved Milevsky and Promislow's approach and sought to model mortality intensity by a fairly general process that includes the mean-reverting Brownian Gompertz model. In [34], a class of processes is developed by supposing that the force of mortality for every fixed $x > 0$ is governed by the SDE

$$d\mu(t, x+t) = \alpha\big(t, x, \mu(t, x+t)\big)dt + \sigma^{\mu}\big(t, x, \mu(t, x+t)\big)dW_t.$$

It is shown that if the drift $\alpha\big(t, x, \mu(t, x+t)\big)$ and volatility $\sigma^{\mu}\big(t, x, \mu(t, x+t)\big)$ satisfy certain regularity conditions, the mortality model possesses an affine structure. That is, the survival probability $p(t, T, x)$ from time t to T $t < T$ for a person aged

$x+t$ given the information up to time t can be expressed as

$$\begin{aligned} p(t,T,x) &= \mathbb{E}\Big[\exp\big(-\int_t^T \mu(u,x+u)du\big)\Big|\mathcal{M}_t\Big] \\ &= \exp\big(A(t,T,x)+B(t,T,x)\mu(t,x+t)\big), \end{aligned} \tag{5.2}$$

where the deterministic functions $A(t,T,x)$ and $B(t,T,x)$ for a fixed x satisfy a system of ordinary differential equations (ODEs) involving α and σ^{μ}. The filtration $\mathcal{M}_t$ is a sequence of non-decreasing sigma-fields generated by the process μ. Since the ODEs associated with (5.2) are generalised Riccati equations (see Biffis [13]) they can be solved using standard numerical methods when explicit solutions are not available.

Recent work on affine mortality model was carried out by Luciano and Vigna [87]. The authors fitted different affine models to observed mortality data and compared their performance. The models considered in [87] are of the affine form, where the force of mortality follows either a CIR-like process or an Ornstein-Uhlenbeck (OU) process. It is established that the best fit is achieved when the force of mortality follows a non-mean reverting OU process, which confirms the theoretical requirement for the mortality models to be non-mean reverting.

In this chapter, we employ affine processes in modelling the stochastic evolution of mortality rates. Affine processes have been widely used in financial modelling and recently such processes have also been applied in describing mortality development. Our proposed model is based on a non-mean-reverting affine process, which was deemed suitable for modelling cohort mortality in previous studies. Within this framework, we include an analysis of trends in mortality behaviour via the HMM filtering techniques of the previous chapters. Optimal estimates of the model parameters are then obtained. Rather than modelling cohorts' survival one at a time, we demonstrate that our approach is able to generate directly the entire mortality surface.

5.2 Modelling framework and affine processes

In this section, the basic components of a stochastic mortality model and notation are introduced. Consider the force of mortality $\mu(t,x)$ for an individual aged x at time t. Traditional mortality models implicitly assume the force of mortality is independent of age (see for example, Bowers *et al* [18]), that is, $\mu(t,x) \equiv \mu(x)$ for all x and t. However, it is now widely recognised that over time mortality evolves in a stochastic manner. This is our motivation to model the force of mortality as a stochastic process in order to capture its time dependency and uncertainty of future developments.

Write

$$S(t,x) := \exp\left(-\int_0^t \mu(u,x+u)du\right) \tag{5.3}$$

for the survival function of a life aged x. Note that if the force of mortality $\mu(t,x)$ is deterministic then the survival function $S(t,x)$ is simply the probability that an individual aged x at time zero will survive until a later time t. Furthermore, if we assume that the force of mortality is time independent, i.e., $\mu(t,x+t) = \mu(x+t)$, expression (5.3) and results of any further analysis will simply reduce to those given in Bowers *et al* [18]. For example, under the assumption of deterministic and time independent force of mortality, formula (5.3) becomes

$$S(t,x) = \exp\left(-\int_x^{x+t} \mu(y)dy\right).$$

However, we intend to make the force of mortality stochastic in this current discussion. Certainly, the survival function $S(t,x)$ is a random variable. Note that $S(t,x)$ is a survival probability whose value can only be observed at time t rather than at the current time 0. In general, under a stochastic framework survival probabilities can be obtained by taking the expected value of the random variable $S(t,x)$.

Let $(\Omega, \mathcal{M}, P)$ be a probability space equipped with the filtration $\mathcal{M}_t$ generated

by the evolution of mortality $\mu(t,x)$ up to time t. In other words, $\mathcal{M}_t$ provides a full information of mortality development up to and including time t, but no information about how mortality rates will progress after time t.

On $(\Omega, \mathcal{M}, \mathbb{P})$ define the real-world or true survival probabilities measured at time t as follows. Let $p(t,T,x)$ be the real-world probability for an individual aged x at time 0 who is still alive at the current time t and survives until later time T. Then

$$p(t,T,x) = E\left[\frac{S(T,x)}{S(t,x)} \;\middle|\; \mathcal{M}_t\right]. \tag{5.4}$$

We note that $p(t,T,x)$ corresponds to ${}_{T-t}p_{x+t}$ in standard actuarial notation as defined for instance in Bowers *et al* [18]. Let $\tau(x)$ be a random residual lifetime of an individual aged x. In other words, $\tau(x)$ represents a future lifetime of an individual aged x. Similar to standard actuarial results (see chapter 3 of Bowers *et al* [18])

$$P\big(\tau(x) > T\big) = E\big[S(T,x)\big]. \tag{5.5}$$

Moreover, the $\mathcal{M}_t$-conditional density $f_t(.)$ of a random residual lifetime $\tau(x)$ of an individual aged x at time 0 on the set $\{\tau(x) > t\}$ is given by

$$f_t(s) = \mathbb{E}\Big[\mu(s,x+s)e^{-\int_t^s \mu(u,x+u)du} \;\Big|\; \mathcal{M}_t\Big]. \tag{5.6}$$

See Biffis [13] for further details concerning equation (5.6).

The financial literature on interest rate modelling is full of examples involving affine processes. One can draw similarities between interest rate and mortality modelling, however there are important differences as noted in Cairns, Blake and Dowd [24]. Whilst mean reverting processes turned out to be most appropriate for interest rate modelling, the force of mortality should not be mean reverting as was practicly confirmed in Luciano and Vigna [87].

It turns out that it is convenient to specify the force of mortality $\mu(t,x)$ via the

SDE

$$d\mu(t,x) = f(\mu(t,x))dt + g(\mu(t,x))dW_t + dJ_t, \tag{5.7}$$

where J is a pure jump process. Also, the drift $f(\mu(x,t))$, jump measure associated with J and covariance matrix $g(\mu(x,t))g(\mu(x,t))^\top$ have affine dependence on $\mu(t,x)$. Such processes are called affine processes; a thorough analysis of which can be found in Duffie, Filipovič and Schachermayer [39].

Affine processes prove to be very analytically tractable since under technical conditions provided in Duffie and Singleton [40]

$$\mathbb{E}\Big[e^{\int_t^T \Lambda(s,\mu(s,x))ds + a\mu(T,x)} \;\Big|\; \mathcal{M}_t\Big] = e^{\alpha(t,T)+\beta(t,T)\mu(t,x)}, \tag{5.8}$$

where $a \in \mathbb{R}$, $\Lambda(t,x)$ is an affine function in x and $\alpha(.)$ and $\beta(.)$ are functions solving uniquely a set of ODEs. These ODEs can be solved at least numerically and in some cases even analytically. Therefore the problem of finding the survival function in equation (5.5) becomes tractable when affine processes are used.

5.3 Mortality model

In an attempt to model mortality surface one needs to specify how mortality develops with time and age of the life in question. In other words, the behaviour of the force of mortality $\mu(t,x)$ in the time parameter t and age variable x needs to be analysed. Based on the previous work of Dahl [34] and Luciano and Vigna [87], we propose to model the cohort mortality rates via a non-mean-reverting affine process of the form (5.7), whilst the improvements in mortality rates for fixed age follow a function of Markov chain. In particular, we assume the Markov chain is independent from the Brownian motion driving the cohort mortality development.

We fitted two models to the observed cohort mortality data for the England and Wales male population born from 1841 to 1903. The first model assumes the cohort

mortality follows an Ornstein-Uhlenbeck (OU) process; section 5.3.1. The second model assumes the cohort mortality follows an OU process with jumps, see section 5.3.2. The mortality data is taken from the Human Mortality Database compiled by the University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). The data is also available at www.mortality.org or www.humanmortality.de. The mortality data set used in this thesis was downloaded on 10 September 2006.

5.3.1 The Ornstein-Uhlenbeck process without jumps

We assume that the cohort force of mortality follows a simple diffusion given by the equation

$$d\mu_t = a\mu_t dt + cdW_t, \tag{5.9}$$

as considered in Luciano and Vigna [87]. It can than be shown that the survival probability (5.4) can be expressed as

$$p(0,t,x) = e^{\alpha(t)+\beta(t)\mu(0,x)}, \tag{5.10}$$

where the functions α and β solve the system of ODEs

$$\begin{aligned} \beta'(t) &= -1 + a\beta(t) \\ \alpha'(t) &= \frac{1}{2}c^2\beta^2(t) \end{aligned} \tag{5.11}$$

with boundary conditions $\beta(0) = 0$ and $\alpha(0) = 0$. By solving the system in (5.11), we find α and β given by

$$\begin{aligned} \alpha(t) &= \frac{c^2}{2a^2}t - \frac{c^2}{a^3}e^{at} + \frac{c^2}{4a^3}e^{2at} + \frac{3c^2}{4a^3} \\ \beta(t) &= \frac{1}{a}\big(1 - e^{at}\big). \end{aligned} \tag{5.12}$$

We observe that the force of mortality μ_t modelled as a stochastic process in (5.9) can take negative values with positive probability. As a consequence we observe that for strictly positive values of c, the survival probability for sufficiently large t becomes an increasing function of age. In addition, the probability of surviving forever tends to infinity. This unrealistic and undesirable feature implies that the Ornstein-Uhlenbeck process might be considered inapropriate to describe the force of mortality.

However, it can be seen that in the applications this model performs very well. Calibration of the model to different mortality tables gives surprisingly good results, leading to very good fits of the survival probabilities $p(0, t, x)$. The reason for its successful application is that the values of a and c resulting from the calibration process are large and small, respectively. This fact is enough to make the probability of negative values of the force of mortality negligible; see further Appendix B and chapter 6. Furthermore, the period under consideration in applications is limited to a few decades during which the survival probability is a decreasing function of age and so, the explosion of the survival probability is avoided in the model. The above was also observed by Luciano and Vigna [87].

For each year with available data for cohort mortality, we fitted the model in (5.9) to the observed data using the least squares (LS) method, considering the spreads between different observed and model survival probabilities. In other words, (5.10) and (5.12) are fitted to the observed data using the LS technique. The initial value for $\mu(65, 0)$ has been chosen equal to $-\ln\big(p(65)\big)$ in the calibration of the mortality model as well as in all subsequent calculations.

For each year of the available data, the model in (5.10) was fitted and optimal parameter values were calculated. The calculated values of the parameters can be found in Table 5.2 (page 84).

Luciano and Vigna [87] showed that the chosen model fits the observed cohort mortality data well, however it does not explain the observed changes over time. As can be seen from Table 5.2 (page 84) the values of optimal parameters are varying

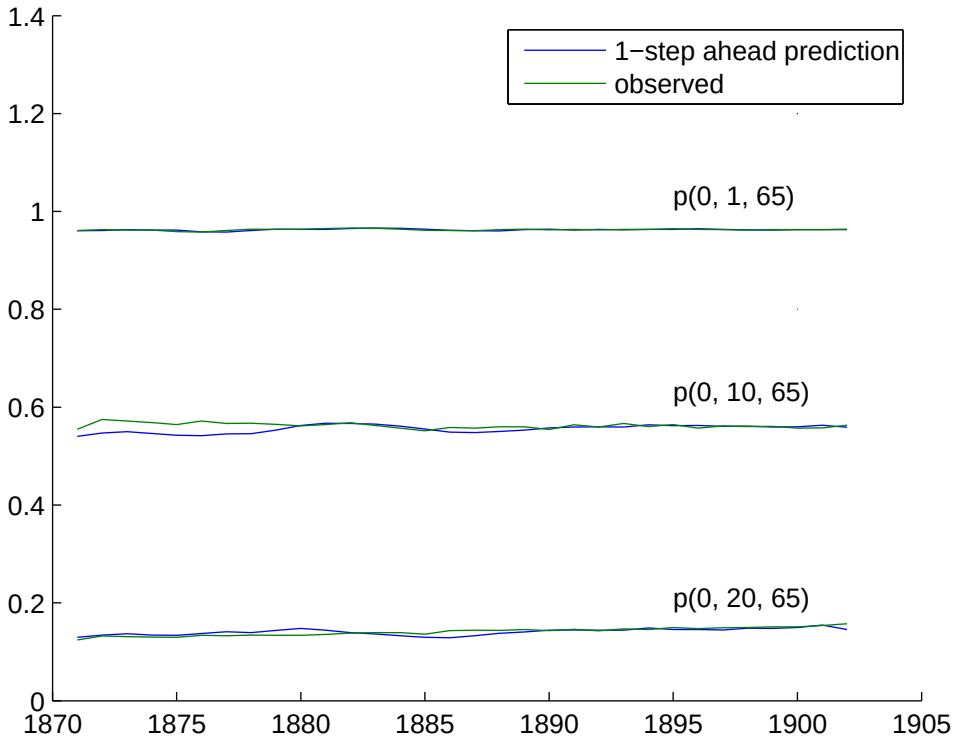

Figure 5.1: Observed and one-step ahead predicted values for $p(0,1,65)$, $p(0,10,65)$ and $p(0,20,65)$ under an OU model without jumps.

erratically over time. This means that not only the cohort mortality develops stochastically, but it also unpredictably varies from one cohort to another as was observed by Cairns, Blake and Dowd [25], [24], Biffis [13] and others.

In equation (5.9) and (5.10) the parameters a and c are assumend constant and positive. Nevertheless, the parameters can be varied through time as long as they are constant for each cohort.

We assume that both parameters of the model in (5.9) are driven by the same discrete-time Markov chain. We model the logarithmic increments of the parameter values simultaneously and employ HMM filtering techniques as proposed in chapter 3 to obtain optimal estimates of the logarithmic increments. These are in turn used to model future mortality rates. We use the observed cohort mortality data for males born from 1841 to 1903 so that we have mortality data for 62 years. We utilise the LS-estimated model parameters (a and c) as input values for the HMM filtering algorithm. In short, we filter the sequence of vector observations (a_k, c_k)

and assume they follow the dynamics

$$a_{k+1} = \langle \alpha_a, X_k \rangle + \langle \beta_a, X_k \rangle z^a_{k+1}$$
$$c_{k+1} = \langle \alpha_c, X_k \rangle + \langle \beta_c, X_k \rangle z^c_{k+1},$$

where X_k is a three-state Markov chain, α_a, α_c, β_a and β_c are real vectors of appropriate dimensions, whilst z^a_k and z^c_k are sequences of IID standard normal random variables.

Due to the small number of available observations, which is one of the problems in mortality modelling (see for example, [25] and [24]), we use half of the data set for "training" the model. After the first 30 data points are processed, one-step ahead predictions for parameters is performed and the predicted values are used to estimate future survival probabilities.

Error measure	OU without jumps	OU with jumps
$p(0,1,65)$:		
SSE	1.542×10^{-4}	9.717×10^{-5}
MSE	2.528×10^{-6}	1.593×10^{-6}
$p(0,10,65)$:		
SSE	1.105×10^{-2}	8.455×10^{-4}
MSE	1.812×10^{-4}	1.386×10^{-5}
$p(0,20,65)$:		
SSE	4.235×10^{-3}	4.939×10^{-4}
MSE	6.944×10^{-5}	8.098×10^{-6}

Table 5.1: Error analysis for cohort mortality predicions.

Figure 5.1 displays the predicted and observed values of survival probabilities $p(0,1,65)$, $p(0,10,65)$ and $p(0,20,65)$. The errors in terms of sum of squared error (SSE) and mean squared error (MSE) for the predicted survival probabilities in comparison to the observed data are reported in Table 5.1. It is clear from Table 5.1 that the errors are increasing with the time horizon of the survival probabilities, which is to be expected. The longer the time horizon, the bigger the uncertainity

and the quality of prediction reflects that. Further, it is apparent from Figure 5.1 that the HMM combined with an affine process can model the evolution of cohort mortality; the trend in mortality is closely matched by the model. It is obvious from Figure 5.1 (bottom most plot) that the 20-year survival probabilities increase with time and the trend is matched by the 1-year ahead prediction.

5.3.2 The Ornstein-Uhlenbeck process with jumps

We consider a model where we add a jump component to the stochastic part of the mortality process. Therefore, the model for the force of mortality has the SDE

$$d\mu(t,x) = a\mu(t,x)dt + c\mu(t,x)dW_t + dJ_t, \tag{5.13}$$

where J is a pure compound Poisson jump process with arrival times of intensity $d > 0$ and exponentialy distributed jump sizes with mean $\lambda < 0$. We also assume independence between the Brownian motion W_t and the Poisson process, as well as between the jump sizes. The choice of negative jump size is motivated by the expectation of sudden improvements in the force of mortality. Jumps should correspond to discontinuity of force of mortality that can be related to medical advancements, for instance. Theoretically, it is possible that the negative jumps in the intensity process could lead to negative intensity. This inconvenience is also observed by Luciano and Vigna [87], however in practical applications the sizes of jumps and jump frequency tend to be relatively small; therefore, the probability of negative values can be considered negligible.

Assuming equation (5.13) models the force of mortality, survival probabilities can be readily computed as in the previous section. This is given by

$$p(0,t,x) = e^{\alpha(t)+\beta(t)\mu(0,x)}, \tag{5.14}$$

where

$$\begin{aligned}\alpha(t) &= \Big(\frac{c^2}{2a^2} + \frac{da}{a-\lambda}\Big)t - \frac{c^2}{a^3}e^{at} + \frac{c^2}{4a^3}e^{2at} + \frac{3c^2}{4a^3} + \frac{d}{a-\lambda}\ln\big(1 - \frac{\lambda}{a} + \frac{\lambda}{a}e^{at}\big)\\ \beta(t) &= \frac{1}{a}\big(1 - e^{at}\big).\end{aligned} \tag{5.15}$$

As in the previous chapter, the parameters of the model (a, c, d, λ) are estimated for each year with the available data using the LS method. Those are in turn modelled as a vector process and we apply the HMM filtering techniques from the previous chapters. Formally, we have a sequence of vector observations $(a_k, c_k, d_k, \lambda_k)$ which are assumed to follow the specifications

$$\begin{aligned}a_{k+1} &= \langle \alpha_a, X_k\rangle + \langle \beta_a, X_k\rangle z^a_{k+1}\\ c_{k+1} &= \langle \alpha_c, X_k\rangle + \langle \beta_c, X_k\rangle z^c_{k+1}\\ d_{k+1} &= \langle \alpha_d, X_k\rangle + \langle \beta_d, X_k\rangle z^d_{k+1}\\ \lambda_{k+1} &= \langle \alpha_\lambda, X_k\rangle + \langle \beta_\lambda, X_k\rangle z^\lambda_{k+1},\end{aligned}$$

where as in the previous section X_k is a three-state Markov chain, α_k and β_k are vectors of appropriate dimensions and z^j_k, $j \in \{a, c, d, \lambda\}$ is a sequence of IID standard normal random variables. The error analysis (SSE and MSE) for this case is reported in Table 5.1 and a graph of the one-step ahead predictions is shown in Figure 5.2.

5.4 Conclusions

In this chapter, a model for the evolution of mortality, on which a financial contract may depend upon, is considered. The proposed model introduces the stochasticity in the evolution of cohort mortality in two directions. First, and in the literature the usual randomness, affects cohort mortality development with age. The observed data however suggests that cohort mortality is also developing in a stochastic fashion through time, which we attempted to model using hidden Markov filtering, thus in effect being able to model the whole mortality surface. We demonstrated with

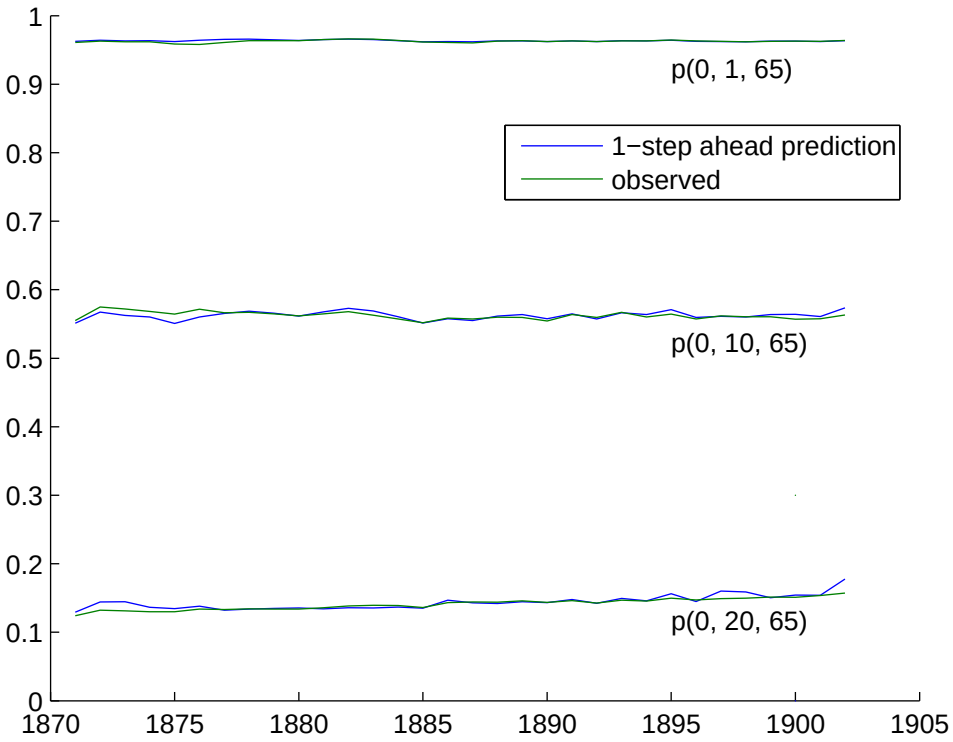

Figure 5.2: Observed and one-step ahead predicted values for $p(0, 1, 65)$, $p(0, 10, 65)$ and $p(0, 20, 65)$ under an OU model with jumps.

numerical examples that such a model can capture mortality developments reasonably well. In the next chapter, we extend the idea and examine how mortality developments coupled with known financial models impact the pricing of common mortality linked derivatives.

year	a	c	year	a	c
1841	0.09344	0.00505	1873	0.09019	0.00216
1842	0.08383	0.00189	1874	0.08854	0.00220
1843	0.08648	0.00376	1875	0.08985	0.00225
1844	0.08177	0.00186	1875	0.08985	0.00225
1845	0.08268	0.00216	1876	0.09098	0.00211
1846	0.08333	0.00189	1877	0.09135	0.00213
1847	0.08410	0.00193	1878	0.09085	0.00213
1848	0.08208	0.00188	1879	0.08863	0.00211
1849	0.08171	0.00192	1880	0.08966	0.00222
1850	0.08367	0.00208	1881	0.09707	0.00432
1851	0.08794	0.00189	1882	0.09418	0.00412
1852	0.09019	0.00221	1883	0.08499	0.00204
1853	0.09454	0.00361	1884	0.08400	0.00222
1854	0.08994	0.00186	1885	0.08317	0.00228
1855	0.09015	0.00218	1886	0.08232	0.00208
1856	0.09696	0.00420	1887	0.08332	0.00209
1857	0.09608	0.00435	1888	0.08351	0.00208
1858	0.09514	0.00427	1889	0.08285	0.00209
1859	0.09478	0.00451	1890	0.08258	0.00208
1860	0.09639	0.00505	1891	0.08324	0.00209
1861	0.09865	0.00469	1892	0.08391	0.00228
1862	0.09356	0.00426	1893	0.08259	0.00209
1863	0.08665	0.00220	1894	0.08638	0.00362
1864	0.08538	0.00221	1895	0.08120	0.00211
1865	0.08582	0.00216	1896	0.08572	0.00382
1866	0.08607	0.00220	1897	0.08113	0.00214
1867	0.08749	0.00220	1898	0.07977	0.00216
1868	0.08785	0.00219	1899	0.07962	0.00217
1869	0.08714	0.00218	1900	0.07828	0.00227
1870	0.08793	0.00216	1901	0.07771	0.00235
1871	0.09063	0.00220	1902	0.07843	0.00234
1872	0.09084	0.00207	1904	0.07752	0.00217

Table 5.2: Values of model parameters for OU-process without jumps using the LS technique and inputs to the HMM filtering algorithm.

Chapter 6

Valuation of contingent claims with mortality and interest rate risks

We consider the pricing of life insurance contracts under stochastic mortality and interest rates assumed not independent of each other. As in the previous chapters, we strongly rely on the method of change of measure together with the Bayes' rule for conditional expectations to obtain solution expressions for the value of common contracts. A demonstration of how to apply our proposed stochastic modelling approach to value survival and death benefits is provided.

In this chapter, we model mortality dynamics and asset prices using affine diffusion processes. We are then able to fully exploit analytical tractability of affine processes and derive pricing expressions for common life insurance contracts. We shall assume the parameters of the models are known and focus on the pricing expressions. In order to filter the parameters for the models, one can fall back on the discussion of filtering in the previous as well as succeeding chapters. We focus on the case where the assumption of independence between financial and mortality risk is dropped. The approach makes use of change of probability measure technique and application of Bayes theorem for conditional expectations, which are also the fundamental

methods employed to obtain the results in HMM filtering in the previous chapters.

This chapter is organised as follows. In section 6.1, we sketch the modelling framework. Section 6.2 provides general pricing formulae using forward and auxiliary measures for a large class of indexed life insurance contracts where both the development of mortality and interest rate dynamics are modelled stochastically. In section 6.3, we present an affine type specification for both interest and mortality rates and examine the corresponding implications in valuation. Section 6.4 presents a numerical implementation of our approach and section 6.5 concludes.

6.1 Modelling framework

In chapter 5, the rudiments of mortality modelling, which is a central concern in pricing pension and insurance contracts, were presented. We now focus our attention to the problem of valuing survival benefits (e.g., pension and life annuities) and death benefits (e.g., life and endowment insurance). To simplify the discussion, we consider generic survival and death contingent claims.

In valuing various products in the financial markets the risk-neutral valuation will be our starting point. In what follows we shall assume that there exists a risk-neutral measure $\mathbb{Q}$, absolutely continuous with respect to the real world measure P when pricing contracts linked to mortality dynamics. Examples of these contracts include endowments, insurance policies, pensions, guaranteed annuity options and mortality-linked bonds. The corresponding survival probability under the risk-neutral measure $\mathbb{Q}$ is given by

$$p_{\mathbb{Q}}(t,T,x) = E_{\mathbb{Q}}\left[\frac{S(T,x)}{S(t,x)} \;\middle|\; \mathcal{M}_t\right]. \tag{6.1}$$

The pricing of these contracts is conveniently simpler when certain assumptions in the underlying stochastic processes driving the financial and mortality variables are imposed. For a certain class of processes such as those which belong to the affine class, we can attain simplified and compact results.

Affine processes belong to a class of Markov processes with conditional characteristic function of the exponential affine form. This kind of processes will be utilised in the specification of mortality and interest rate dynamics in the succeeding sections. A thorough and comprehensive study and review of such processes are provided in Duffie *et al* [39] and Filipovič [55]. However, we adopt a more common approach in financial applications based on the definition of stochastic processes in terms of the solutions to specific SDEs in a given filtered probability space. An important result that we shall need later on processes with an affine structure is summarised in the following Theorem.

Theorem 6.1

Let $(\Omega, \mathcal{F}, P)$ be a probability space equipped with a filtration $\mathcal{F}_t$, large enough to support a standard d-dimensional Brownian motion W_t and suppose $\alpha(t,x)\colon [0,\infty]\times \mathbb{R} \to \mathbb{R}$ has an affine dependency on the second argument. In other words, let $\alpha(t,x) = \gamma(t) + \delta(t)x$. Assume further that $\beta(t)$ is an $\mathcal{F}$-previsible d-dimensional vector process which satisfies the growth condition

$$\mathbb{E}_P\left[\exp\big(\frac{1}{2}\int_0^T |\beta(u)|^2 du\big)\right] < \infty. \tag{6.2}$$

(i). If x_t is a stochastic process admitting the dynamics

$$dx_t = \alpha(t,x_t)dt + \beta(t)^\top \cdot dW_t, \tag{6.3}$$

it follows that conditional characteristic function has the exponential affine form

$$X(t,T) = \mathbb{E}\Big[e^{-\int_t^T x_u du} \;\Big|\; \mathcal{F}_t\Big] = e^{A(t,T)x_t + B(t,T)}, \tag{6.4}$$

where $A(t,T)$ and $B(t,T)$ are deterministic functions and $\top$ denotes the transpose of a vector.

(ii). If we define an equivalent measure $\mathbb{Q}$ on Ω via a Radon-Nikodym derivative

$$\Lambda_T := \frac{d\mathbb{Q}}{dP}\Big|_{\mathcal{F}_T} = \frac{\exp(-\int_0^T x_u du)}{X(0,T)},$$

then

$$W_t^{\mathbb{Q}} = W_t^P - \int_0^t A(u,T)\beta(u)du$$

is a standard n-dimensional Brownian motion under measure $\mathbb{Q}$.

Proof

The first part of the theorem is a known fact; a proof for which can be found in various sources. We refer to Biffis [13] and the references therein; specifically, Duffie *et al* [39] and Filipovič [55] provide a thorough treatment of affine processes.

To prove (ii) consider the Radon-Nikodŷm process

$$\Lambda_t = \mathbb{E}[\Lambda_T \mid \mathcal{F}_t] = \mathbb{E}\Big[\frac{d\mathbb{Q}}{dP} \;\Big|\; \mathcal{F}_t\Big] = \frac{\exp\big(-\int_0^t x_u du\big)X(t,T)}{X(0,T)}. \tag{6.5}$$

Therefore for every $s < t < T$

$$\Lambda_{s,t} = \frac{\exp\big(-\int_s^t x_u du\big)X(t,T)}{X(s,T)}.$$

A straightforward calculation shows that $\Lambda_{s,t}$ conditioned on $\mathcal{F}_s$ satisfies

$$\frac{d\Lambda_t}{\Lambda_t} = \frac{dX(t,T)}{X(t,T)} - x_t dt. \tag{6.6}$$

Now we calculate $\frac{dX(t,T)}{X(t,T)}$. Applying Itô formula to the representation in (6.4), we

have

$$\begin{aligned}\frac{dX(t,T)}{X(t,T)} &= \big(A_t(t,T)x_t + B_t(t,T)\big)dt + A(t,T)dx_t + \frac{1}{2}A(t,T)^2\langle dx_t, dx_t\rangle = \\ &= \big(A_t(t,T)x_t + B_t(t,T)\big)dt + A(t,T)\big(\alpha(t,x_t)dt + \beta(t)^\top \cdot dW_t\big) \\ &\quad + \frac{1}{2}A(t,T)^2\beta(t)^\top \cdot \beta(t)dt \\ &= \big(A_t(t,T)x_t + B_t(t,T) + A(t,T)\alpha(t,x_t) \\ &\quad + \frac{1}{2}A(t,T)^2\beta(t)^\top \cdot \beta(t)\big)dt + A(t,T)\beta(t)^\top \cdot dW_t.\end{aligned} \tag{6.7}$$

Taking into account equations (6.9) and (6.10), it may be verified that

$$\big(A_t(t,T)x_t + B_t(t,T) + A(t,T)\alpha(t,x_t) + \frac{1}{2}A(t,T)^2\beta(t)^\top \cdot \beta(t)\big)dt = x_t dt.$$

Consequently,

$$\begin{aligned}\frac{d\Lambda_t}{\Lambda_t} &= \frac{dX(t,T)}{X(t,T)} - x_t dt = x_t dt + A(t,T)\beta(t)^\top \cdot dW_t - x_t dt \\ &= A(t,T)\beta(t)^\top \cdot dW_t.\end{aligned} \tag{6.8}$$

Using a Cameron-Martin-Girsanov theorem (see Theorem 2.4), the result follows.

□

Remark 6.2

The deterministic functions $A(t,T)$ and $B(t,T)$ satisfy the system of ODEs

$$A_t(t,T) = 1 - \delta(t)A(t,T) \tag{6.9}$$

$$B_t(t,T) = -\gamma(t)A(t,T) - \frac{1}{2}\beta(t)^* \cdot \beta(t)A(t,T)^2, \tag{6.10}$$

where

$$A(T,T) = B(T,T) = 0.$$

Another approach to prove Theorem 6.1 (i), which solely relies on the method of stochastic flows and forward measure and does not involve the solution of the above

Ricatti equation can be found in Elliot and van der Hoek [47].

6.2 Integrating the interest and force of mortality models

In this section, we introduce the general setup of a combined modelling framework integrating both mortality and interest rate processes. We consider a filtered probability space $(\Omega, \mathcal{G}, \{\mathcal{G}_t\}, P)$, large enough to support a process r representing the evolution of interest rates and a process μ representing the evolution of mortality. The filtration $\{\mathcal{G}_t\}$ represents the information available up to time t revealed by the processes r and μ.

Write $\mathcal{M}_t \subset \mathcal{G}_t$ for the filtration generated by the evolution of mortality up to time t as in section 6.1. Similarly let $\mathcal{F}_t \subset \mathcal{G}_t$ be the filtration generated by the evolution of the short rate process r up to time t. Then we can express the filtration $\mathcal{G}_t$ as the smallest sigma-algebra generated by $\mathcal{M}_t$ and $\mathcal{F}_t$. Formally, we write

$$\mathcal{G}_t := \mathcal{M}_t \vee \mathcal{F}_t,$$

where $\mathcal{M}_t \vee \mathcal{F}_t$ is the filtration generated by $\sigma(\mathcal{M}_t \cup \mathcal{F}_t)$.

6.2.1 Interest rate model

We fix the stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ and take as given an adapted short rate process r_t such that it satisfies the technical condition $\int_0^t r(s)ds < \infty$ for all $t > 0$. The process r_t represents the continuously compounded rate of interest of a riskless security. Consider a riskless money market account B_t. The amount of money available at time t from investing one unit at time 0 is given by $B_t = \exp\Big(\int_0^t r(s)ds\Big)$. We suppose that an equivalent martingale measure $\mathbb{Q}$ exists, under which the gain from holding a risky security is a martingale after discounting by the money market account. From now on, we assume that the dynamics of all

security processes are specified under a risk-neutral measure $\mathbb{Q}$ unless otherwise stated. We note that the zero-coupon bond price $B(0,T)$ is given by

$$B(0,T) = \mathbb{E}_{\mathbb{Q}}\Big[\exp\big(-\int_0^T r(s)ds\big)\Big]. \tag{6.11}$$

6.2.2 Mortality model

When we assume that the force of mortality is governed by an affine process, closed-form solutions for the survival probabilities are explicitly obtained (Biffis [13] and Dahl [34]). This is also the implication of Theorem 6.1. Schrager [106] proposed to model the force of mortality $\mu(x,t)$ according to the specification

$$\mu(x,t) = g_0(x) + \sum_{i=1}^{M} X_i(t)g_i(x),$$

where $g_i : \mathbb{R} \to \mathbb{R}$ is some deterministic function and $X_i(t)$ is an affine process. Such set-up encapsulates traditional mortality models of Gompertz, Makeham or Thiele where the parameters are assumed to follow certain stochastic processes. Biffis [13], and Luciano and Vigna [87] also studied the force of mortality as a stochastic process of the form originally suggested by Dahl [34], which is

$$d\mu(x,t) = a\big(x,t,\mu(x,t)\big)dt + \sigma\big(x,t,\mu(x,t)\big)dW_t, \tag{6.12}$$

where $a(.)$ and $\sigma(.)$ are deterministic functions, and W_t is a standard Brownian motion. Although Biffis [13], and Luciano and Vigna [87] studied simplified versions of the model (6.12) they did not include age dependency in the drift and volatility functions of equation (6.12).

Needless to say, the force of mortality hugely depends on the age of the observed population. Therefore, the model would be more realistic if it includes this dependency. In addition, observed data suggest that the volatility of the force of mortality is highly age-dependent. In fact, there is evidence that it appears to be an exponential function of the age. In contrast to the approach of Schrager [106]

and Biffis [13], we make use of a model for mortality that is not mean-reverting to a pre-determined target level in this chapter.

Consider a general multivariate model for the force of mortality. Specifically, let process $\mu(x,t)$ be a non-mean-reverting and age dependent with affine functional form

$$d\mu(x,t) = \alpha\big(x,t,\mu(x,t)\big)dt + \sigma\big(x,t,\mu(x,t)\big)^{\top} \cdot dW_t, \tag{6.13}$$

where W_t is a standard d-dimensional Brownian motion. Here, α and σ are deterministic maps.

Special cases of the general model stated in (6.13) are two well-known processes. The first one is the simple Itô process with time and age dependent parameters which follows the SDE

$$d\mu(x,t) = \alpha\big(x,t\big)dt + \sigma^{\mu}\big(x,t\big)^{\top} \cdot dW_t, \tag{6.14}$$

and the second one is the Ornstein-Uhlenbeck type process

$$d\mu(x,t) = \alpha\big(x,t\big)\mu(x,t)dt + \sigma^{\mu}\big(x,t\big)^{\top} \cdot dW_t. \tag{6.15}$$

The processes in (6.14) and (6.15) are affine (Björk [14]). Hence, the survival probabilities can be expressed as

$$p(t,T,x) = \mathbb{E}\left[e^{-\int_t^T \mu(x+s,s)ds}\right] = e^{A(t,T,x)\mu(x,t)+B(t,T,x)}, \tag{6.16}$$

where $A(t,T,x)$ and $B(t,T,x)$ are deterministic functions. Under certain technical conditions stated in Nielsen [96], it is possible to derive explicit expressions for $A(t,T,x)$ and $B(t,T,x)$ in equation (6.16). When the force of mortality follows an

Itô process, then

$$\begin{aligned} A(t,T,x) &= -\int_t^T \int_t^u \alpha(x+s,s)ds\,du + \sum_{i=1}^{d} \int_t^T \sigma_i^{\mu}(x+s,s)(T-s)^2 ds \\ B(t,T,x) &= -(T-t). \end{aligned} \tag{6.17}$$

On the other hand, when the force of mortality is driven by an OU-like process in (6.15), we have

$$\begin{aligned} A(t,T,x) = &-\int_t^T e^{-K(x,u)} \int_t^u e^{K(x,s)} \alpha(x+s,s)ds\,du \\ &+ \sum_{i=1}^{d} \int_t^T \sigma_i^{\mu}(x+s,s)B(s,T,x)ds \\ B(t,T,x) = &-e^{K(x,t)} \int_t^T e^{-K(x,u)} du, \end{aligned} \tag{6.18}$$

where $K(x,t) = \int_0^t \alpha(x,u)du$. For the derivation of equation (6.18), we refer to Nielsen [96].

6.2.3 Independent case

In this subsection, we develop the fair valuation of two basic actuarial benefits involved in standard insurance contracts. The payoffs are contingent on survival or death of an individual over a pre-specified period of time and may be linked to other security prices. So far, no references were made to any specific model for interest rate dynamics; we only specified that the force of mortality is driven by an affine process.

We simply suppose as well that the dynamics of financial parameters are independent of the mortality development as assumed in previous investigations, such as in Cairns *et al* [24] and [25], Ballotta and Haberman [5] or Biffis [13]. The assumption that the short rate r_t and the force of mortality $\mu(x,t)$ are independent will significantly reduce the complexity of the pricing equations. It allows the separate pricing of mortality risk from the pricing of financial risks. This is reasonable for

a relatively short time horizon. We are aware though that in the long-run interest rates can be influenced by the relative size of population, which in turn, is influenced by mortality development (as well as fertility). Also in the short term, we recognise that a catastrophe event that seriously affects the size of population, such as major natural disasters or a nuclear war, can also affect interest rates.

Survival Benefit

Let C be a bounded $\mathcal{G}$-adapted process. The fair value $B_S(t,T,C_T)$ at time t of a survival benefit C_T to be paid out at time T $(t<T)$ for an individual aged x at time 0 can be written as

$$\begin{aligned} B_S(t,T,C_T) &= \mathbb{E}\Big[e^{-\int_t^T r(s)ds}1_{\{\tau>T\}}C_T \;\Big|\; \mathcal{G}_t\Big] \\ &= 1_{\{\tau>t\}}\mathbb{E}\Big[e^{-\int_t^T (r(s)+\mu(s,x+s))ds}C_T \;\Big|\; \mathcal{G}_t\Big], \end{aligned} \tag{6.19}$$

where $\tau=\tau(x)$ is the residual lifetime of a life aged x as defined in (5.5). When the dynamics of the interest rates are independent of the dynamics of mortality, then

$$B_S(t,T,C_T) = 1_{\{\tau>t\}}\mathbb{E}\Big[e^{-\int_t^T r(s)ds}C_T \;\Big|\; \mathcal{F}_t\Big]\mathbb{E}\Big[e^{-\int_t^T \mu(s,x+s)ds} \;\Big|\; \mathcal{M}_t\Big]. \tag{6.20}$$

Death Benefit

Assume again that C is a bounded $\mathcal{G}$-adapted process. The fair value at time t of a death benefit $B_D(t,T,C_\tau)$ of amount C_τ payable at the time of death in case the insured aged x at time 0 dies before time T, with $0\le t\le T$ can be expressed as

$$\begin{aligned} B_D(t,T,C_\tau) &= \mathbb{E}\Big[e^{-\int_t^\tau r(s)ds}1_{\{t<\tau\le T\}}C_\tau \;\Big|\; \mathcal{G}_t\Big] \\ &= 1_{\{\tau>t\}}\int_t^T \mathbb{E}\Big[e^{-\int_t^u (r(s)+\mu(s,x+s))ds}\mu(u,x+u)C_u \;\Big|\; \mathcal{G}_t\Big]du. \end{aligned} \tag{6.21}$$

The result in (6.21) is a direct consequence of expression (5.6). Again, if we assume

independence of interest rate and mortality dynamics, we have

$$B_D(t,T,C_\tau) = 1_{\{\tau>t\}} \int_t^T \mathbb{E}\Big[e^{-\int_t^u r(s)ds} C_u \;\Big|\; \mathcal{F}_t\Big] \times \mathbb{E}\Big[e^{-\int_t^u \mu(s,x+s)ds} \mu(u,x+u) \;\Big|\; \mathcal{M}_t\Big] du. \tag{6.22}$$

6.2.4 Dependent case

In subsection 6.2.3, we made the assumption that the dynamics of interest rates and mortality are independent. This assumption permits us to separate the evaluation of mortality risks from financial risks, thus enabling us to derive general pricing formulae for a generic class of life insurance contracts.

We drop this assumption in this section. We consider stochastic processes r_t and $\mu(x,t)$ to be dependent. We use a change of measure technique to derive pricing equations for the survival benefit $B_S(t,T,C_T)$ and death benefit $B_D(t,T,C_\tau)$. Once we obtain $B_S(t,T,C_T)$ and $B_D(t,T,C_\tau)$, it is straightforward to derive expressions for the values of life insurance contracts, such as endowments annuities and various types of insurance.

We shall be working under the forward measure P^T defined on a filtration $\mathcal{G}_t$ by setting the Radon-Nikodým derivative of P^T with respect to the risk-neutral measure $\mathbb{Q}$ as

$$\left.\frac{dP^T}{d\mathbb{Q}}\right|_{\mathcal{G}_t} = \Lambda_{0,T} = \frac{\exp\big(-\int_0^T r(s)ds\big)}{B(0,T)}, \tag{6.23}$$

where $B(0,T)$ is defined as in (6.11). Let $\mathbb{E}^T$ denote the expectation under the forward measure P^T. From Bayes' rule

$$\mathbb{E}^T[H \mid \mathcal{G}_t] = \frac{\mathbb{E}[\Lambda_{0,T} H \mid \mathcal{G}_t]}{\mathbb{E}[\Lambda_{0,T} \mid \mathcal{G}_t]}, \tag{6.24}$$

for a contingent claim H. Equation (6.24) together with equation (6.23) implies

that

$$\mathbb{E}^T[H \mid \mathcal{G}_t] = \frac{\mathbb{E}\Big[\exp\big(-\int_t^T r(s)ds\big) H \;\Big|\; \mathcal{G}_t\Big]}{B(t,T)},$$

or

$$\mathbb{E}\Big[e^{-\int_t^T r(s)ds} H \;\Big|\; \mathcal{G}_t\Big] = B(t,T)\mathbb{E}^T[H \mid \mathcal{G}_t]. \tag{6.25}$$

Equation (6.25) can provide an alternative formula for both $B_S(t,T,C_T)$ and $B_D(t,T,C_\tau)$. Taking $H = \exp\big(-\int_t^T \mu(s,x+s)ds\big)C_T$ in equation (6.25) and plugging this into equation (6.19), we obtain

$$B_S(t,T,C_T) = 1_{\{\tau>t\}} B(t,T)\mathbb{E}^T\big[e^{-\int_t^T \mu(s,x+s)ds} C_T \;\big|\; \mathcal{G}_t\big]. \tag{6.26}$$

The goal is to separate the calculation of expectations of interest rate dynamics, mortality and the benefit process under the forward measure.

To explicitly solve (6.26), we define an auxiliary measure $\tilde{P}^T$ that is absolutely continuous with respect to the forward measure P^T via the Radon-Nikodŷm derivative

$$\left.\frac{d\tilde{P}^T}{dP^T}\right|_{\mathcal{G}_t} = \tilde{\Lambda}_{0,T} = \frac{\exp\Big(-\int_0^T \mu(s,x+s)ds\Big)}{\tilde{p}(0,T,x)}, \tag{6.27}$$

where $\tilde{p}(0,T,x) = E^T\big[\exp(-\int_0^T \mu(s,x+s)ds)\big]$.

Let $\tilde{\mathbb{E}}^T$ denote the expectation under the auxiliary measure $\tilde{P}^T$. Then invoking equations (6.25) and (6.26) the survival benefit value $B_S(t,T,C_T)$ can be written as

$$B_S(t,T,C_T) = 1_{\{\tau>t\}} B(t,T)\tilde{p}(t,T,x)\tilde{\mathbb{E}}^T\big[C_T \;\big|\; \mathcal{G}_t\big]. \tag{6.28}$$

The main advantage of working with the forward measure is that the valuation is simpler when dealing with stochastic interest rates and force of mortality even without the independence assumption. In particular, we succeeded in splitting the

expectation of a product (equation (6.19)) into a product of expectations, where the expectation of the second term must be taken under the auxiliary measure. Since we started with the assumption that both the short rate process and force of mortality follow affine processes, an explicit expression can be derived for the dynamics of the benefit process C under an auxiliary measure $\tilde{P}^T$ given its dynamics under risk-neutral measure. This is made possible with the aid of Theorem 6.1.

In a similar manner, using a forward measure P^T we can also derive valuation expression for the death benefit value $B_D(t,T,C_\tau)$. Under the measure P^T, the expression for the value of death benefit in (6.21) is

$$B_D(t,T,C_\tau) = 1_{\{\tau>t\}} \int_t^T B(t,u)\mathbb{E}^T\Big[e^{-\int_t^u \mu(s,x+s)ds}\mu(u,x+u)C_u \;\Big|\; \mathcal{G}_t\Big]du.$$

In order to separate the dynamics of mortality from the dynamics of the benefit process, we define another auxiliary measure $\overline{P}^u$ via the Radon-Nikodým derivative

$$\left.\frac{d\overline{P}^u}{dP^T}\right|_{\mathcal{G}_t} = \overline{\Lambda}_{0,u} = \frac{\mu(u,x+u)\exp\left(-\int_0^u \mu(s,x+s)ds\right)}{\overline{f}_t(u)}, \tag{6.29}$$

where $t \le u \le T$ and $\overline{f}_t(u)$ is the $\mathcal{G}_t$-conditional density of a random residual lifetime $\tau(x)$ (see expression (5.6)) taken under the forward measure P^T. Applying the Bayes rule (6.24) and changing measure to the auxiliary measure $\overline{P}^u$ the value of death benefit is given by

$$B_D(t,T,C_\tau) = 1_{\{\tau>t\}} \int_t^T B(t,u)\overline{f}_t(u)\overline{E}^u\Big[C_u \;\Big|\; \mathcal{G}_t\Big]du, \tag{6.30}$$

where $\overline{E}^u$ denotes the expectation taken under the new auxiliary measure $\overline{P}^u$.

6.3 Example and illustration

An example depicting the applicability as well as demonstrating the flexibility of the proposed pricing approach within the context of affine models is provided in

this section. Suppose that both the short rate r_t and the force of mortality $\mu(t)$ follow stochastic processes of an affine type in its drift and volatility specification. Assume, for example, that the short rate follows a Vasiček model [110]with constant parameters, i.e., its dynamics is given by the SDE

$$dr_t = (a^r - b^r\, r_t)dt + c^r\, dW_t^r. \tag{6.31}$$

Assume further that the force of mortality follows a relatively simple diffusion given by the equation

$$d\mu_t = a^\mu \mu_t dt + c^\mu dW_t^\mu, \tag{6.32}$$

as considered in Luciano and Vigna [87]. In contrast to other developments of mortality modelling using affine processes and other attempts to price mortality-linked contracts, we do not assume the dynamics of the short rate and mortality development to be independent. Instead, we use the change of measure approach developed in subsection 6.2.4 to find explicit solutions for valuing basic mortality-linked instruments.

As shown in subsection 6.2.4, see equation (6.28), the price of survival benefit can be expressed as

$$B_S(t,T,C_T) = 1_{\tau>t} B(t,T)\tilde{p}(t,T,x)\tilde{\mathbb{E}}^T\big[C_T \,\big|\, \mathcal{G}_t\big]. \tag{6.33}$$

Assuming further that the policy holder has survived to current time t ($\tau > t$) equation (6.33) simplifies to

$$B_S(t,T,C_T) = B(t,T)\tilde{p}(t,T,x)\tilde{\mathbb{E}}^T\big[C_T \,\big|\, \mathcal{G}_t\big]. \tag{6.34}$$

If r_t has Vasiček dynamics then $B(t,T)$ has the explicit solution

$$B(t,T) = e^{A_r(t,T)r_t + B_r(t,T)}, \tag{6.35}$$

where $A_r(t,T)$ and $B_r(t,T)$ are deterministic functions. The second factor $\tilde{p}(t,T,x)$ is quite complex to evaluate. The difficulty comes from the fact that the expectation in

$$\tilde{p}(t,T,x) = \mathbb{E}^T\left[\exp\left(-\int_t^T \mu(s,x+s)ds\right)\middle|\mathcal{G}_t\right] \tag{6.36}$$

is taken under the T-forward measure P^T; see equations (6.23) to (6.26). In order to analytically derive the expression for $\tilde{p}(t,T,x)$, we need the dynamics of the force of mortality under the measure P^T. By Theorem 6.1, the evolution of μ_t under the T-forward measure follows the SDE

$$d\mu_t = (c^\mu c^r A_r(t,T) + a^\mu \mu_t)dt + c^\mu dW_t^{P^T}, \tag{6.37}$$

where $W_t^{P^T}$ is a standard Brownian motion under a measure P^T. Since the function $A_r(t,T)$ is deterministic, the dynamics of the force of mortality under a T-forward measure P^T admits an affine form. Therefore $\tilde{p}(t,T,x)$ can be expressed as

$$\tilde{p}(t,T,x) = e^{A_\mu(t,T)\mu_t + B_\mu(t,T)}. \tag{6.38}$$

Plugging (6.35) and (6.38) into equation (6.34), the value of survival benefit can be written as

$$B_S(t,T,C_T) = e^{A_r(t,T)r_t + B_r(t,T) + A_\mu(t,T)\mu_t + B_\mu(t,T)}\tilde{\mathbb{E}}^T\big[C_T \;\big|\; \mathcal{G}_t\big]. \tag{6.39}$$

From Theorem 6.1, given the dynamics of the benefit process C_t under a risk-neutral measure we can also derive the dynamics of the C_t under the auxiliary measure $\tilde{P}^T$. Consequently, given the dynamics of the actual benefit C_t, the survival benefit can be priced in an analytically tractable way without assuming the independence of the dynamics of the short rate from the dynamics of mortality development.

Suppose the benefit consists of a fixed unit amount plus a variable amount equal

to a percentage λ of the level of the short rate at the policy date. In other words

$$C_t = 1 + \lambda r_t. \tag{6.40}$$

In equation (6.39), we need to derive the dynamics of the short rate r_t under an auxiliary measure $\tilde{P}^T$. However, we know that under a risk-neutral measure $\mathbb{Q}$, the dynamics of the short rate is given by equation (6.31). Furthermore, it follows from Theorem 6.1 that

$$W_t^{P^T} = W_t^{\mathbb{Q}} - \int_0^t A_r(u,T)c^r du \tag{6.41}$$

is a standard Brownian motion under the T-forward measure P^T. A similar line of reasoning shows that

$$W_t^{\tilde{P}^T} = W_t^{P^T} - \int_0^t A_\mu(u,T)c^\mu du \tag{6.42}$$

is a standard Brownian motion under the auxiliary measure $\tilde{P}^T$. Combining (6.41) and (6.42) we can write down the dynamics of the short rate process r_t under the auxiliary measure $\tilde{P}^T$ as

$$dr_t = (a^r + (c^r)^2 A_r(t,T) + c^\mu c^r A_\mu(t,T) - b^r\, r_t)dt + c^r\, dW_t^{\tilde{P}^T}.$$

Given the dynamics of r_t under an auxiliary measure, it is straightforward to calculate the expectation in equation (6.39). Therefore, the value of survival benefit can be fully determined given a chosen functional form for the benefit process as in (6.40).

In a similar manner, one can also use the change of measure technique to derive an explicit expression for the price of death benefit in (6.30). With the same assumptions regarding the dynamics of the short rate and the force of mortality processes we get similar expression for the price of the death benefit $B_D(t,T,C_\tau)$. The only factor in the product under the integral in equation (6.30) that we need to

consider is $\overline{f}_t(u)$. Employing Theorem 6.1 and the fact that the force of mortality is an affine process, the $\mathcal{M}_t$-conditional density under the T-forward measure of a random residual lifetime in (5.6), can be written as

$$\overline{f}_t(T) = e^{A_f(t,T)\mu(t)+B_f(t,T)}\big(C_f(t,T)\mu(t) + D_f(t,T)\big) \tag{6.43}$$

following Duffie *et al* [39]. Therefore the death benefit in (6.30) has a fair value of

$$\begin{aligned} B_D(t,T,C_\tau) = \int_t^T & e^{A_r(t,u)r_t+B_r(t,u)}e^{A_f(t,u)\mu(t)+B_f(t,u)} \\ & \big(C_f(t,u)\mu(t) + D_f(t,u)\big)\overline{\mathbb{E}}^u\big[C_u \;\big|\; \mathcal{G}_t\big]du. \end{aligned} \tag{6.44}$$

6.4 Implementation

In order to calculate the prices of basic mortality-linked instruments we need to calibrate the mortality and interest rate models first. The parameters estimated from fitting the models to observed data can then be readily used to value instruments in both dependent and independent case. We choose a well-known Vasiček model (6.31) as a representation of interest rates. For the mortality model an Ornstein-Uhlenbeck (OU) process (5.9) is used; the OU process was found to generate the best fit to observed data amongst all tested affine processes without jumps in the study of Luciano and Vigna [87].

The mortality table selected for calibration is the observed UK generation tables for males born 1900. The observed mortality table is taken from the Human Mortality Database compiled by the University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). The data is also available at www.mortality.org or www.humanmortality.de. The mortality data set used in this implementation was downloaded on 10 September 2006. In fitting the data, we adopt the least squares method, considering the spreads between different observed and model survival probabilities. The initial value for $\mu(65,0)$ is set to $-\ln\big(p(65)\big)$ in the calibration of the mortality model as well as in all subsequent calculations.

T (years)	independent case $B_{S_i}(0,T,1)$	dependent case $B_S(0,T,1)$
1	0.9221	0.9347
2	0.8457	0.8736
3	0.7714	0.8163
4	0.6997	0.7624
5	0.6310	0.7119
6	0.5655	0.6645
7	0.5036	0.6200
8	0.4456	0.5783
9	0.3915	0.5391
10	0.3414	0.5024
11	0.2955	0.4679
12	0.2537	0.4356
13	0.2160	0.4053
14	0.1822	0.3770
15	0.1522	0.3505
16	0.1259	0.3257
17	0.1030	0.3025
18	0.0833	0.2808
19	0.0665	0.2605
20	0.0524	0.2416

Table 6.1: Actuarial fair prices of survival benefit for different times to maturity, both for independent and dependent case.

For the estimation of interest rate model parameters we use a different technique. A model was fitted to 1-month UK inter-bank loan data for the year 2003. The method selected in calibrating the interest rate model to LIBOR data is the Maximum Likelihood Method (see James and Webber [74]). We note that the data used for calibration of mortality and interest rate model are not consistent. However, as mentioned before, our intention here is to show how the fitted parameters can be used to calculate the prices of insurance products.

Once we calibrated the parameters of the interest rate and mortality models we can proceed to calculate the values of the contracts. For the data considered in this

chapter, the parameters for the interest rate model with specification in equation (6.31) are $a^r = 0.004089$, $b^r = 0.045398$ and $c^r = 0.003789$, whilst the parameters for the mortality model with specification (6.32) are $a^\mu = 0.078282$ and $c^\mu = 0.002271$. Table 6.1 displays the calculated prices of survival benefit with time to maturity between 1 and 20 years which pays 1 contingent on the survival to time T of an individual aged 65 (at current time 0). In the second column we present the actuarial fair values of survival benefit calculated under the assumption of independence between financial and demographic factors as in equation (6.20). The prices calculated without the assumption of independence between interest and mortality rates are depicted in the third column. In order to calculate values of the survival benefit $B_S(0, T, 1)$, the expression in (6.39) needs to be implemented. In our fairly simple case we assume the survival benefit to be deterministic, which simplifies the calculation of expected value under the auxiliary measure in (6.39). We note that working with random payoffs is straightforward provided we can express their dynamics under a risk neutral measure, as explained in section 6.3.

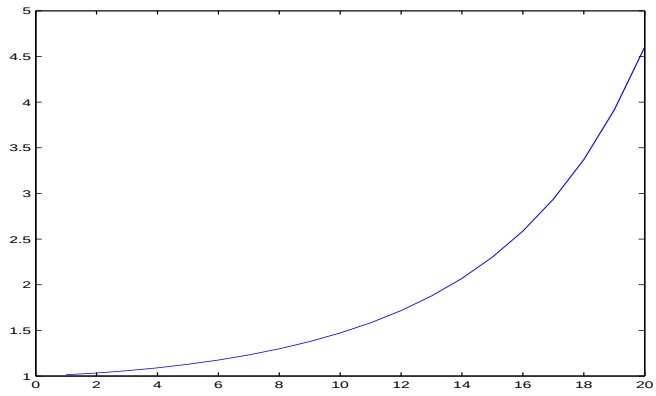

Figure 6.1: Relative difference $(B_S(0, T, 1)/B_{S_i}(0, T, 1))$ with respect to maturity.

From Table 6.1, it is apparent that the calculated survival benefit values differ considerably between the dependent and independent case. It can also be seen that the value of the survival benefit is noticeably higher when the independent assumption is dropped. The latter is somewhat expected since the independence assumption is relatively strong and disregards any correlation between demographic

and financial factors. Furthermore, it is evident from Table 6.1 that the relative difference between the prices calculated with and without an independence assumption is increasing. Indeed, the relative difference increases exponentially with time to maturity as shown in Figure 6.1.

The assumption of independence between mortality development and financial factors leads to a considerable lower prices of mortality-linked contracts compared to the prices generated under the dependence case. Therefore, based on the evidence suggested by our empirical work one cannot ignore the dependence between interest and mortality risks in pricing instruments with long term maturities.

For the calculation of the values of death benefit in (6.44), the expected value under an auxiliary measure needs to be evaluated numerically. Nonetheless, since the Radon-Nikodŷm derivative $\frac{d\overline{P}^u}{dP^u}$ can readily be expressed (see equation (6.29)), the calculation of the expected value of the benefit itself under an auxiliary measure is achievable; of course, this is assuming its dynamics under a risk-neutral measure $\mathbb{Q}$ is explicitly given.

We stress that the analytical solutions for pricing survival and death benefits might not be available when more complicated mortality and interest rate models are assumed. However, as long as both interest and mortality rate models are of affine type, analytical tractability can still be achieved. Equations (6.39) and (6.43) are essentially linked to generalised Ricatti equations, which can be solved using standard numerical methods as noted in Biffis [13].

6.5 Conclusions

In this chapter we extended previous works on stochastic mortality modelling. Independence between financial and demographic risk factors is not assumed. By using change of probability measures technique and employing affine diffusion processes in the description of both the evolution of mortality and interest rates, we showed that the framework developed could lead to analytical expressions for the pricing

problem for a number of life insurance contracts, either traditional, unit-linked or indexed. We presented numerical examples that demonstrate the applicability of our results and theoretical contributions. Numerical results clearly suggest that the dependence between mortality and financial factors should not be ignored when pricing and reserving for the instruments with long term maturities.

Part II

Contributions to sigma point filtering

Chapter 7

A new moment matching algorithm for sampling from partially specified symmetric distributions

In this chapter, we start with a brief review of Kalman filter and it best known extensions to aid the succeeding discussion. The main topic of this chapter however, is a development of a new algorithm for generating scenarios from a partially specified symmetric multivariate distribution. In particular, the algorithm generates samples which match the first two moments exactly and match the marginal fourth moment approximately, using a semidefinite programming procedure. The performance of the algorithm is also illustrated by a numerical example.

7.1 A short review of Kalman filter

Chapters 7 to 9 of this thesis will focus on sigma point filtering. We make theoretical and practical contributions to sample point generation and filtering heuristics. The fundamental background in this research endeavour is Kalman filtering and its

extensions, which are briefly reviewed in this chapter for a self contained presentation. We give a short account of the Kalman filter, its extension to cover nonlinear processes and this is being referred to as the extended Kalman filter (EKF), and the unscented or sigma point filter. This will motivate and aid the discussions in the succeeding chapters.

7.1.1 Kalman filter

The Kalman filter is an efficient recursive filter which estimates the state of a linear Gaussian state space system from a series of noisy observations. It is employed in a wide variety of engineering applications ranging from radar navigation, computer vision, climatology to financial modelling. It has also practical relevance to control theory and control systems engineering.

Kalman filters are linear dynamical systems discretised in the time domain and are modelled on a Markov chain built on linear operators perturbed by a Gaussian noise. At each time increment, a linear operator is applied to the state in order to generate a new state corrupted by noise, then another linear operator is applied to the generated state corrupted by noise to produce the observable outputs from the hidden state.

In particular, the Kalman filter model assumes the state at time k evolves from the state at time $k-1$ according to

$$\mathcal{X}(k) = F_k\mathcal{X}(k-1) + B_k\mathcal{U}(k) + G_k\mathcal{W}(k),$$

where F_k is the state transition matrix, B_k is the control input matrix applied to the control vector $\mathcal{U}(k)$ and G_k is the covariance matrix applied to multivariate standard normal process noise $\mathcal{W}(k)$.

At time k, the observation $\mathcal{Y}(k)$ of the true state $\mathcal{X}(k)$ is obtained according to

$$\mathcal{Y}(k) = H_k\mathcal{X}(k) + Q_k\mathcal{V}(k),$$

where H_k is the observation model matrix mapping the true state into the observed space and Q_k is the covariance matrix for the standard normal multivariate Gaussian noise process $\mathcal{V}(k)$.

For additional details on filtering and update equations for standard Kalman filters, refer to [4] for instance.

Remark 7.1

It is worth noting that Kalman filter can be regarded as analogous to HMM that was analysed in the previous chapters. The key difference is that the hidden state variables in Kalman filtering take values in a continuous state as opposed to the discrete state space in HMM filtering considered in this thesis. It is nevertheless possible to extend the HMM framework to the continuous state space (see [43]) providing a greater interlink between the two approaches. It is therefore not surprising that there exists a strong duality between the equations of the Kalman filter and those of the HMM. This fact is duly emphasised in [1] whilst the review of Linear Gaussian and other models can be found in [103].

7.1.2 Extended Kalman filter

The standard Kalman filter is limited by its linearity assumption. We know that many non-trivial systems are nonlinear. The nonlinearity in the system can be associated with either the observation model, the process model or in the worst case with both.

The extended Kalman filter is able to deal with nonlinearities in both the process and observation models and allows them to be nonlinear. The differentiability of the models, is however, required for EKF. The filter model takes the form

$$\begin{aligned} \mathcal{X}(k) &= \mathbf{f}(\mathcal{X}(k-1), \mathcal{U}(k)) + \mathcal{W}(k) \\ \mathcal{Y}(k) &= \mathbf{h}(\mathcal{X}(k)) + \mathcal{V}(k), \end{aligned} \tag{7.1}$$

where $\mathbf{f}$ and $\mathbf{h}$ are differentiable functions and $\mathcal{W}(k)$, $\mathcal{V}(k)$ are the process and observation noises both assumed to be zero mean multivariate Gaussian processes. At each time the Jacobian, a matrix of all first-order partial derivatives of a vector-valued function, is used with the current predicted states and these matrices can be used in Kalman filter equations. EKF then essentially linearises the nonlinear functions around the current estimate. The actual filtering and updating equations can be found in [4] and will not be elaborated here.

EKF can give reasonable performance and is very popular in the filtering of nonlinear processes. It does have its drawbacks though. Unlike the standard Kalman filter, EKF is not an optimal estimator because the mean and covariance are propagated through linearisation of the underlying nonlinear model. In addition, owing to its linearisation, the filter can diverge quickly if the initial estimate of the state is wrong or if the process is not modelled correctly.

7.1.3 Unscented Kalman filter

When the state transition and observation models, that is the predict and update functions $\mathbf{f}$ and $\mathbf{h}$ in the equation (7.1) are highly nonlinear, the EKF can produce poor results. Attempts to improve the performance of EKF thus give rise to the development of unscented Kalman filter (UKF), also a nonlinear filter. The poor performance of EKF is a consequence of propagating the mean and variance through the linearisation of the underlying nonlinear model. Under the UKF approach, a deterministic sampling technique is used to pick a minimal set of sample points (also called sigma points) around the mean. These are then propagated through the nonlinear functions from which the mean and covariance of the estimate can be recovered. This results to a filter that is able to capture the true mean

and covariance more accurately. Another benefit of the UKF over EKF is that it removes the requirement to explicitly calculate the Jacobian; for complex functions its calculation can be a difficult task on its own. The actual filtering formulae as well as the generation of sigma points needed in UKF can be found for example in [4], [76] and [54].

The UKF performs better than the EKF for highly nonlinear processes. Unfortunately, it still has drawbacks related to sigma point generation as well as the filtering heuristics when the noise term is not normally distributed. These drawbacks and problems are specified in detail and addressed in the succeeding chapters.

7.2 Introduction

Sampling from a partially specified multivariate distribution is a problem that arises in many different areas. Research work in this chapter was inspired by stochastic programming-based optimisation models in operations research, in which the key computational challenge is to generate scenarios from a distribution of the underlying random variables. For a large number of random variables, the scenario generation can be computationally very challenging. The distribution to be sampled from may not be available in closed-form and it may instead be characterised by moments obtained from empirical data. Even if the distribution is available in closed-form, it may be very difficult to sample and an approximation may be necessary. To deal with generation of scenarios under partially specified distributions one has a choice of several heuristic methods. Due to the burgeoning amount of literature on sampling from such probability distributions, we restrict our attention to the approaches used in operational research and finance only. These approaches can be divided roughly into two main classifications:

1. Under the first approach, the statistical properties of the joint distribution are specified in terms of moments, usually including the covariance matrix.

In [71], cubic transformation of univariate, standard normal random variables and Cholesky factorisation of covariance matrix are used to produce a multivariate distribution which approximately matches a given set of marginal central moments and the covariance matrix. Similar moment matching approach is employed to generate probability weights and support points using non-convex optimisation in [61]. In [107], entropy maximisation method is used to generate a discrete approximation to a given continuous distribution.

2. In the second approach, specified (parametric) marginal distributions are sampled independently and the samples are then used along with Cholesky factorisation of the covariance matrix to generate the necessary multivariate distribution. An iterative procedure of this type is described in [89].

Other approaches to scenario generation with specific emphasis on operations research applications include principal component analysis based simulation [109] and stochastic approximation based on transportation metrics ([98], [68]). A detailed survey of different scenario generation methods also appears in [79].

The approaches described above do have a considerable success in the practical applications of stochastic optimisation. Nevertheless, the procedures involved in the approaches of drawing samples from a partially specified multivariate distribution when it is specified in terms of moments have several limitations as detailed below:

1. All the moment-matching procedures in the above mentioned papers use non-convex optimisation to generate scenarios which match a specified set of statistical properties, in addition to a needed factorisation of the covariance matrix. Given a univariate random variable with known first 12 central moments, the approach used in [71] and [72] finds a cubic polynomial function of this random variable which has the required four central moments. This requires a non-convex optimisation in terms of the coefficients of the polynomial. The procedure has to be repeated iteratively for each marginal distribution. Similarly, the algorithm in [89] requires a non-convex optimisation over the space of lower triangular matrices.

2. The achieved moments of the generated samples match the target moments only approximately. There are two sources of error in these moment matching methods: one is due to the fact that only the local optima are found for the non-convex optimisation problem and the other is the inexact starting moments of samples of univariate random variables. Since these procedures employ samples from a known, "simple" univariate distribution, the achieved moments usually depend on the sample moments of univariate random variables used.

The primary objective of this chapter is to develop an algorithm based on convex optimisation which matches exactly the mean, covariance matrix and marginal (zero) skewness of a symmetric distribution and also matches the marginal fourth moments approximately (by minimising the worst case error between the achieved and the target marginal fourth moments). An analytic solution to this optimisation is known in the scalar case, as illustrated in subsection 7.3.3.

This algorithm may be used as a scenario generator on its own or its scalar version may be adopted to produce an initial guess for the optimisation routines proposed by other authors. Being able to match a small set of statistical properties exactly, possibly with a very small set of scenarios, may be preferable to generating a very large number of scenarios to model the entire distribution. This is especially true when the scenarios are to be used in stochastic optimisation procedures.

The rest of this chapter is organised as follows. In the next section, we introduce the notation, develop the main sampling algorithm of this chapter and provide a discussion of its properties. Section 7.4 presents a numerical study demonstrating the utility and effciency of the algorithm. Finally, section 7.5 concludes and outlines certain directions of further research.

7.3 The sampling algorithm

7.3.1 Notation

The following notation will be used in the development of the sampling algorithm.

m	number of random variables,
s	number of scenarios,
$\mathcal{X}$	discrete m-dimensional random variable,
$\mathcal{X}_i$	i^{th} random variable,
Φ	target mean vector for $\mathcal{X}$,
R	target covariance matrix for $\mathcal{X}$,
κ_i	target marginal 4^{th} central moment for the i^{th} random variable $\mathcal{X}_i$,
L_{ij}	entry in the i^{th} row and j^{th} column of a matrix L,
$\mathbb{P}(A)$	probability of an event A,
$\mathbb{E}[\mathcal{Y}]$	expected value of a random variable $\mathcal{Y}$.

Furthermore, let 1_s denote s-dimensional vector with all entries 1, $\text{diag}(x_i)$ denote a diagonal matrix with $x_1, x_2, \ldots, x_s$ on the diagonal. For a symmetric matrix M, we write $M \geq 0$ to indicate that the matrix is positive semi-definite, i.e. it has all non-negative eigenvalues.

To reiterate our objective, we aim to generate samples from a symmetric distribution with a specified mean vector Φ and a specified (positive definite) covariance matrix R. These target moments will usually be obtained from the data. If the covariance matrix obtained from the data is not positive definite, an adjustment may be necessary, such as the one suggested in [89]. In addition, we wish to minimise the worst case mismatch between the achieved marginal fourth moments and the target marginal fourth moments. We shall describe the algorithm from an optimi-

sation point of view first and then provide the closed-form solution in the scalar case. The rationale behind the key steps in the algorithm will become clear from the proofs of subsequent results and accompanying discussion.

7.3.2 Algorithm for moment matching scenario generation

The algorithm for generation of moment matching scenarios can be summarised in 4 steps given below.

1. Find a symmetric positive definite matrix L such that $R = LL^\top$. For a symmetric positive definite matrix R, matrix L is unique. If in addition matrix R has distinct eigenvalues, this may be found using singular value decomposition; see for example [69] and the references therein for methods of finding L. This matrix L is usually referred to as the square root of the matrix R.

2. Solve the following optimisation problem:

$$\min_{\epsilon, q_1, q_2, \ldots, q_s} \quad \epsilon \tag{7.2}$$

subject to

$$\begin{bmatrix} \epsilon & \psi_i \\ \psi_i & \epsilon \end{bmatrix} \geq 0, \quad i \in \{1, 2, \ldots, n\}, \tag{7.3}$$

$$\operatorname{diag}(q_k) \geq 0, \tag{7.4}$$

$$\begin{bmatrix} 1 & \mathbf{1}_s^\top \\ \mathbf{1}_s & \frac{1}{2m}\operatorname{diag}(q_k) \end{bmatrix} \geq 0, \tag{7.5}$$

where

$$\psi_i = \frac{1}{2s^2} \sum_{j=1}^{m} L_{ij}^4 \sum_{k=1}^{s} q_k - \kappa_i \quad \text{for} \quad i \in \{1, 2, \ldots, m\}.$$

Note that the above is a convex optimisation problem with a linear objective function and affine matrix inequality (AMI) constraints. These problems can be solved in polynomial time using interior point methods and extensive software packages are available to implement interior point methods for solving convex problems of this type (which are also called semidefinite programming problems); see for example [19], [56] and [112]. Let $\hat{q}_k; k = 1, 2, \ldots, s$ and $\hat{\epsilon}^2$ be the arguments which solve the above problem within a specified degree of accuracy.

3. Set $p_i = \frac{1}{\hat{q}_i}$, $i = 1, 2, \ldots, s$ and $p_{s+1} = 1 - 2m \sum_{i=1}^{s} p_i$.

4. Define a discrete m-dimensional random variable $\mathcal{X}$ over a support of $2ms+1$ points as follows:

$$\begin{aligned} &\mathbb{P}\left(\mathcal{X} = \Phi \pm \frac{1}{\sqrt{2sp_i}} L_j\right) = p_i, \\ &j = 1, 2, \ldots, m \quad \text{and} \quad i = 1, 2, \ldots, s, \\ &\mathbb{P}(\mathcal{X} = \Phi) = p_{s+1}. \end{aligned} \tag{7.6}$$

where L_j denotes the j^{th} column of matrix L.

Steps 1 – 4 constitute the entire set of procedures needed to construct the required samples. Before we prove that it has the required moment properties, we need to show that $p_i : i \in \{1, 2, \ldots, s+1\}$ defines a probability measure over the chosen $2ms + 1$ support points. Since $\hat{q}_i$ satisfies (7.4), it is immediate that $p_i = \frac{1}{\hat{q}_i} \geq 0, \quad \forall \quad i \in \{1, 2, \ldots, s\}$. It only remains to show that $p_{s+1} = 1 - 2m \sum_{i=1}^{s} p_i$ is non-negative. This is demonstrated in the following lemma.

Lemma 7.2

For $\hat{q}_i$ as defined above (see step 2 of the moment matching scenario generation algorithm),

$$2m \sum_{i=1}^{s} p_i = 2m \sum_{i=1}^{s} \frac{1}{\hat{q}_i} \leq 1.$$

Proof

The proof of Lemma 7.2 relies on the well-known property of the positive definite block matrices; namely,

$$M := \begin{bmatrix} A & B \\ C & D \end{bmatrix} \geq 0 \text{ and } D \geq 0 \iff A - BD^{-1}C \geq 0.$$

The block matrix $A - BD^{-1}C$ is called the Schur complement of D in M; see for example [19] for more details. In the present case, write

$$Q = \frac{1}{2m}\text{diag}(q_k).$$

Since by construction, for $i \in \{1, 2, \ldots, s\}$ q_i satisfies (7.4) and (7.5), we can write

$$\begin{bmatrix} 1 & \mathbf{1}_s^\top \\ \mathbf{1}_s & Q \end{bmatrix} \geq 0 \quad \text{and} \quad \text{diag}(q_k) \geq 0$$
$$\iff 1 - \mathbf{1}_s^\top Q^{-1} \mathbf{1}_s \geq 0,$$

from which the result follows directly.

□

Next, we shall establish a relationship between the optimal argument $\hat{\epsilon}$ (which is also equal to the optimal cost in step 2 of the moment matching scenario generation algorithm) and the target fourth marginal moments κ_i.

Lemma 7.3

For a discrete m-dimensional random variable $\mathcal{X}$ as defined in step 4 of the algorithm,

$$\max_i \left| \kappa_i - \mathbb{E}(\mathcal{X}_i - \Phi_i)^4 \right| \leq \hat{\epsilon}_i. \tag{7.7}$$

Proof

From the construction of sample points of a random variable $\mathcal{X}$, it follows that

$$\mathbb{E}(\mathcal{X}_i - \Phi_i)^4 = \sum_{k=1}^{s} p_k \frac{1}{2p_k^2 s^2} \sum_{j=1}^{m} L_{ij}^4 = \left(\sum_{k=1}^{s} \hat{q}_k \right) \frac{1}{2s^2} \left(\sum_{j=1}^{m} L_{ij}^4 \right).$$

Write

$$\hat{\psi}_i = \frac{1}{2s^2} \sum_{j=1}^{m} L_{ij}^4 \sum_{k=1}^{s} \hat{q}_k - \kappa_i.$$

By observing that $\hat{q}_k$ and $\hat{\epsilon}$ satisfy the constraint specified in (7.3) and

$$\begin{bmatrix} \hat{\epsilon} & \hat{\psi}_i \\ \hat{\psi}_i & \hat{\epsilon} \end{bmatrix} \geq 0 \quad \Longleftrightarrow \quad \hat{\epsilon}^2 - \hat{\psi}_i^2 \geq 0,$$

the result follows.

At this point, we collect all the moment matching properties of a random variable $\mathcal{X}$ as constructed by the algorithm in the following result.

Theorem 7.4

The distribution defined in (7.6) *satisfies the following properties:*

$$\mathbb{E}\big[\mathcal{X}\big] = \Phi \tag{7.8}$$

$$\mathbb{E}\big[(\mathcal{X} - \Phi)(\mathcal{X} - \Phi)^\top\big] = R \tag{7.9}$$

$$\mathbb{E}\big[(\mathcal{X}_i - \Phi_i)^3\big] = 0 \tag{7.10}$$

$$\max_i \big|\kappa_i - \mathbb{E}[(\mathcal{X}_i - \Phi_i)^4]\big| \leq \hat{\epsilon}. \tag{7.11}$$

Proof

Equations (7.8) and (7.10) follow immediately from the fact that the support points are symetrical around the mean vector Φ and (7.11) was proven in Lemma 7.3. We

therefore only need to prove (7.3) which follows by noting that

$$\mathbb{E}\big[(\mathcal{X}-\Phi)(\mathcal{X}-\Phi)^\top\big] = 2\sum_{i=1}^{s}\frac{p_i}{2sp_i}\left(\sum_{j=1}^{m} L_j L_j^\top\right) = \sum_{j=1}^{m} L_j L_j^\top = R.$$

□

There are several remarks that we would like to make concerning the above results.

(a) Note that the optimisation problem (7.3) – (7.5) finds the smallest $\hat{\epsilon}$ and the corresponding $\hat{q}_k$ such that (7.11) holds. In other words, the algorithm minimises an upper bound on the worst case error in matching the fourth marginal moment. A small value of $\hat{\epsilon}$ thus indicates that the fourth moment is approximately matched (with the maximum approximation error being $\hat{\epsilon}$ itself). This upper bound can be made zero in the scalar case, as will be seen in the next subsection.

(b) Even if the chosen $\hat{q}_k$ are not optimal, equations (7.8) – (7.11) will still hold provided $\hat{q}_k$'s satisfy the condition in Lemma 7.2 to define a probability measure. If we are not concerned with matching the fourth marginal moment, we may choose not to solve the optimisation problem and choose any $\hat{q}_k$ such that the condition in Lemma 7.3 holds, e.g., we can choose $\hat{q}_k > 2ms, \ \ \forall k$ which automatically satisfy the required condition. The actual choice of $\hat{q}_k$ subject to the lower bound $2ms$ can be made using any deterministic or stochastic algorithm. This provides s additional degrees of freedom, which may, in principle, be used to match other statistical properties (e.g., certain quantiles of interest). We have restricted our attention to matching fourth marginal moment only since matching these moments is relevant from a practical point of view and the associated optimisation, being convex, is numerically tractable.

(c) The downside, of course is that the algorithm is limited to symmetric distributions. However, even in cases when the underlying distribution is known to be asymmetric, the proposed algorithm may still have a useful role to play.

In the computation or optimisation of risk of a financial portfolio, the leptokurtic behavior of the loss distribution is often far more important than the asymmetry and a symmetric approximation which captures the tail behavior of loss distribution correctly may be admissible.

(d) In a somewhat unrelated field, similar sample point generation methods are also employed in the development of sigma point filters (also called unscented filters) widely adopted in engineering; see [76] and the references therein. These methods have become quite popular as a computationally cheaper alternative to particle filters for state estimation problems in nonlinear systems. However, the sampling methods in the existing sigma point filtering techniques do not guarantee that the weights assigned to each sample point will always be nonnegative. Our proposed algorithm for sampling distributions avoids this problem and its application in sigma point filtering has now been reported in [35] and is also explored in the remaining chapters (8 and 9) of this thesis.

7.3.3 Closed-form solution for the scalar case

We have demonstrated that finding positive q_k satisfying $2m\sum_{k=1}^{s} q_k^{-1} < 1$ and minimising the worst case error in matching the fourth marginal moment is a convex optimisation problem. A natural question to ask is whether it is possible to find a closed-form solution to this problem in specific instances. As mentioned earlier, choosing $q_k > 2ms, \ \forall k$ will automatically satisfy the necessary constraint on the sum of q_k^{-1}. At this stage, it still remains to be seen whether we can choose $q_k > 2ms$ which will also satisfy the condition for matching the fourth moment, i.e., whether we can choose q_k such that $\kappa_i = \mathbb{E}(\mathcal{X} - \Phi)^4$ holds. In the scalar case, i.e., when $m = 1$, the answer is affirmative as shown in the following Lemma.

Lemma 7.5

Suppose that $m = 1$ *and that* $\frac{\kappa_1}{L_{11}^4} > 1$. *In addition, let* $\hat{q}_i \in [2ms, \psi]$, $i \in$

$\{1, 2, \ldots, s-1\}$ be $s-1$ real numbers, let the constant ψ be given by

$$\psi = \frac{2s^2\kappa_1}{(s-1)L_{11}^4} - \frac{2s}{s-1},$$

and let

$$\hat{q}_s = \frac{2s^2\kappa_1}{L_{11}^4} - \sum_{i=1}^{s-1} \hat{q}_i. \tag{7.12}$$

Finally, let $\mathcal{X}$ be as in (7.6) *with $m = 1$. Then the random variable $\mathcal{X}$ satisfies the properties given in* (7.8) – (7.10) *of Theorem 7.4. In addition,*

$$\kappa_1 = \mathbb{E}\big[(\mathcal{X} - \Phi)^4\big].$$

Proof

Verifying that the random variable $\mathcal{X}$ satisfies properties (7.8) – (7.10) is straightforward. In fact, one can prove this using exactly the same arguments as in the proof of Theorem 7.4. To verify that $\kappa_1 = \mathbb{E}\big[(\mathcal{X}-\Phi)^4\big]$ holds, first note that $\hat{q}_i \leq \psi$ for $i \in \{1, 2, \ldots, s-1\}$ from the set-up of Lemma 7.5, which ensures that $\hat{q}_s > 2ms$. Therefore,

$$\mathbb{E}\big[(\mathcal{X} - \Phi)^4\big] = \sum_{k=1}^{s} \frac{p_k L_{11}^4}{2p_k^2 s^2} = \left(\sum_{k=1}^{s} \hat{q}_k\right) \frac{L_{11}^4}{2s^2}.$$

Plugging in the definition of $\hat{q}_s$ from (7.12) one gets

$$\mathbb{E}\big[(\mathcal{X} - \Phi)^4\big] = \left(\sum_{k=1}^{s-1} \hat{q}_k + \frac{2s^2\kappa_1}{L_{11}^4} - \sum_{i=1}^{s-1} \hat{q}_i\right) \frac{L_{11}^4}{2s^2} = \kappa_i,$$

which completes the proof.

□

Remark 7.6

Note that the condition $\frac{\kappa_1}{L_{11}^4} > 1$ is not particularly restrictive, and is in fact satisfied by all elliptic distributions including Gaussian distribution and t-distribution; see

[12].

The above result gives an optimisation-free methodology of matching the first four moments of a symmetric scalar random variable. This fact is quite important in itself and our algorithm as proposed above can be used as an efficient alternative to cubic transformation-based approaches for generating random samples which match a given set of four central moments. Furthermore, there is a lot of extra freedom in the choice of q_i which may be utilised to match further higher moments. Alternatively, q_i may be generated using any random number generator or using an appropriate deterministic algorithm.

7.4 Numerical experiments

To test the computational efficiency of the optimisation procedure, we use LMI toolbox of MATLAB (version 6.5), running on a desktop with a 3 GHz Pentium processor. To derive a covariance matrix and marginal kurtosis, which is guaranteed to correspond to a feasible distribution, we use MATLAB's random number generator for t-distribution with 10 degrees of freedom. The sample covariance matrix of the resulting random samples was used as a target covariance matrix and the sample marginal kurtosis values were used as the target kurtosis in our optimisation. We ran the numerical experiments for various combinations of number of variables (m), scenarios ($2ms + 1$) and kurtosis values. Some of the results are reported in Table 7.1 with $\hat{\epsilon}$ defined in Theorem 7.4. We report only the mean target kurtosis, rather than the individual kurtosis values, for brevity. Note that the mean vector, the covariance matrix and the zero skewness are exactly matched in all cases. The specific choice of these first three moments has very little impact on the matching of the fourth marginal moment. It can be seen that it took less than 15 seconds to generate 7201 samples for 60 random variables. The worst kurtosis matching error over the scenarios and dimensions under consideration was around 15%. The average error between the target and the achieved kurtosis values over m

dimensions was significantly smaller and was under 5% in all cases. The computation times can easily be improved by employing a higher specification machine and a purpose-written optimisation code, i.e., one that exploits the sparsity in (7.5).

m	s	$\frac{1}{m}\sum_{i=1}^{m}\kappa_i$	$\hat{\epsilon}$	time in seconds
2	20	5.6118	0.3037	0.16
4	5	6.1847	0.4875	0.03
10	2	6.3868	0.6344	0.04
50	50	6.0837	0.8347	8.01
60	60	6.2017	0.7547	14.21

Table 7.1: Results of numerical experiments.

7.5 Future research

Our proposed method deals only with single stage scenarios. An extension of this algorithm to generation of scenario trees for multi-stage decision problems and an implementation of a large scale stochastic programming model demonstrating the use of this method in financial optimisation are topics of ongoing research. From a theoretical point of view, the relationship between the proposed optimisation procedure and the semi-definite optimisation procedures to determine whether a given vector of moments would be feasible (e.g., as discussed in chapter 16 of [112]) is worth investigating.

Chapter 8

A new algorithm for latent state estimation in nonlinear time series models

In this chapter, we consider the problem of optimal state estimation for a wide class of nonlinear time series models. A modified sigma point filter is proposed, which uses a new procedure for generating sigma points as detailed in chapter 7. Unlike the existing sigma point generation methodologies in engineering where negative probability weights may occur, we develop an algorithm capable of generating sample points that always form a valid probability distribution whilst still allowing the user to sample using a random number generator. The effectiveness of the new filtering procedure is in turn assessed through simulation examples.

8.1 Introduction

We consider the problem of latent state estimation in discrete, nonlinear time series. The modelling of financial and economic variables is an important consideration in the pricing, hedging and optimisation of a portfolio of financial contracts. Many of the financial models that have been put forward in the finance literature and

successfully applied in the industry can be encapsulated within the generalised specification given below. We consider a general class of systems having the state space form:

$$\mathcal{X}(k+1) = \mathbf{f}(\mathcal{X}(k)) + \mathbf{g}(\mathcal{X}(k))\,\mathcal{W}(k+1), \tag{8.1}$$

$$\mathcal{Y}(k) = \mathbf{h}(\mathcal{Y}(k)) + \mathcal{V}(k), \tag{8.2}$$

where $\mathcal{X}(k)$ is the state vector at time t_k, $\mathcal{Y}(k)$ is the measurement vector at time t_k, $\mathbf{f}, \mathbf{g}, \mathbf{h}$ are given nonlinear (vector-valued) functions and $\mathcal{V}(k), \mathcal{W}(k)$ are symmetric vector-valued random variables with bounded mean, variance and marginal kurtosis. We assume that $t_k - t_{k-1}$ is constant for all k. At each time t_k, the noisy measurement vector $\mathcal{Y}(k)$ is assumed to be available and an estimate of the random vector $\mathcal{X}(k)$ based on information up to (and including) time t_k is desired.

Examples, which are special cases of the specification in (8.1), include the constant elasticity of variance (CEV) model in stock option pricing described in Cox [31] and several exponential affine term structure models including the Cox, Ingersoll and Ross model [32] and the mean-reverting Vasicek model [110] amongst others.

The state estimation problems for these nonlinear models are practically important and occur in a wide spectrum of research areas such as radar navigation, climatology, geosciences and financial modeling, amongst others. These problems can be quite challenging numerically since the optimal recursive solution to the state estimation problem requires the propagation of full probability density; see for example, [82], for an approximate solution to a more general nonlinear filtering problem. In the special case of linear Gaussian state space models, a closed-form expression exists for the conditional state density and is given by the linear Kalman filter.

In practice, the current approaches addressing the nonlinear filtering type problems make use of one of the following ways of approximation:

- One may use the EKF, which utilises local linearisation of equation (8.1). This

leads to the derivation of a linear state space system and then a Kalman filter is employed to derive the conditional state density of $\mathcal{X}(k)$. This approach has been used in engineering for more than three decades [75] and has been extensively discussed in [4], which provides an example of the use of extended Kalman filtering. EKF works well if the system is, indeed, approximately linear. This assumption is often extremely difficult to verify. A successful implementation of EKF for a nonlinear interest rate model is given in [88].

- Another approach for nonlinear filtering is sequential Monte Carlo filtering (also called *particle filtering*), where the required density functions are represented by a set of random samples (or *particles*) with discrete probability weights and these samples are then used to compute the necessary conditional moment estimates. As the number of samples becomes large, the estimate approaches the optimal Bayesian estimate under fairly general conditions; see [81], [85], [108] and the references therein for more details on this technique. Whilst this method can perform significantly better than EKF for highly nonlinear systems, it is computationally quite expensive since a large number of samples need to be generated at each time t_k. Some computational saving is possible if the system contains a linear substructure which can be dealt with linear Kalman updates. These *marginalised* filters, which are the combination of standard particle fitler by Gordon *et al* [59] and the Kalman filter by Kalman [77], have found some applications in engineering; see [78] and the references therein.

- A modification of EKF in terms of *unscented filter* or *sigma point filter* has become popular in recent years. In [76], a survey of several applications of sigma point filters in engineering is provided, specifically in communication, tracking and navigation (also see [54]). Other reported applications of this filtering technique include the modelling of population dynamics [111] and state estimation in electrochemical cells for battery management [99]. Approximate methods to deal with multiplicative uncertainty in the observation

equation under sigma point filtering framework are discussed in [67]. This type of filters may be seen as a compromise between an EKF and a particle filter.

Similar to the propagation equations in EKF, the sigma point filters use closed-form recursive formulae based on the linear Kalman filter to propagate the mean and the covariance of state vector. However, the system equations are *not* linearised in this case. Instead, a small set of sample points (or *sigma points*) is generated and propagated through the nonlinear transformation to compute the conditional moment estimates. Instead of using a large number of points and matching the distributions asymptotically (as in a particle filter), the sigma point filter uses a small set of points which are chosen such that some of the moment properties of the *a priori* distribution are matched exactly. The main problem with this type of filters is that the sample points do not necessarily define a valid distribution since the weights corresponding to probability masses are not guaranteed to be non-negative. Furthermore, the algorithms for generating samples are purely deterministic and do not allow for a source of randomness in the filtering procedure.

A sigma point filter requires computing the square root of the state covariance matrix at each time step. This may not be computationally feasible if the number of states is very large, which is the case for most problems in geosciences. A variant of sigma point filter, usually called the *ensemble filter*, is used in geosciences where the state is not sampled at all and only the noise distributions are sampled using traditional Monte Carlo sampling techniques. This technique was introduced in [52] and has also been employed in [70]. The method we propose is closer in spirit to ensemble filters. We discuss the similarities and differences between the two filtering methods, i.e., ensemble filter vis-a-vis our proposed new method later in section 8.5.

The purpose of this chapter is to propose a new filtering algorithm for state estimation in nonlinear time series which addresses the above-mentioned deficiencies of

sigma point filters and to assess the performance of this algorithm through numerical examples. The sigma point generation step in this algorithm is adopted from a recently proposed method for generating samples from a discrete distribution with specified moment properties [37] which was also presented in chapter 7. The rest of the chapter is organised as follows. The next section outlines the recursive equations for a linear Kalman filter, which are then used in the development of subsequent sections. Section 8.3 outlines the use of the proposed sigma point filter whilst section 8.4 outlines the underlying algorithm for sigma point generation. The operation of the algorithm is demonstrated through two examples in section 8.6. Finally, section 8.7 concludes and outlines some directions for future research.

8.2 Linear Kalman filter

For a linear state space system of the form

$$\mathcal{X}(k+1) = A\mathcal{X}(k) + B + U_w\mathcal{W}(k+1), \tag{8.3}$$

$$\mathcal{Y}(k) = C\mathcal{X}(k) + D + U_v\mathcal{V}(k), \tag{8.4}$$

where A, B, C, D, U_v and U_w are constant matrices, assume that the conditional expectation $\hat{\mathcal{X}}(k \mid k)$ and its covariance matrix $P_{xx}(k \mid k)$ at time t_k (derived after measuring $\mathcal{Y}(k)$) are known. The Kalman filtering algorithm for finding conditional moments at the next time t_{k+1} proceeds as follows.

$$\hat{\mathcal{X}}(k+1 \mid k) = A\hat{\mathcal{X}}(k \mid k) + B, \tag{8.5}$$

$$P_{xx}(k+1 \mid k) = AP_{xx}(k \mid k)A^\top + U_wU_w^\top, \tag{8.6}$$

$$\hat{\mathcal{V}}(k+1) = \mathcal{Y}(k+1) - C\hat{\mathcal{X}}(k+1 \mid k) - D, \tag{8.7}$$

$$P_{xv}(k+1 \mid k) = AP_{xx}(k+1 \mid k)C^\top, \tag{8.8}$$

$$P_{vv}(k+1 \mid k) = CP_{xx}(k+1 \mid k)C^\top + U_vU_v^\top, \tag{8.9}$$

$$\hat{\mathcal{X}}(k+1 \mid k+1) = \hat{\mathcal{X}}(k+1 \mid k) + P_{xv}(k+1 \mid k) P_{vv}^{-1}(k+1 \mid k) \hat{\mathcal{V}}(k+1), \tag{8.10}$$

$$P_{xx}(k+1 \mid k+1) = P_{xx}(k+1 \mid k) - P_{xv}(k+1 \mid k) P_{vv}(k+1 \mid k)^{-1} P_{xv}(k+1 \mid k)^{\top}, \tag{8.11}$$

where $\hat{\mathcal{X}}(k+1 \mid k)$ denotes the optimal estimate of $\mathcal{X}$ at time $k+1$ given the measurements and other available values up to time k. The terms P_{xx}, P_{xv} and P_{vv} are covariance matrices under this single factor, single measurement system.

Equation (8.10) is an optimal linear filter in the sense that it yields the minimum variance over all linear filters even when $\mathcal{V}(k)$, $\mathcal{W}(k)$ are not Gaussian. When $\mathcal{V}(k)$, $\mathcal{W}(k)$ are Gaussian, $\hat{\mathcal{X}}(k+1 \mid k)$ is the conditional mean estimator for $\mathcal{X}(k+1)$, given $\mathcal{Y}(k)$. In fact, equation (8.10) may be derived using a standard conditional mean relationship for two Gaussian variables $\mathcal{X}, \mathcal{Z}$ [60]:

$$\mathbb{E}\left(\mathcal{X} \mid \mathcal{Z}\right) = \mathbb{E}\left(\mathcal{Z}\right) + \Sigma_{XZ}\Sigma_{YY}^{-1}, \big(\mathcal{Z} - \mathbb{E}(\mathcal{Z})\big), \tag{8.12}$$

where Σ_{YY} and Σ_{XZ} are covariance matrices.

The main idea of sigma point filters as well as ensemble filters is to derive approximations to the quantities on the right hand side of (8.10) through sampling the distributions of $\mathcal{V}(k)$ and $\mathcal{W}(k)$ and then use the same, closed-form update formula (8.10), which is known to be optimal for the linear Gaussian case. As mentioned earlier, the number of generated samples is kept small for computational reasons instead of matching the distributional properties asymptotically with a large number of samples.

In sigma point filters, certain moment properties of the prior distribution are matched exactly using determinstic sigma point generation. In ensemble filters on the other hand, pseudo-random number generators are used to sample the known distributions of the noise terms.

The next section details the sigma point filtering algorithm outlined above. In

the meantime, we set aside the question of which statistical properties to match and assume that a method for generating samples matching appropriate statistical properties is available. We shall consider the problem of generating samples in section 8.4.

8.3 A sigma point filter

At time t_{k+1}, assume that sample points (or *sigma points*)

$$\left[\mathcal{W}^{(i)}(k+1)^\top \quad \mathcal{V}^{(i)}(k+1)^\top\right]^\top,\ i = 1, 2, \ldots 2ms+1$$

are available for the discrete time state space system (8.1) – (8.2), along with the associated joint probability weights p_i, $i = 1, 2, \ldots s$. Here, m is the dimension of the composite vector

$$\left[\mathcal{W}^{(i)}(k+1)^\top \quad \mathcal{V}^{(i)}(k+1)^\top\right]^\top.$$

As will be seen in the next section, some of the probability weights are common to two or more support points and the set of s probability weights determine the $2ms+1$ support points above. Further, the sample points of the updated state estimate $\mathcal{X}^{(i)}(k \mid k)$ are available.

Remark 8.1

Note that $\mathcal{X}^{(i)}(k \mid k)$ is not sampled and therefore the joint probability p_i for $\left[\mathcal{W}^{(i)}(k+1)^\top \quad \mathcal{V}^{(i)}(k+1)^\top\right]^\top$ at each i is effectively assigned as the joint probability of occurrence of $\left[\mathcal{W}^{(i)}(k+1)^\top \quad \mathcal{V}^{(i)}(k+1)^\top \mathcal{X}^{(i)}(k \mid k)^\top\right]^\top$. In this respect the procedure is similar to an ensemble filter.

To initialise the procedure, we assume that $\mathcal{X}(0)$ is a random vector with a known mean, known covariance matrix and zero marginal skewness. The sample points

$\mathcal{X}^{(i)}(0 \mid 0)$ can be generated from this prior knowledge about the moments of $\mathcal{X}(0)$ using the procedure outlined in section 8.4. For $k \geq 0$, the steps involved in the computation of sigma points at time t_{k+1} once the measurement $\mathcal{Y}(k+1)$ becomes available are as follows:

$$\mathcal{X}^{(i)}(k+1 \mid k) = \mathbf{f}\left(\mathcal{X}^{(i)}(k \mid k)\right) + \mathbf{g}\left(\mathcal{X}^{(i)}(k \mid k)\right) \mathcal{W}^{(i)}(k+1), \tag{8.13}$$

$$\mathcal{Z}^{(i)}(k+1 \mid k) = \mathbf{h}\left(\mathcal{X}^{(i)}(k+1 \mid k)\right) + \mathcal{V}^{(i)}(k+1), \tag{8.14}$$

$$\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k) = \mathcal{Z}^{(i)}(k+1 \mid k) - \mathcal{Y}(k+1), \tag{8.15}$$

$$\hat{\mathcal{X}}(k+1 \mid k) = \sum_{i=1}^{2ms+1} p_i \mathcal{X}^{(i)}(k+1 \mid k), \tag{8.16}$$

$$\hat{\mathcal{V}}_{\mathcal{X}}^{(i)}(k+1 \mid k) = \mathcal{X}^{(i)}(k+1 \mid k) - \hat{\mathcal{X}}(k+1 \mid k), \tag{8.17}$$

$$P_{xx}(k+1 \mid k) = \sum_{i=1}^{2ms+1} p_i \big(\hat{\mathcal{V}}_{\mathcal{X}}^{(i)}(k+1 \mid k)\big)\big(\hat{\mathcal{V}}_{\mathcal{X}}^{(i)}(k+1 \mid k)\big)^{\top}, \tag{8.18}$$

$$P_{xv}(k+1 \mid k) = \sum_{i=1}^{2ms+1} p_i \big(\hat{\mathcal{V}}_{\mathcal{X}}^{(i)}(k+1 \mid k)\big)\big(\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k)\big)^{\top}, \tag{8.19}$$

$$P_{vv}(k+1 \mid k) = \sum_{i=1}^{2ms+1} p_i \big(\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k)\big)\big(\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k)\big)^{\top}, \tag{8.20}$$

$$\begin{aligned}\mathcal{X}^{(i)}(k+1 \mid k+1) = \hat{\mathcal{X}}(k+1 \mid k) \\ + P_{xv}(k+1 \mid k) P_{vv}^{-1}(k+1 \mid k) \hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k).\end{aligned} \tag{8.21}$$

Note the similarity between equations (8.10) and (8.21). Implementing the above algorithm yields the sigma points $\mathcal{X}^{(i)}(k+1 \mid k+1)$, $i = 1, 2, \ldots, 2ms+1$. Note that the heuristics we use to generate the samples for the measurement innovations $\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k)$ is different from the more common approach in the literature on sigma point filtering, which replaces (8.15) and (8.21) by

$$\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k) = \mathcal{Z}^{(i)}(k+1 \mid k) - \hat{\mathcal{Z}}(k+1 \mid k) \tag{8.22}$$

and

$$\mathcal{X}^{(i)}(k+1 \mid k+1) = \hat{\mathcal{X}}(k+1 \mid k) + P_{xv}(k+1 \mid k)P_{vv}^{-1}(k+1 \mid k)(\mathcal{Z}^{(i)}(k+1 \mid k) - \mathcal{Y}(k+1)) \tag{8.23}$$

respectively. In equation (8.22),

$$\hat{\mathcal{Z}}(k+1 \mid k) = \sum_{i=1}^{2ms+1} p_i \mathcal{Z}^{(i)}(k+1 \mid k).$$

However, we found that using the heuristics (8.15) and (8.21) in lieu of equations (8.22) and (8.23) improves the state estimation performance significantly. Intuitively, our choice may be justified by the fact that $\hat{\mathcal{Z}}(k+1 \mid k)$ is simply an estimate of $\mathcal{Y}(k+1)$ and it makes sense to use the actual measurement value when it is available rather than using its estimate.

To reiterate the point of this exercise, we can preserve the nonlinearity in the system dynamics whilst generating the state estimate and can do better than linear filters *without* having to resort to the computationally expensive sequential Monte Carlo-based estimation. In particular, instead of asymptotically generating entire distributions which requires a large number of samples, we use only a small number of samples but reproduce some statistical properties exactly.

In the above algorithm, we assume that a procedure to generate a set of sigma points

$$\begin{bmatrix} \mathcal{W}^{(i)}(k+1)^\top & \mathcal{V}^{(i)}(k+1)^\top \end{bmatrix}^\top, \quad i = 1, 2, \ldots 2ms+1$$

with the desired mean vector, covariance matrix and zero marginal skewness is available, and where possible, the desired sum of marginal kurtosis. The next section outlines such a procedure to generate a symmetric discrete distribution to match a given mean vector and covariance matrix exactly and also match the

sum of marginal kurtosis exactly in some cases, without requiring an additional optimisation.

8.4 Generation of sigma points

8.4.1 Notation

For completeness and to facilitate the discussion, we list the notation used in our development of the sigma point generation algorithm.

m	number of random variables (or dimension of a random vector),
s	number of samples,
Φ	target mean vector,
R	target covariance matrix,
κ_i	target marginal 4^{th} central moment for the i^{th} random variable,
L_{ij}	entry in the i^{th} row and j^{th} column of a matrix L,

We aim to generate samples from a symmetric distribution with a specified mean vector and a specified (positive definite) covariance matrix. In the case of sigma point filtering algorithm described in the last section, the purpose is to generate

$$\mathcal{G} := \begin{bmatrix} \mathcal{W}^{(i)}(k+1)^\top & \mathcal{V}^{(i)}(k+1)^\top \end{bmatrix}^\top, \; i = 1, 2, \ldots 2ms+1,$$

which match a given mean vector Φ, a given covariance matrix R and have symmetric marginal distribution.

8.4.2 Algorithm for generating sigma points

The algorithm described below is adopted from the scenario generation algorithm from chapter 7.

(i) Find a symmetric positive definite matrix L such that $R = LL^\top$. For a symmetric positive definite R, L is unique. If R has distinct eigenvalues, this may be found using singular value decomposition.

(ii) If $\frac{\sum_{i=1}^{m} \kappa_i}{\sum_{i,j=1}^{m}(L_{ij}^4)} > m$, generate $s-1$ numbers $q_i \in [2ms, \psi], i = 1, 2, \ldots, s-1$, where the constant ψ is given by

$$\psi = \frac{2s^2 \sum_{i=1}^{m} \kappa_i}{(s-1)\sum_{i,j=1}^{m}(L_{ij}^4)} - \frac{2ms}{s-1},$$

and set $q_s = \frac{2s^2 \sum_{i=1}^{m} \kappa_i}{\sum_{i,j=1}^{m}(L_{ij}^4)} - \sum_{i=1}^{s-1} q_i$. It can be easily shown that if $\frac{\sum_{i=1}^{m} \kappa_i}{\sum_{i,j=1}^{m}(L_{ij}^4)} > m$, it is always possible to choose s such that $\psi > 2ms$. Further, due to the definition of the upper bound on q_i, it can be shown that $q_s > 2ms$ holds. For a scalar random variable, the constraint $\frac{\kappa_1}{(L_{11}^4)} > 1$ implies that the kurtosis is greater than unity, which is always true for elliptic distributions [12].

If the condition above is not satisfied, i.e., if $\psi \leq 2ms$, generate s numbers $q_i \in [2ms, \infty], i = 1, 2, \ldots, s-1$. Given its lower bound and (possibly) upper bound in either case, q_i may be generated using any deterministic algorithm or using a random number generator.

(iii) Set $p_i = \frac{1}{q_i}$, $i = 1, 2, \ldots, s$ and $p_{s+1} = 1 - 2m\sum_{i=1}^{s} p_i$.

(iv) Define a multivariate discrete distribution $\mathcal{G}$ over a support of $2ms+1$ points as follows:

$$\begin{aligned} &\mathbb{P}\left(\mathcal{G} = \Phi + \frac{1}{\sqrt{2sp_i}} L_j\right) = \mathbb{P}\left(\mathcal{G} = \Phi - \frac{1}{\sqrt{2sp_i}} L_j\right) = p_i, \\ &j = 1, 2, \ldots, m,\ i = 1, 2, \ldots, s, \\ &\mathbb{P}\left(\mathcal{G} = \Phi\right) = p_{s+1}, \end{aligned} \tag{8.24}$$

where L_j denotes the j^{th} column of a matrix L.

Steps (i)-(iv) constitute the entire set of procedures needed to construct the required samples. Step (i) need not be repeated in a sequential procedure, if the covariance matrix is to remain the same through multiple time steps. This is practically important since noise covariance matrices are usually assumed to be constant and they need not be factorised at each time step during the filtering.

The following result collects together the distributional properties of these samples.

Theorem 8.2

(i). For p_i defined as above, $p_i \geq 0$, $i = 1, 2, \ldots, s$ and $2m \sum_{i=1}^{s} p_i + p_{s+1} = 1$.

(ii). For $\mathcal{G}$ defined as above,

$$\mathbb{E}[\mathcal{G}] = \Phi, \tag{8.25}$$

$$\mathbb{E}\left[(\mathcal{G} - \Phi)(\mathcal{G} - \Phi)^{\top}\right] = R, \tag{8.26}$$

$$\mathbb{E}\left[(\mathcal{G}_i - \Phi_i)^3\right] = 0. \tag{8.27}$$

Furthermore, if $\psi > 2ms$ holds,

$$\sum_{i=1}^{m} |\kappa_i - \mathbb{E}\left(\mathcal{G}_i - \Phi_i\right)^4| = \sum_{i=1}^{m} \kappa_i. \tag{8.28}$$

Proof

Since $q_i \geq 2ms$, $i = 1, 2 \ldots, s$, and $p_{s+1} = 1 - 2n \sum_{i=1}^{s} p_i$ by definition, we need to show that $p_{s+1} \geq 0$ for part (i) to hold. This follows by noting that

$$p_{s+1} \geq 1 - 2ms \frac{1}{\min_i q_i} \geq 0.$$

As far as part (ii) of Theorem 8.2 is concerned, equations (8.25) and (8.27) are immediate due to the symmetry of the support points around the target mean

vector Φ. Equation (8.26) follows by noting that

$$\mathbb{E}\left[(\mathcal{G}-\Phi)(\mathcal{G}-\Phi)^\top\right] = 2\sum_{i=1}^{s} p_i \frac{1}{2sp_i}\left(\sum_{j=1}^{m} L_j L_j^\top\right) = \sum_{j=1}^{m} L_j L_j^\top = R.$$

Finally,

$$\mathbb{E}\left[(\mathcal{G}_i-\Phi_i)^4\right] = \sum_{k=1}^{s} p_k \frac{1}{2p_k^2 s^2}\sum_{j=1}^{m} L_{ij}^4 = \left(\sum_{k=1}^{s} q_k\right)\frac{1}{2s^2}\left(\sum_{j=1}^{m} L_{ij}^4\right),$$

so that, when $\psi > 2ms$ holds,

$$\sum_{i=1}^{m}\mathbb{E}\left[(\mathcal{G}_i-\Phi_i)^4\right] = \sum_{i=1}^{m}\kappa_i,$$

where the last equality follows from the definition of q_s in step (ii) of the algorithm.

□

Theorem 8.2 and its proof above demonstrate one of the main advantages of our method: provided that the weights p_i form a valid probability measure, their exact values have no impact on the exact matching of Φ and R. In particular, p_i's get cancelled in forming the covariance matrix from the support points and the associated probability weights of $\mathcal{G}$. If $\mathcal{G}(k)$ itself represents a discrete time stochastic process, this crucial fact allows us to choose *random* probability weights $\{p_i\}$ within the specified bounds, $\left(\frac{1}{2ms}, \frac{1}{\psi}\right)$ or $\left(\frac{1}{2ms}, \right)$, at each time k, thereby generating a different realization of $\mathcal{G}(k)$ at each time k. Of course, we may choose to use deterministic p_i's instead if desired.

8.5 What is new in our approach?

The filtering algorithm described in section 8.3 is similar to various sigma point and ensemble filtering algorithms described elsewhere. However, the sampling approach involved in our method differs radically from those of others as described in section 8.4. The main differences between sample generation methods in traditional sigma

point filters and the filter using the proposed sampling method, which we shall henceforth refer to as modified sigma point filter (MSPF) are as follows.

- In the existing methods, the sigma points or samples generated do not *necessarily* form a valid probability distribution, as some of the probability weights can be negative. This puts the probabilistic interpretation of the whole procedure into question. This problem does not arise in MSPF as can be seen from part (ii) of the Theorem 8.2.

- The existing sigma point generation algorithms are deterministic, with no way of incorporating random behaviour. The randomness may be desirable to reflect the real dynamics of the system, especially when the system is assumed to have explicit sources of randomness. This is especially true when modelling econometric or financial time series. As mentioned earlier, we have the flexibility of using either a deterministic or probabilistic sigma point generation in our algorithm since q_i may be generated in either way.

- We can also match the sum of fourth marginal central moments in cases when the condition $\psi > 2ms$ holds, as detailed in section 8.2. This criterion will often hold for a system with a single measurement and a single state. In case it does not hold, it is possible to minimise the worst case error in matching the marginal fourth central moments of the components of $\mathcal{G}$ using a convex optimisation procedure. The details of this procedure were described in chapter 7 and hence are omitted here.

- Similar to ensemble filters, we do not sample the state $\mathcal{X}(k \mid k)$. This makes intuitive sense since we are sampling all the exogenous sources of randomness and $\mathcal{X}(k \mid k)$ is simply a function of these exogenous random processes. This has also a very significant computational advantage over traditional sigma point filters in terms of not having to compute a new matrix square root of a covariance matrix at each time step (which would be the case if we were sampling $\mathcal{X}(k \mid k)$). The ensemble filters, however, use random number

generators with the specified distributional properties to generate samples for

$$\begin{bmatrix} \mathcal{W}^{(i)}(k+1)^\top & \mathcal{V}^{(i)}(k+1)^\top \end{bmatrix}^\top$$

and then use the corresponding sample mean and the sample covariances in (8.21). However, sampling from an underlying distribution with a very small number of samples and then using the sample moments may yield misleading results and it may be better to match a few moments exactly instead, which is what the MSPF is designed to do.

MSPF integrates the numerical simplicity of the ensemble filter with the exact moment matching of the traditional sigma point filter. Additionally, it extends the moment matching method in traditional sigma point filters beyond matching the mean vector and the covariance matrix. In MSPF, the third marginal moment is matched exactly in all cases and the sum of the fourth marginal moments can also be matched exactly in some cases. Achieving even approximate matching of these higher moments in the case of existing sigma point generation algorithms requires non-convex optimisation.

8.6 Numerical examples

We consider two different examples to illustrate the proposed filtering method.

8.6.1 CEV-type time series model

The first numerical example is a nonlinear, non-Gaussian time series given by

$$\begin{aligned} x(k+1) &= ax(k) + b + \sigma_w \left\{ (x(k))^2 \right\}^{\frac{\gamma}{2}} w(k+1), \\ y(k+1) &= cx(k) + d + \sigma_v v(k+1). \end{aligned} \tag{8.29}$$

Examples of the specification in (8.29) include the constant elasticity of variance (CEV) model in stock option pricing described in Cox [31] and several exponential affine term structure models including the Cox, Ingersoll and Ross model [32]. We consider a univariate example for simplicity. The parameters of the model are $a = 0.9$, $b = 0.1$, $\sigma_w = 0.01$, $c = 1$, $d = 0.1$, $\sigma_v = 0.01$. The noise terms $w(k+1)$ and $v(k+1)$ are IID with standard normal distribution. We consider three different values for γ: $\gamma = 0.125, 0.25, 0.375$. Sample paths of the state x and observation y are generated by sampling $w(k)$ and $v(k)$. Based on the observation sample path, we wish to see whether we can predict $x(k+1|k)$ accurately using the MSPF method proposed here and we also would like to compare its predictive ability with that of the EKF and the ensemble filter. As a measure of performance of a filter, we consider the average of root mean squared error (AvRMSE) in one-step ahead predictions. The root mean squared error (RMSE) for a filter and for a particular sample path i is given by

$$RMSE^{(i)} = \sqrt{\frac{1}{T}\sum_{k=1}^{T}\left(x^{(i)}(k+1) - \hat{x}^{(i)}(k+1|k)\right)^2},$$

where the superscript i denotes the i^{th} sample path and T is the time horizon. AvRMSE, as the sample mean $\frac{1}{S}\sum_{i=1}^{S} RMSE^{(i)}$, and VarRMSE as the corresponding sample variance,

$$\text{VarRMSE} = \frac{1}{S-1}\sum_{i=1}^{S}\left(RMSE^{(i)} - AvRMSE\right)^2,$$

are computed over $S = 100$ sample paths with each path consisting of $T = 100$ time steps. At each time step, only 10 samples or sigma points are generated. Table 8.1 illustrates the results of this error analysis. We see that the MSPF yields the lowest AvRMSE and the lowest VarRMSE amongst the three filtering methods, for all values of γ. Moreover, the VarRMSE is the lowest for MSPF. Whilst the difference in the performance of ensemble filter and MSPF as measured by AvRMSE is small, MSPF has far more predictable MSE, as indicated by significantly lower

VarRMSE.

Filtering method		γ		
		0.125	0.25	0.375
EKF	AvRMSE	2.9563×10^{-2}	2.9584×10^{-2}	2.9606×10^{-2}
	VarRMSE	8.8845×10^{-5}	8.9478×10^{-5}	9.0154×10^{-5}
Ensemble filter	AvRMSE	2.6251×10^{-2}	2.6258×10^{-2}	2.6267×10^{-2}
	VarRMSE	3.9879×10^{-5}	3.9882×10^{-5}	3.9916×10^{-5}
MSPF	AvRMSE	2.4960×10^{-2}	2.4967×10^{-2}	2.4974×10^{-2}
	VarRMSE	2.7946×10^{-5}	2.7888×10^{-5}	2.7859×10^{-5}

Table 8.1: Comparison of prediction errors using different filters for system in (8.29) for a specific filter and a value of γ.

8.6.2 Univariate non-stationary growth model

The second numerical example is a univariate non-stationary growth model given by

$$x(k) = \alpha x(k-1) + \beta \frac{x(k-1)}{1+x^2(k-1)} + \gamma \cos(1.2(k-1)) + \sigma_w w(k), \tag{8.30}$$

$$y(k) = \frac{x^2(k)}{20} + \sigma_v v(k), \tag{8.31}$$

where $v(k)$ are IID $N(0,1)$ random variables and $w(k)$ are IID random variables with a zero mean and unit variance following a t-distribution having 10 degrees of freedom. A nonlinear model of this type was discussed in [27] and [80], where amongst other things, it was shown to exhibit a long memory property.

We use the parameters $\alpha = 0.5, \beta = 28, \gamma = 8, \sigma_w = 0.1, \sigma_v = 0.1$. In this case, we compare the performance of the ensemble filter and MSPF for one-step ahead predictions, using AvRMSE and VarRMSE as defined in the previous example. As in the previous case, AvRMSE is computed over 100 sample paths; each path consisting of 100 time steps, and 10 sigma points are generated at each time step. The AvRMSE for MSPF is 0.61172 whilst that for the ensemble filter is 0.61558.

The VarRMSE for MSPF is 0.00028 whilst that for the ensemble filter is 0.00107. As in the previous example, it is seen that the proposed method outperforms the ensemble filter, with a lower AvRMSE and a lower VarRMSE.

Whilst VarRMSE reflects variation of RMSE across different sample paths, it is also worthwhile to comment on the variation in one-step ahead prediction errors across different time steps along the same sample path. Even though both the filtering methods, viz. ensemble filter and MSPF use a small number of randomly generated samples for computing moments, the samples in MSPF are likely to yield locally closer predictions due to moment matching constraints. To illustrate this point, a plot of a simulated sample path (denoted by a blue solid line) and the plot corresponding to the one step-ahead predictions, using MSPF (denoted by green dashed line) are depicted in Figure 8.1, whilst the same sample path and one-step ahead predictions using ensemble filter are plotted in Figure 8.2. Indeed, this intuition is confirmed by comparing Figures 8.1 and 8.2 and observing the contrast between the behaviours of the two predictions occurring most notably within the first 20 time steps.

We do not compare the methods with a linearised filter in this example since the system is too nonlinear for a local linearisation procedure to be effective.

8.7 Concluding remarks

We have developed a modified sigma point filtering algorithm for nonlinear and non-Gaussian systems. This algorithm combines the numerical simplicity of the ensemble filter (in the sense that the state covariance matrix need not be factorised) along with the exact moment matching properties of the traditional sigma point filter. Furthermore, whilst the traditional sigma point filtering methods match only the first two moments, the exact moment matching is extended to three and in some cases four moments in the MSPF. The use of the algorithm is demonstrated through numerical examples.

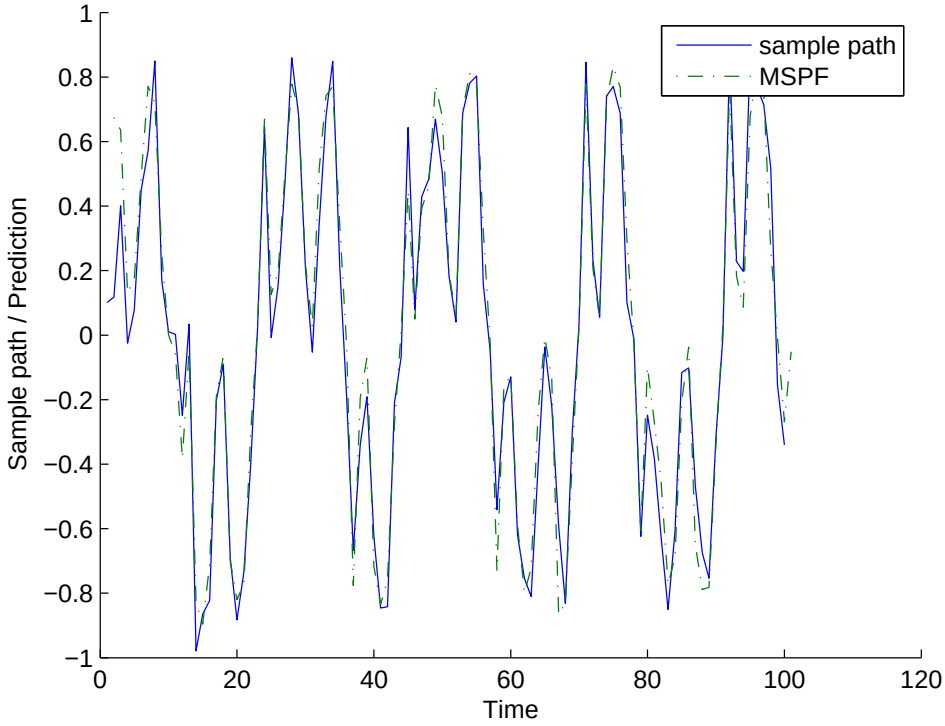

Figure 8.1: Plot of simulated sample paths and one step-ahead prediction for univariate non-stationary growth model using MSPF.

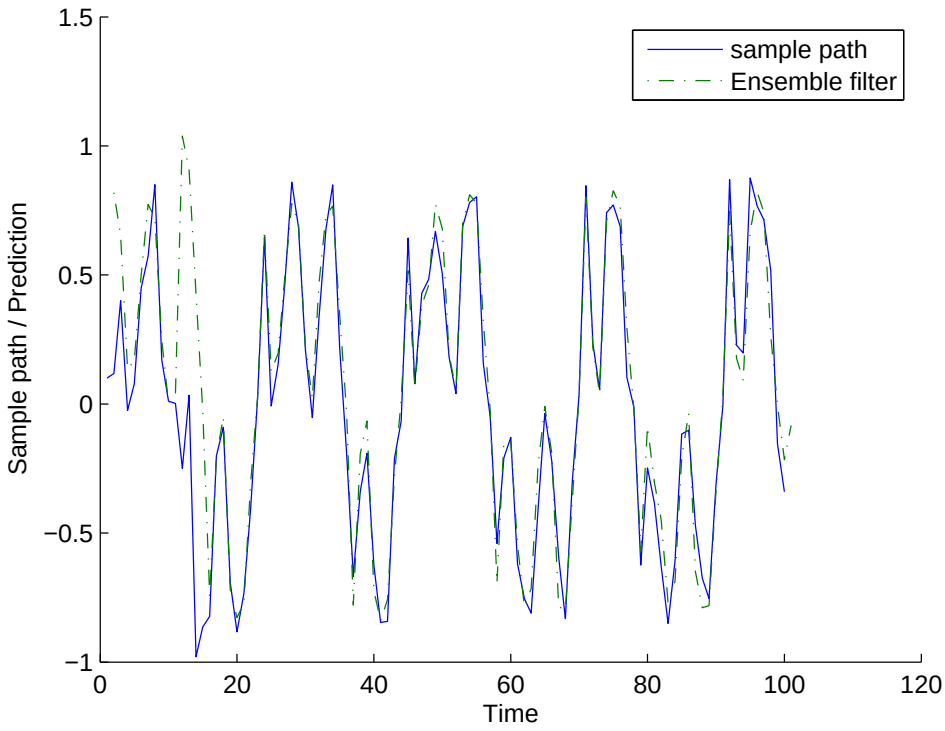

Figure 8.2: Plot of simulated sample paths and one step-ahead prediction for univariate non-stationary growth model using Ensemble filter.

To the best of our knowledge, sigma point filters have not been adopted widely by researchers in fields such as econometrics, finance and actuarial science. We feel that with the methodology we develop in this paper that addresses some of the major shortcomings of sigma point filters, the method of sigma point filtering will be more attractive and useful to researchers and practitioners alike in these fields of research. The proposed method further provides a very useful alternative to traditional sigma point filters in engineering and ensemble filters in geosciences. In both of these fields of research, speed of computation for nonlinear filtering is crucial. In engineering, the time steps can be too small to allow intensive computation in filtering, whilst in geosciences, the typical number of states is too large to carry out a Monte-Carlo-based filtering. The proposed algorithm fills in the gap left by sigma point and Monte-Carlo-based filtering.

Chapter 9

A partially linearised sigma point filter for latent state estimation in nonlinear time series models

We propose an alternative technique for the optimal state estimation of a wide class of nonlinear time series models. In particular, we develop a partially linearised sigma point filter in which sigma points for generating samples of possible state values are employed at the prediction step and then a linear programming-based procedure is used in the update step of the state estimation. The effectiveness of the new filtering procedure is assessed via a simulation example that deals with a highly nonlinear, multivariate interest rate process.

9.1 Introduction

This chapter is concerned with the problem of latent state estimation for a nonlinear time series in discrete time. As in chapter 8, our analysis will focus on the general

class of systems with the state space form:

$$\mathcal{X}(k+1) = \mathbf{f}(\mathcal{X}(k)) + \mathbf{g}(\mathcal{X}(k))\,\mathcal{W}(k+1), \tag{9.1}$$

$$\mathcal{Y}(k) = \mathbf{h}(\mathcal{Y}(k)) + \mathcal{V}(k), \tag{9.2}$$

where $\mathcal{X}(k)$ and $\mathcal{Y}(k)$ are the respective state vector and measurement vector at time t_k; $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{h}$ are given nonlinear (vector-valued) functions, and both $\mathcal{V}(k)$ and $\mathcal{W}(k)$ are symmetric vector-valued random variables. The time increment $t_k - t_{k-1}$ is assumed constant for all k. Moreover, we assume the noisy measurement vector $\mathcal{Y}(k)$ is available for every t_k. We wish to find an estimate of the random vector $\mathcal{X}(k)$ based on information up to (and including) time t_k.

Current approaches in practice designed to address the nonlinear filtering type problems usually fall under one of the following approximation methods:

- Extended Kalman filter (EKF),
- Sequential Monte Carlo filtering,
- Unscented filter.

More details and a brief description of the above listed approximation methods were provided in chapter 8, specifically in section 8.1.

The sigma point filters use closed-form recursive formulae based on the linear Kalman filter to propagate the mean and the covariance of state vector; this is essentially similar to the propagation equations in EKF. The system equations, nonetheless, are not linearised in this case. A small set of sample points (or sigma points) is generated and propagated through the nonlinear transformation to compute the conditional moment estimates. In lieu of using a large number of points and matching the distributions asymptotically (as in a particle filter), the sigma point filter uses a small set of points which are chosen such that some of the moment properties of the a priori distribution are matched exactly. Whilst these filters

have prevailed in engineering, they suffer from several shortcomings as elaborated below:

1. The weights corresponding to probability masses are not guaranteed to be non-negative. Thus, the sample points generated in these filters do not necessarily define a valid distribution.

2. There is no source of randomness in the filtering procedure because the algorithms for generating samples are purely deterministic.

3. The computation of a square root of the state covariance matrix at each time-step is required in a sigma point filter. If the number of states is very large, this would present a hurdle in its computational feasibility. This is usually the case for most problems in geosciences. The ensemble filter, a variant of sigma point filter, is commonly used in geosciences under which the state is often not sampled but only the noise distributions are sampled via the traditional Monte Carlo sampling techniques. This technique was developed in [52] and could also be found in [70]. A review of ensemble filtering techniques appears in [53]. However, as indicated in chapter 8, using a small number of samples (typically around 10) as a discrete proxy for a continuous distribution may lead to misleading results.

4. Lastly, Kalman filter has a very clear interpretation only in the linear case. For linear Gaussian systems, Kalman filter is a conditional mean estimator. It is worth recalling that even when $\mathcal{V}(k)$ and $\mathcal{W}(k)$ are not Gaussian, Kalman filter is an optimal linear filter, in the sense that it yields the minimum variance over all linear filters. However, neither of these properties are relevant if the system is nonlinear. As well, the motivation of using Kalman filtering state estimator equations based on (8.12) is not always clear.

The first three shortcomings listed above were addressed in chapter 8, where a new sigma point generation procedure was employed to match the first three moments exactly (as in the case of sigma point filters) at the same time using randomly

generated samples (as in the case of ensemble filters). However, in the previous chapter we were still using a formula based on (8.12) for the update step in filtering. The purpose of this chapter is to present an alternative heuristic for the state estimation of a nonlinear time series which does not use (8.12) in state estimation at all and and seeks a state estimate which best matches the observations in an appropriate deterministic sense. The algorithm uses linearised measurement equation but preserves the nonlinearity of the state evolution equation. Hence, we shall refer to this new filter as partially linearised sigma point filter (PLSPF).

In PLSPF, we generate samples of exogenous noise in the state evolution equation (9.1) using the exact moment-matching procedure from chapter 7. These noise samples are used to obtain samples of state prediction. The measurement equation is linearised (similar to EKF) and a set of linear programming problems is solved to obtain samples of the updated state which best match the observations.

This chapter is organised as follows. Section 9.2 sets out the algorithm in implementing the partially linearised sigma point filter whilst section 9.3 outlines briefly the underlying algorithm for sigma point generation. We include a demonstration of the algorithm's operation through a numerical example in section 9.4. More specifically, we illustrate the filtering procedure with a multivariate, nonlinear time series. Finally, some concluding remarks and certain directions for future research are given in section 9.5.

9.2 A partially linearised sigma point filter

Suppose at time t_{k+1}, the sample points (or sigma points)

$$\mathcal{W}^{(i)}(k+1), \quad i \in \{1, 2, \ldots, 2ms+1\}$$

are available for the discrete time state space system (9.1) – (9.2) together with their associated probability weights p_i, $i \in \{1, 2, \ldots, 2ms+1\}$. Here, m is the dimension of the vector $\mathcal{W}(k)$ (or in other words, the dimension of the state space

in (9.1)). We assume that the collection of samples $\mathcal{W}(i)(k+1)$ matches a given mean vector, covariance matrix and zero marginal skewness. The discussion of how to generate $\mathcal{W}(i)(k+1)$ is postponed until the next section.

A set of s probability weights determines the $2ms+1$ support points above; details are given in the next section. In addition, the sample points of the updated state estimate $\mathcal{X}(i)(k \mid k)$ are assumed to be available at time t_{k+1}.

Remark 9.1

Note that similar to the procedure described in the previous chapter $\mathcal{X}^{(i)}(k \mid k)$ is not sampled. The joint probability for $\left[\mathcal{W}^{(i)}(k+1)^\top \;\; \mathcal{V}^{(i)}(k+1)^\top\right]^\top$ is at each i effectively assigned as the joint probability of occurence of $\left[\mathcal{W}^{(i)}(k+1)^\top \;\; \mathcal{V}^{(i)}(k+1)^\top \;\; \mathcal{X}^{(i)}(k \mid k)\right]^\top$. In this respect our algorithm is similar to an ensemble filter.

We assume that $\mathcal{X}(0)$ is a random vector with a known mean, known covariance matrix and zero marginal skewness in the initialisation stage of the procedure. Section 9.3 describes a procedure that can be employed to generate the sample points $\mathcal{X}^{(i)}(0 \mid 0)$ from a prior knowledge about the moments of $\mathcal{X}(0)$. For $k \geq 0$ and whenever the measurement $\mathcal{Y}(k+1)$ becomes available, we present the steps in the computation of sigma points at time t_{k+1}:

$$\mathcal{X}^{(i)}(k+1 \mid k) = \mathbf{f}\left(\mathcal{X}^{(i)}(k \mid k)\right) + \mathbf{g}\left(\mathcal{X}^{(i)}(k \mid k)\right) \mathcal{W}^{(i)}(k+1) \tag{9.3}$$

$$\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k) = \mathcal{Y}(k+1) - \mathbf{h}\left(\mathcal{X}^{(i)}(k+1 \mid k)\right) \tag{9.4}$$

$$\hat{\mathcal{X}}(k+1 \mid k) = \sum_{i=1}^{2ms+1} p_i \mathcal{X}^{(i)}(k+1 \mid k) \tag{9.5}$$

$$\hat{\delta}^{(i)}(k+1 \mid k+1) = \arg \min_{\delta^{(i)}(k+1|k+1)} ||\hat{\mathcal{V}}_{\mathcal{Y}}^{(i)}(k+1 \mid k) - H^{(i)}(k+1 \mid k)\delta^{(i)}(k+1 \mid k+1)||_1 \tag{9.6}$$

$$\mathcal{X}^{(i)}(k+1 \mid k+1) = \mathcal{X}^{(i)}(k+1 \mid k) + \hat{\delta}^{(i)}(k+1 \mid k+1) \tag{9.7}$$

$$\hat{\mathcal{X}}(k+1 \mid k+1) = \sum_{i=1}^{2ms+1} p_i \mathcal{X}^{(i)}(k+1 \mid k+1), \tag{9.8}$$

where the gradient matrix $H^{(i)}(k+1 \mid k)$ for the vector valued function $\mathbf{h}$ at time

t_{k+1} is defined by

$$\left[H^{(i)}(k+1 \mid k)\right]_{jl} = \left.\frac{\partial \mathbf{h}_j}{\mathcal{X}_l}\right|_{\mathcal{X}^{(i)}(k+1|k)}$$

and $||\cdot||_1$ denotes the 1-norm of a vector (which equals the summation of absolute values of all elements).

Implementing the above algorithm yields the sigma points $\mathcal{X}^{(i)}(k+1 \mid k+1)$, $i \in \{1, 2, \ldots, 2ms+1\}$, along with the expected values of the predicted and the updated state estimate, i.e., $\hat{\mathcal{X}}(k+1 \mid k)$ and $\hat{\mathcal{X}}(k+1 \mid k+1)$, respectively. Note that the 1-norm minimisation in (9.6) can be achieved by linear programming (LP).

If $\hat{\epsilon}^{(i)}$ is the minimum cost and if $\hat{\delta}^{(i)}(k+1 \mid k+1)$ are the decision variables which achieve this minimum, it is easy to see that there exists $\mathcal{V}^{(i)}(k+1)$ such that

$$\mathcal{Y}(k+1) = \mathbf{h}\big(\mathcal{X}^{(i)}(k+1 \mid k)\big) + H^{(i)}(k+1 \mid k)\delta^{(i)}(k+1 \mid k+1) + \mathcal{V}^{(i)}(k+1)$$

holds and

$$||\mathcal{V}^{(i)}(k+1)||_1 \leq \hat{\epsilon}^{(i)}.$$

In other words, corresponding to each $\mathcal{X}^{(i)}(k+1 \mid k)$, the procedure finds (vector-valued) measurement noise which causes the smallest error as measured by the 1-norm between the linearised prediction of $\mathbf{h}(\cdot)$ around $\mathcal{X}^{(i)}(k+1 \mid k)$ and the actual observation $\mathcal{Y}(k+1)$.

We re-emphasise that the main idea of this exercise is to preserve some of the nonlinearity in the system dynamics whilst generating the state estimate and can (possibly) do better than the EKF without having to resort to the computationally expensive sequential Monte Carlo-based estimation. Note that solving a small number of LP-based optimisation problems with m decision variables will usually be cheaper than doing a Monte Carlo simulation with m correlated random variables. LP problems can be solved extremely effciently (theoretically, in polynomial time)

and several good LP solvers are commercially available; see [105] for more details on linear programming algorithms. Solving an LP problem with several hundred variables and constraints in a few seconds on an ordinary desktop is a reasonable expectation given today's technological advancement in computing. Furthermore, this procedure eliminates the need of knowing the information about the parametric form of distribution of the measurement noise, which is not always available.

Finally, the special case when **h** is linear is worth mentioning. When **h** is linear, H is a constant matrix and the measurement equation can be written as

$$\mathcal{Y}(k+1) = H\mathcal{X}(k+1) + \mathcal{V}(k+1).$$

In this case, the 1-norm minimisation problem

$$\min_{\mathcal{X}(k+1|k+1)} ||\mathcal{Y}(k+1) - H\mathcal{X}(k+1 \mid k+1)||_1$$

has a unique solution. One simply needs to solve this single linear programming problem and need not use the samples $\mathcal{X}^{(i)}(k+1 \mid k)$ in computing $\mathcal{X}(k+1 \mid k+1)$. Moreover,

$$\mathcal{X}^{(i)}(k+1 \mid k+1) = \mathcal{X}^{(j)}(k+1 \mid k+1) =: \hat{\mathcal{X}}(k+1 \mid k+1)$$

holds. The sampling procedure is still required for $\mathcal{W}^{(i)}(k+1)$ if the expected value of prediction, $\hat{\mathcal{X}}(k+1 \mid k)$, in (9.5) needs to be predicted. Time series models with nonlinear **f** and **g** in (9.1), but a linear **h** in (9.2) commonly occur in financial economics and econometrics. The most prominent class of models with this structure includes the Cox-Ingersoll-Ross (CIR) model, which is employed to model interest rates. This popular class of models has been widely discussed in the literature; see [32] and [57], amongst others. The instantaneously compounded interest rate in these types of models is unobservable and has to be inferred from observed interest rates using a nonlinear filter; see [57] for the use of EKF in CIR-type interest rate models. Clearly, the algorithm proposed here can provide an

intuitively attractive and computationally affordable alternative to EKF, which does not rely on linearisation of the state evolution equation.

In the above algorithm, we assume that a procedure to generate a set of sigma points $\mathcal{W}^{(i)}(k+1)$ for $i \in \{1, 2, \ldots, 2ms+1\}$ with the desired statistical properties is available. The next section outlines such a procedure in generating a symmetric discrete distribution that matches a given mean vector and covariance matrix exactly without requiring an additional optimisation. This procedure was first suggested in chapter 7 and was used in nonlinear filtering context in chapter 8. A very brief summary of this procedure is provided here for a self-contained presentation of our proposed method in latent state estimation.

9.3 Generation of sigma points

Our aim is to generate samples from a symmetric distribution with a specified mean vector and a specified (positive definite) covariance matrix. The sigma point generation algorithm given in the next subsection forms a part of the filtering procedure described in section 9.2, as it is used to generate $\mathcal{G} := \mathcal{W}^{(i)}(k+1)$, for $i \in \{1, 2, \ldots, 2ms+1\}$ which match a given mean vector Φ, a given covariance matrix R and have a symmetric marginal distribution.

As mentioned before, a brief outline of sigma point generation will be provided here for a complete presentation. A more thorough elaboration of the algorithm together with the notation is given in section 8.4. All notation used here correspond to the ones already defined in the previous chapter.

Below is a short and general description of the algorithm for sigma point generation.

1. Decompose a matrix R as $R = LL^\top$ where L is a symmetric positive definite matrix. For a symmetric positive definite R, L is unique.

2. Generate an s number of $q_i \in [2ms, 1]$ for $i \in \{1, 2, \ldots, s-1\}$. The q_i's may be generated using any deterministic algorithm or using a random number generator.

3. Write $p_i := \frac{1}{q_i}$ for $i \in \{1, 2, \ldots, s\}$ and $p_{s+1} := 1 - 2m \sum_{i=1}^{s} p_i$.

4. Define a multivariate discrete distribution $\mathcal{G}$ over a support of $2ms+1$ points as follows:

$$\begin{aligned} &\mathbb{P}\left(\mathcal{G} = \Phi + \frac{1}{\sqrt{2sp_i}} L_j\right) = \mathbb{P}\left(\mathcal{G} = \Phi - \frac{1}{\sqrt{2sp_i}} L_j\right) = p_i, \\ &j = 1, 2, \ldots, m,\ i = 1, 2, \ldots, s, \\ &\mathbb{P}\left(\mathcal{G} = \Phi\right) = p_{s+1}. \end{aligned} \tag{9.9}$$

   where L_j denotes the j^{th} column of a matrix L.

Steps 1 – 4 comprise of the procedure to generate sigma points having Φ as the mean vector, R as the covariance matrix and zero marginal skewness. In sequential state estimation, step 1 is not necessary to be repeated when the covariance matrix has to remain the same throughout multiple time steps. In various important practical applications, the noise covariance matrices are usually assumed to be constant. Hence, they need not be factorised at each time step in the filtering process.

The distributional properties of these samples are summarised in Theorem 8.2. In particular, Theorem 8.2 demonstrates one of the main advantages of our method. In particular, provided that the weights p_i form a valid probability measure, their exact values have no impact on the exact matching of Φ and R. Note that p_i's get cancelled in forming the covariance matrix from the support points and the associated probability weights of $\mathcal{G}$. A notable fact to mention is that in situations where $\mathcal{G}(k)$ itself represents a discrete time stochastic process, we could choose random probability weights $\{p_i\}$ within the specified bounds $(0, \frac{1}{2ns})$ at each time k, thereby generating a different realization of $\mathcal{G}(k)$ at each time k. Of course, we may choose to use deterministic p_i's instead if desired.

9.4 Numerical example

We consider an Euler-discretised version of a 2-factor, square root affine interest rate model as described in [57] to demonstrate the implementation of the new filtering method. This model is a generalisation of the CIR model first proposed in [32]. In this model, the two unobservable states $\mathcal{X}_1(k)$ and $\mathcal{X}_2(k)$ are assumed to evolve according to the equation

$$\begin{aligned}\mathcal{X}_j(k+1) &= \kappa_j\theta_j\Delta + (1-\kappa_j\Delta)\mathcal{X}_j(k) \\ &\quad + \sigma_j\sqrt{\mathcal{X}_j(k)\Delta}\mathcal{W}_j(k+1), \quad \text{for} \quad j \in \{1,2\},\end{aligned} \tag{9.10}$$

where $\mathcal{W}_1(k)$ and $\mathcal{W}_2(k)$ are independent standard normal random variables at each time t_k and $\Delta := t_k - t_{k-1}$. The time period between two successive samples $\Delta = 1/250$ is assumed to be constant. The measurement functions of these states, $\mathcal{Y}_i(k)$, are given by

$$\mathcal{Y}_i(k) = \prod_{j=1}^{2} A_{i,j} \exp\left(-\sum_{j=1}^{2} B_{i,j}\mathcal{X}_j(k)\right) + \mathcal{V}_i(k), \tag{9.11}$$

where

$$A_{i,j} = \left(\frac{2\phi_{j,1}\exp\left(\phi_{j,2}\frac{T_i}{2}\right)}{\phi_{j,4}}\right)^{\phi_{j,3}} \quad \text{and} \quad B_{i,j} = \frac{2\big(\exp(\phi_{j,1}T_i)-1\big)}{\phi_{j,4}}.$$

In addition,

$$\begin{aligned}\phi_{j,1} &= \sqrt{(\kappa_j+\lambda_j)^2 + 2\sigma_j^2} \\ \phi_{j,2} &= \kappa_j + \lambda_j + \phi_{j,1} \\ \phi_{j,3} &= \frac{2\kappa_j\theta_j}{\sigma_j^2} \\ \phi_{j,4} &= 2\phi_{j,1} + \phi_{j,2}\big(\exp(\phi_{j,1}T_i) - 1\big).\end{aligned}$$

In these equations, κ_j, θ_j σ_j and λ_j are constants. Here, T_i is a non-negative num-

ber which, in practice, represents the time to maturity of a pure discount bond and $\mathcal{Y}_i(k)$ is the corresponding price of the bond at time t_k. Note that each T_i only appears in the measurement equation for $\mathcal{Y}_i(k)$. We assume that the bond prices $\mathcal{Y}_1(k)$, $\mathcal{Y}_2(k)$, etc are observed in noise $\mathcal{V}_i(k)$ which is assumed to be bounded and have a mean of zero.

Remark 9.2

One may use $-\log(\mathcal{Y}_i(k))$ *as a measurement, which yields a linear measurement equation in* $\mathcal{X}_j(k)$. *We shall use* $\mathcal{Y}_i(k)$ *as a measurement to illustrate the performance of the proposed filter wherein the state space system involves a nonlinear unobservable dynamics as well as a nonlinear measurement equation.*

The parameters used for this model are the same as those used in the numerical demonstration in [57] and are presented in Table 9.1.

κ_1	0.0718	σ_1	0.2160
κ_2	0.7830	σ_1	1.2200
θ_1	4.3000	λ_1	-0.2130
θ_2	1.6400	λ_2	-0.9140

Table 9.1: Parameters in the implementation of the system specified in (9.10) – (9.11).

We use $T_1 = 0.5$, $T_2 = 1$, $T_3 = 2$ and employ the corresponding $\mathcal{Y}_1(k)$, $\mathcal{Y}_2(k)$ and $\mathcal{Y}_3(k)$ as the observations at each time t_k. This gives a two-state, three-measurement state space system. Based on a simulated observation sample path, we wish to see whether we can predict $\mathcal{Y}_i(k+1 \mid k)$ at each t_k accurately, where

$$\mathcal{Y}_i(k+1 \mid k) = \sum_{l=1}^{4s+1} p_l \prod_{j=1}^{2} A_{i,j} \exp\left(- \sum_{j=1}^{2} B_{i,j} \mathcal{X}_j^{(l)}(k+1 \mid k) \right)$$

is the predicted bond price. Here, $4s+1$ is the total number of sigma points for $\mathcal{W}^{(i)}(k+1)$ (since $m = 2$) and p_l's are the corresponding $4s+1$ probability weights.

We would like to compare the predictive ability of the PLSPF proposed here to that of the EKF. At time $t_k + 1$, EKF uses linearised versions of equations (9.10) – (9.11) around the updated state estimate $\hat{\mathcal{X}}(k \mid k)$ at time t_k and then uses the standard Kalman filter for state prediction and update. The formulae for the EKF based on

$$\mathbb{E}\left(\mathcal{X} \mid \mathcal{Z}\right) = \mathbb{E}\left(\mathcal{X}\right) + \Sigma_{XZ}\Sigma_{ZZ}^{-1}\big(\mathcal{Z} - \mathbb{E}(\mathcal{Z})\big)$$

are not repeated here; they are provided in chapter 8.

To measure the performance of a filter, we consider the average of root mean squared error (AvRMSE) as well as the average of mean relative absolute error (AvMRAE) in one-step ahead predictions. The root mean squared error (RMSE) for a measurement $\mathcal{Y}_j$ and for a particular sample path i is given by

$$\text{RMSE}_{(i,j)} = \sqrt{\frac{1}{T}\sum_{k=1}^{T}\Big(\big(\mathcal{Y}_j(k+1)\big)_i - \big(\mathcal{Y}_j(k+1 \mid k)\big)_i\Big)^2},$$

where T is the time horizon. Here, $\big(\mathcal{Y}_j(k+1)\big)_i$ (respectively, $\big(\mathcal{Y}_j(k+1 \mid k)\big)_i$) denotes the noisy observation of bond price, $\mathcal{Y}_j(k+1)$ (respectively, the prediction of bond price, $\mathcal{Y}_j(k+1 \mid k)$) for the i-th sample path. AvRMSE_j is computed as the sample mean of $\text{RMSE}_{(i,j)}$ over different sample paths i, that is,

$$\text{AvRMSE}_j = \frac{1}{S}\sum_{i=1}^{S}\text{RMSE}_{(i,j)} \quad \text{for} \quad j \in \{1,2,3\}.$$

In a similar fashion, MRAE for measurement $\mathcal{Y}_j$ and sample path i is defined by

$$\text{MRAE}_{(i,j)} = \frac{1}{T}\sum_{k=1}^{T}\frac{\big|\big(\mathcal{Y}_j(k+1)\big)_i - \big(\mathcal{Y}_j(k+1 \mid k)\big)_i\big|}{\big(\mathcal{Y}_j(k+1)\big)_i}$$

and AvMRAE_j is computed as the sample mean of $\text{MRAE}_{(i,j)}$ over different sample paths i, i.e.,

$$\text{AvMRAE}_j = \frac{1}{S}\sum_{i=1}^{S} \text{MRAE}_{(i,j)} \quad \text{for} \quad j \in \{1, 2, 3\}.$$

Both these functions of prediction error, AvRMSE_j and AvMRAE_j, are computed over $S = 100$ sample paths, with each path consisting of $T = 250$ time steps, for each of the three measurements $\mathcal{Y}_1(k)$, $\mathcal{Y}_2(k)$ and $\mathcal{Y}_3(k)$. At each time step only 13 samples or sigma points are generated, which corresponds to choosing $s = 3$ for the algorithm in section 9.3. The results of this error analysis for PLSPF are reported in Table 9.2. Figure 9.1, on the other hand, displays a graphical comparison between the simulated $\mathcal{Y}_i(k)$ (blue, solid line) and the predicted $\mathcal{Y}_j(k+1 \mid k)$ (green, solid line) for one particular sample path. The mean computation time per sample path for PLSPF is 65.88 seconds, with the maximum time per sample path being 71.03 seconds. In other words, the performance with PLSPF is achieved at the cost of only around 0.27 seconds per time step. The experiments were also repeated for four measurements and three states and the mean computation time per sample path in this case is 117.15 seconds, with the maximum time per sample path at 129.32 seconds. The results of error analysis for the case of four measurements and three states are reported in Table 9.3. This computation was carried out on a desktop with Pentium IV core duo processor (2.4 Ghz), running MATLAB version R2007b on Windows XP. The computation time can easily be improved by employing a purpose-written optimisation code or a higher specification machine. Clearly, this computation time is affordable even for real time processing involving applications where the estimation of state dynamics is suffciently slow, such as on-line estimation problems in many chemical processes.

The state estimation results with EKF in the present example are significantly worse, with the filter diverging in 60 out of 100 sample paths and yielding extremely large errors. The average errors over the remaining 40 sample paths are still high,

	$j=1$	$j=2$	$j=3$
AvMRAE_j	0.000498	0.000589	0.000795
AvRMSE_j	0.000525	0.000616	0.000764

Table 9.2: Average errors in predicting $\mathcal{Y}_j(k+1)$ with PLSPF for three measurement, two-state case (average over 100 sample paths, with 250 time steps in each sample path).

	$j=1$	$j=2$	$j=3$	$j=4$
AvMRAE_j	0.000690	0.000960	0.001941	0.002954
AvRMSE_j	0.000468	0.000665	0.001394	0.002174

Table 9.3: Average errors in predicting $\mathcal{Y}_j(k+1)$ with PLSPF for four measurement, three-state case (average over 100 samplepaths, with 250 time steps in each sample path).

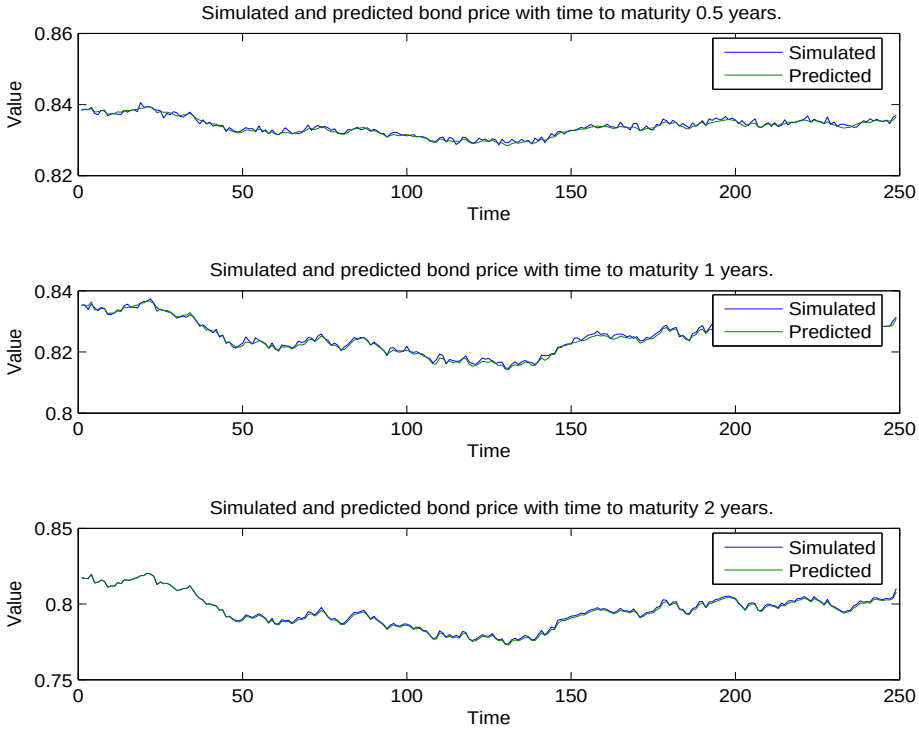

Figure 9.1: Prediction for $\mathcal{Y}_1(k+1)$, $\mathcal{Y}_2(k+1)$ and $\mathcal{Y}_3(k+1)$ using PLSPF.

even with the lowest average error being over ten times the corresponding error with the PLSPF, as can be seen from Table 9.4.

	$j=1$	$j=2$	$j=3$
AvMRAE_j	0.006915	0.014729	0.034515
AvRMSE_j	0.013974	0.030556	0.074106

Table 9.4: Average errors in predicting $\mathcal{Y}_j(k+1)$ with EKF (average over 40 sample paths on which the filter did not diverge, with 250 time-steps in each sample path).

These numerical experiments clearly indicate the superiority of the proposed algorithm over the EKF for nonlinear systems of the form (9.1), in the case when the measurement equation is sufficiently smooth.

9.5 Concluding remarks

In this chapter, we put forward new filtering heuristics for nonlinear and non-Gaussian systems, which we refer to as PLSPF. This algorithm shares some of the advantages of the modified sigma point filter proposed in chapter 8, in the sense that the state covariance matrix need not be factorised at each step and the first three moments are exactly matched. However, unlike conventional sigma point filters, the state update step in PLSPF does not use closed-form formula based on the Gaussianity assumption. Instead, a simple and intuitively appealing optimisation is utilised where the measurement equation is linearised and the updated state which best matches the given observations in an appropriate deterministic sense is found. We demonstrated the implementation of the algorithm through a detailed numerical example involving a nonlinear, multivariate time series. As shown, the proposed method is a computationally simpler and attractive alternative to particle filtering for nonlinear time series in econometrics. Furthermore, it also provides a very useful alternative to traditional sigma point filters in engineering and ensemble filters in geosciences.

Chapter 10

Conclusions and directions for future research

10.1 Summary of contributions

In this thesis, various filtering approaches were put forward and multiple extensions as well as new methods proposed. The first part of this thesis focuses on the analysis of models where the noisy observation process, which can either be univariate or multivariate, is driven by a Markov chain. Such models, also called regime-switching models, are extremely flexible allowing model parameters to take values in accordance with the dynamic changes in different regimes or states of an economy. The main area of focus here was on parameter estimation. Throughout the entire process of estimating parameters, the change of probability measure technique in conjunction with the application of Bayes' theorem and the EM algorithm was employed. We provided refined, extended and new HMM-based optimal parameter estimation procedures via the development of adaptive filters both in the univariate and multivariate cases. In particular, the parameter updating expressions were given for the case where the drift and volatility have independent probabilistic behaviour for both univariate and multivariate observation settings. In addition, we explored how non-normal noise distributions affect the filtering process. More

specifically the formulae for updating the parameters of the model were derived when the noise term has a student's t-distributed noise term. Numerical examples were included to illustrate the applicability of the filters to financial and mortality data.

The second part of this thesis moves away from Markov chain filtering and deals with filtering methods for general nonlinear time series. It addresses problems embedded in extended Kalman filtering and its modifications when these are applied to highly nonlinear observations. Specifically, we examined unscented or sigma point filtering which has become popular in the recent years. Chapter 7.1 outlines the standard Kalman filter as well as its extensions, the extended Kalman filter and unscented Kalman filter. In addition, the similarities between HMM filtering and Kalman filtering are discussed. The two can be viewed as analogous algorithms utilising the same approaches and constructs to estimate optimally the unobservable state, as well as in recursive parameter estimation (see for example [1]). In chapter 7, a new algorithm for generation of sigma points from partially specified symmetric multivariate distribution was developed. The sigma points generated using such approach match the first two moments exactly, whilst the fourth moment is matched approximately. Moreover, this algorithm rectifies the problem of negative probability weights which arises in existing sigma point generation methodologies. Chapter 8 contains a new sigma point filtering technique which utilises the sample points generation algorithm developed in chapter 7 and is essentially still based on the idea underlying the Kalman filtering. Finally, an alternative technique for the optimal state estimation was proposed in chapter 9. It is based on a linear programming-based procedure during the update step and hence, it considerably departs from current methods based on Kalman filtering.

Each of the filtering techniques discussed and developed in this thesis was tested on observed or simulated data sets. The various empirical investigations afforded us with many important insights regarding the performance of the filtering methods that were advanced in this research work.

Additionally, this thesis made contributions to the field of actuarial science. In chapter 5, an HMM-based estimation of a mortality model is developed and tested on observed mortality data. Chapter 6, in turn, examined the valuation of contingent claims taking into account the integration of mortality and interest rate risks. Closed-form expressions for generic mortality-linked contingent claims were established and these were accompanied by some numerical demonstrations. It has to be noted that the valuation approach uses several changes of reference probability along with the Bayes' rule and therefore, the idea behind this approach is similar in spirit to the approach that underpins the HMM filtering.

10.2 Future directions

From the results presented in this thesis and outlined in the previous section, several research questions naturally arise. These questions lead to several directions that could be studied in the future.

- In the study of HMMs presented in chapters 3 and 4, novel extensions to the HMM filtering were developed. However, more empirical works are needed. There have been numerous studies based on HMM filtering with the usual assumption of equal number of states for the drift and volatility and standard normal noise term (see for example, [45], [49], [50] and [51]) but without a doubt, they could be improved further with the results presented in this thesis. Applications, especially to econometrics data involving an HMM-driven model with more general noise term, could spur more advancements in the area of financial modelling.

- The mortality derivative pricing equations developed in chapter 6 produce very considerable differences in prices obtained between two cases: namely the cases of independent and dependent assumptions on interest rate and mortality risks. It would be an interesting endeavour to determine the extent of this difference with observed data.

- If generation of scenario trees is needed for multi-stage decision problems, then an extension of the sigma point generation algorithm presented in chapter 7 is required. An implementation of a large scale stochastic programming model demonstrating the use of this method in financial optimisation is another possible research topic, which is a natural ramification of the moment matching algorithm presented in this thesis.

- New sigma point filtering techniques were designed in chapters 8 and 9. Probing the performance of these filtering methods using observed data along with the estimation of model parameters is definitely a new practical and important research consideration.

Appendix A

Additional plots for Chapter 4

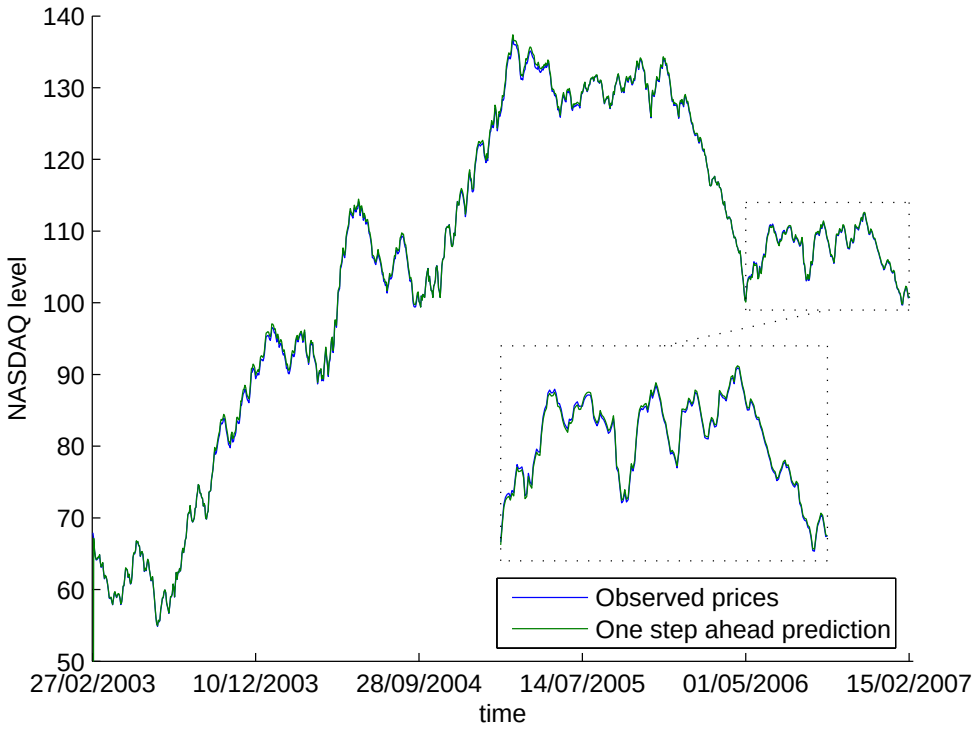

Figure A.1: NASDAQ actual series (blue) and one-step ahead predictions (green).

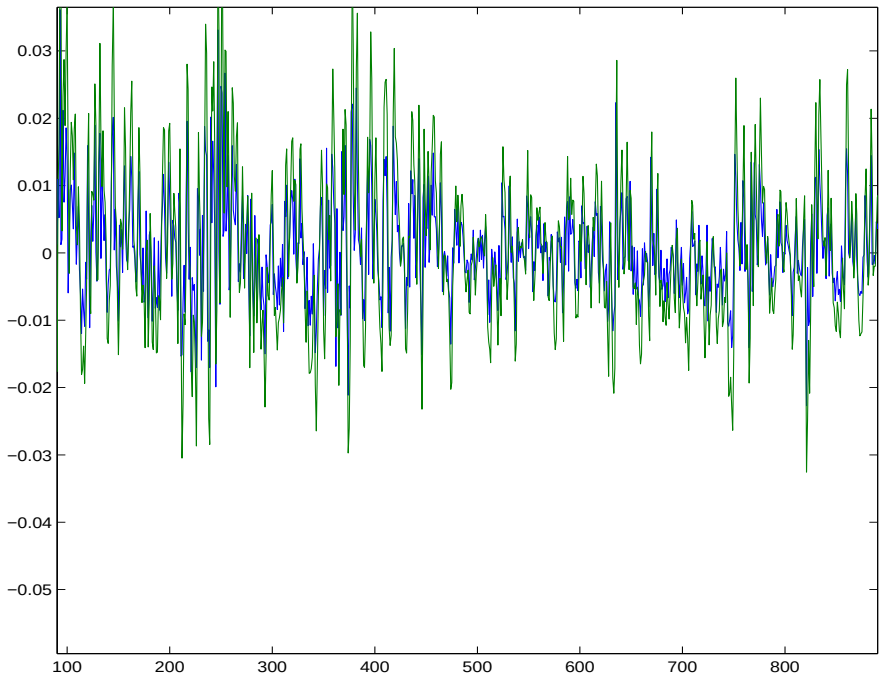

Figure A.2: NASDAQ returns (blue) and one-step ahead predictions (green).

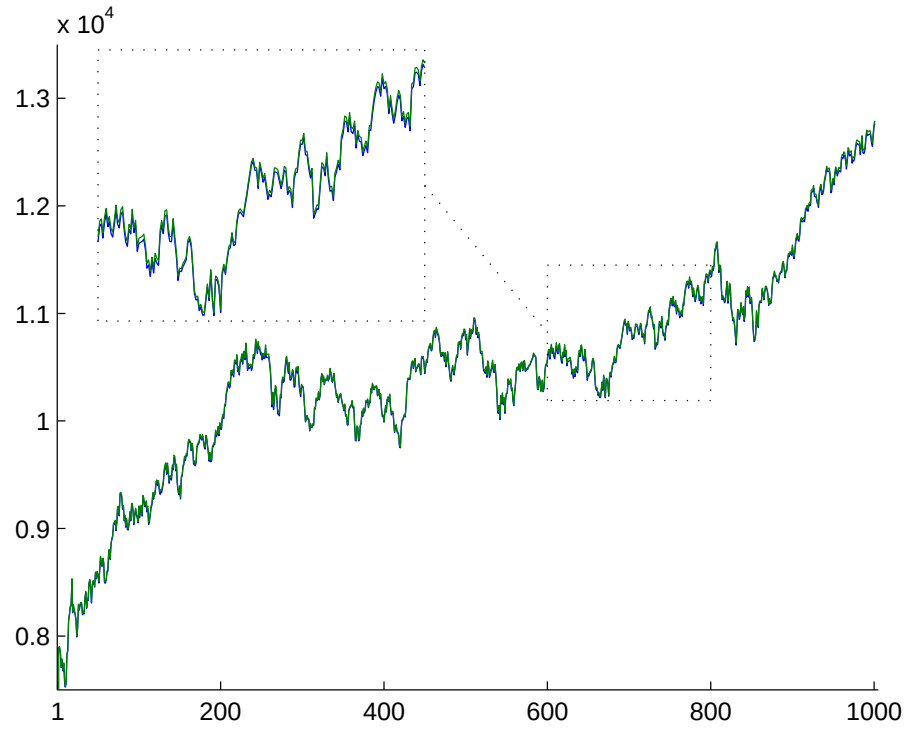

Figure A.3: DOW JONES actual series (blue) and one-step ahead predictions (green).

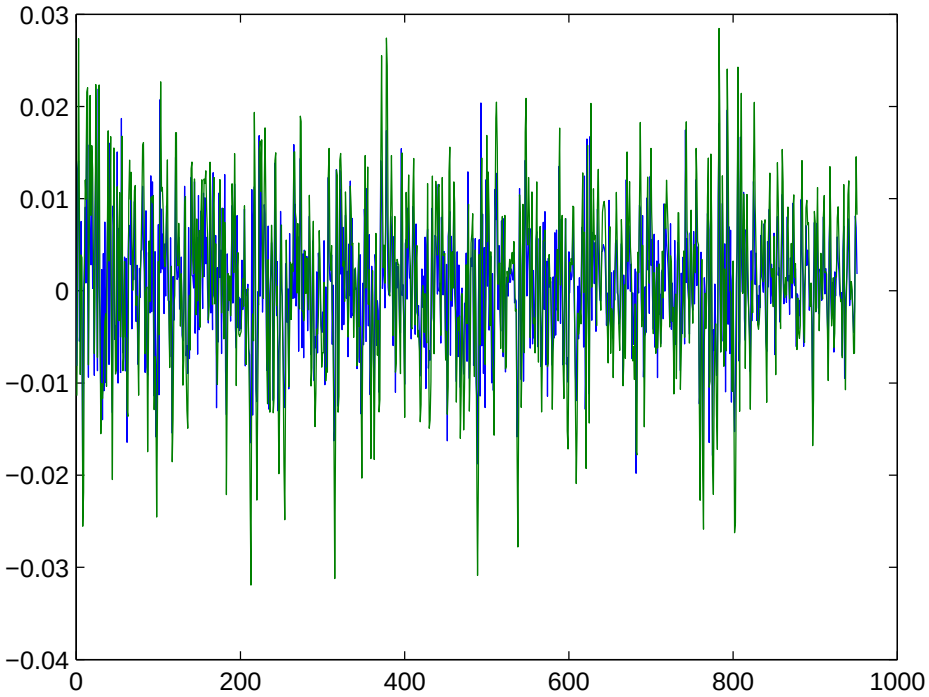

Figure A.4: DOW JONES returns (blue) and one-step ahead predictions (green).

Appendix B

First hitting time density of an Ornstein-Uhlenbeck process with constant parameters

Let W_t be a standard Brownian motion. The associated Ornstein-Uhlenbeck (OU) process U_t with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}$, is defined as the solution to the stochastic differential equation

$$dU_t = \mu U_t dt + \sigma dW_t, \quad U_0 \in \mathbb{R}. \tag{B.1}$$

Furthermore, the process U_t is a strong Markov process with infinitesimal generator, denoted by A, given by

$$Af(x) = \mu x \frac{\partial f}{\partial x} + \frac{\sigma^2}{2} \frac{\partial^2 f}{\partial x^2}, \quad x \in \mathbb{R}. \tag{B.2}$$

When integrated, the stochastic differential equation (B.1) yields the realisation

$$U_t = e^{\mu t}\Big(U_0 + \sigma \int_0^t e^{-\mu s} dW_s\Big) \tag{B.3}$$

for $t \geq 0$. By the Dambis, Dubins-Schwartz theorem, there exists a Brownian motion $\tilde{W}_t$, such that

$$\int_0^t e^{-\mu s} dW_s = \tilde{W}_{\tau(t)},$$

for any $t \geq 0$, where $\tau(t) = \big(2\mu\big)^{-1}\big(e^{2\mu t} - 1\big)$. Therefore, the representation

$$U_t = e^{\mu t}\big(U_0 + \sigma \tilde{W}_{\tau(t)}\big)$$

which is also known as Doob's transform holds. For a fixed real number a, define the stopping time

$$\lambda_a = \inf\big\{t > 0 \mid U_t = a\big\}.$$

The law of λ_a, denoted by $p^{\mu}_{x\to a}(t)$ is absolutely continuous with respect to a Lebesgue measure. We start with the assumption that μ is negative so that U_t is a recurrent process and therefore λ_a is finite.

Finding the density of a first barrier hitting time of an OU-process is generally a difficult problem. There are several approximation methods to calculate the density, however they are fairly complicated. See for example, Alili *et al* [3], Alili and Patie [2], Lo and Hui [86], amongst others.

For the special case $a = 0$ however, there is a simple expression for $p^{\mu}_{x\to 0}(t)$. Set $\tilde{\lambda_a} = \inf\big\{t > 0 \mid \tilde{W}_t = a\sqrt{1 - 2\mu t}\big\}$. As noted in Breiman [22], Doob's transform implies the identity $\lambda_a = \tilde{\lambda}_a$ a.s. Therefore, we can deduce

$$p^{\mu}_{x\to 0}(t) = \tau'(t) p^{0}_{x\to 0}\big(\tau(t)\big). \tag{B.4}$$

Furthermore, by letting $\mu \to 0$ we recover U_t as a Brownian motion rescaled by σ, i.e., $dU_t = \sigma dW_t$. Hence,

$$p^{0}_{x\to a}(t) = \frac{|a - x|}{\sigma\sqrt{2\pi t^3}} \exp\left(-\frac{(a-x)^2}{2\,\sigma^2 t}\right). \tag{B.5}$$

It follows from equations (B.4) and (B.5) that

$$p^{\mu}_{x\to 0}(t) = \frac{|x|}{\sigma\sqrt{2\pi}}\left(\frac{\mu}{\sinh(\mu t)}\right)^{3/2}\exp\left(-\frac{x^2\mu e^{\mu t}}{2\sigma^2\sinh(\mu t)} - \frac{\mu t}{2}\right) \tag{B.6}$$

Recall that if μ is positive, the process U_t is transient and formula (B.6) no longer applies. Nevertheless, one can still find fairly simple formulae for the density of the first hitting time. Denote by P^{μ}_x the law of U_t where $x = U_0 \in \mathbb{R}$. As before, letting $\mu \to 0$ we retrieve the law P^0_x of a σW_t started at x. Due to Girsanov's Theorem 2.4, the absolute-continuity relationship

$$dP^{\mu}_{x|\mathcal{F}_t} = \exp\left(-\frac{\mu}{2\sigma^2}\big(W_t^2 - x^2 - t\big) - \frac{\mu^2}{2\sigma^2}\int_0^t W_s^2 ds\right) dP_{x|\mathcal{F}_t} \tag{B.7}$$

holds for every $t \geq 0$. From the chain rule and equation (B.7) we can deduce as in [17] that

$$dP^{\mu}_{x|\mathcal{F}_t} = \exp\big(\mu(W_t^2 - x^2 - t)\big) dP^{-\mu}_{x|\mathcal{F}_t} \tag{B.8}$$

for $t > 0$. The expression in (B.8) combined with the optional stopping theorem yields

$$p^{\mu}_{x\to a}(t) = \exp\Big(\frac{\mu}{\sigma^2}\big(a^2 - x^2 - t\big)\Big) p^{-\mu}_{x\to a}(t). \tag{B.9}$$

Bibliography

[1] L. Aggoun and R.J. Elliott, *Measure theory and filtering: Introduction and applications*, Cambridge University Press, Cambridge, 2004.

[2] L. Alili and P. Patie, *On the first crossing times of a Brownian motion and a family of continuous curves*, Comptes Rendus de l'Academie des Sciences-Mathematiques **340** (2005), no. 3, 225–228.

[3] L. Alili, P. Patie, and JL Pedersen, *Representations of the First Hitting Time Density of an Ornstein-Uhlenbeck Process 1*, Stochastic Models **21** (2005), no. 4, 967–980.

[4] B. Anderson and J. Moore, *Optimal filtering*, Prentice-Hall, New Jersey, 1979.

[5] L. Ballotta and S. Haberman, *Valuation of guaranteed annuity conversion options*, Insurance: Mathematics and Economics **33** (2003), 87–108.

[6] R. Bansal and H. Zhou, *Term structure of interest rates with regime shifts*, Journal of Finance (2002), 1997–2043.

[7] L.E. Baum, *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*, Inequalities **3** (1972), no. 1, 1–8.

[8] L.E. Baum and J.A. Eagon, *An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology*, Bulletin of the American Mathematical Society **73** (1967), no. 360-363, 212.

[9] L.E. Baum and T. Petrie, *Statistical inference for probabilistic functions of finite state Markov chains*, Annals of Mathematical Statistics (1966), 1554–1563.

[10] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*, Annals of Mathematical Statistics (1970), 164–171.

[11] L.E. Baum and G.R. Sell, *Growth transformations for functions on manifolds*, Pacific Journal of Mathematics **27** (1968), no. 2, 211–227.

[12] P.M. Bentler and M. Berkane, *Greatest lower bound to the elliptical theory kurtosis parameter*, Biometrika **73** (1986), no. 1, 240–241.

[13] E. Biffis, *Affine processes for dynamic mortality and actuarial valuations*, Insurance: Mathematics and Economics **37** (2004), 443–468.

[14] T. Björk, *Interest rate theory: Financial mathematics, Springer lecture notes in mathematics*, Springer-Verlag, Berlin, 1997.

[15] N.P. Bollen, S.F. Gray, and R.E. Whaley, *Regime switching in foreign exchange rates: Evidence from currency option prices*, Journal of Econometrics **94** (2000), no. 1-2, 239–276.

[16] M.J. Bolton, D.H. Carr, P.A. Collis, C.M. George, V.P. Knowles, and A.J. Whitehouse, *Reserving for annuity guarantees*, Report of the Annuity Guarantees Working Party, Institute of Actuaries, London, UK (1997), 1–36.

[17] A.N. Borodin and P. Salminen, *Handbook of Brownian motion: Facts and formulae*, Birkhäuser, Basel, 2002.

[18] N.L. Bowers, H.U. Gerber, J.C. Hickman, D.A. Jones, and C.J. Nesbitt, *Actuarial mathematics, 2nd edition*, Society of Actuaries, Schaumburg, 1997.

[19] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*, SIAM Studies in Applied Mathematics, Philadelphia, 1994.

[20] P. Boyle and T. Draviam, *Pricing exotic options under regime switching*, Insurance: Mathematics and Economics **40** (2007), no. 2, 267–282.

[21] P. Boyle and M. Hardy, *Guaranteed annuity options*, ASTIN Bulletin **33** (2003), no. 2, 125–152.

[22] L. Breiman, *First exit times from a square root boundary*, Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability **2** (1967), 1–16.

[23] N. Brouhns, M. Denuit, and J.K. Vermunt, *A Poisson log-bilinear regression approach to the construction of projected lifetables*, Insurance: Mathematics and Economics **31** (2002), no. 3, 373–393.

[24] A.J. Cairns, D. Blake, and K. Dowd, *Pricing death: Frameworks for the valuation and securitisation of mortality risk*, ASTIN Bulletin **36** (2006), no. 1, 79–120.

[25] A.J. Cairns, D.P. Blake, K. Dowd, and J. Campus, *A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration*, Journal of Risk & Insurance **73** (2006), no. 4, 687–718.

[26] O. Cappé, E. Moulines, and T. Ryden, *Inference in hidden Markov models*, Springer Verlag, New York, 2005.

[27] B.P. Carlin, N.G. Polson, and D.S. Stoffer, *A Monte Carlo approach to non-normal and nonlinear state-space modeling*, Journal of the American Statistical Association (1992), 493–500.

[28] J.F. Carriere, *An investigation of the Gompertz law of mortality*, Actuarial Research Clearing House **2** (1994), 161–177.

[29] W. Chang and H. Tsai, *On dynamic tax reform with regime switching*, Public Finance Review **34** (2006), no. 3, 306–327.

[30] C.S. Chu, G.J. Santoni, and T. Liu, *Stock market volatility and regime shifts in returns*, Information Sciences **94** (1996), no. 1, 179–190.

[31] J.C. Cox, *The constant elasticity of variance option pricing model*, Journal of Portfolio Management, December (1996).

[32] J.C. Cox, J.E. Ingersoll Jr, and S.A. Ross, *A theory of the term structure of interest rates*, Econometrica **53** (1985), no. 2, 385–407.

[33] I.D. Currie, M. Durban, and P.H. Eilers, *Smoothing and forecasting mortality rates*, Statistical Modelling **4** (2004), 279–298.

[34] M. Dahl, *Stochastic mortality in life insurance: Market reserves and mortality-linked insurance contracts*, Insurance: Mathematics and Economics **35** (2004), 113–136.

[35] P. Date, L. Jalen, and R.S. Mamon, *A new algorithm for latent state estimation in non-linear time series models*, Applied Mathematics and Computation **203** (2008), no. 1, 224–232.

[36] ———, *A partially linearised sigma point filter for latent state estimation in nonlinear time series models*, Applied Mathematical Modelling (2009), submitted.

[37] P. Date, R.S. Mamon, and L. Jalen, *A new moment matching algorithm for sampling from partially specified symmetric distributions*, Operations Research Letters **36** (2008), no. 6, 669–672.

[38] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B (Methodology) (1977), 1–38.

[39] D. Duffie, D. Filipovič, and W. Schachermayer, *Affine processes and applications in finance*, Annals of Applied Probability **13** (2003), no. 3, 984–1053.

[40] D. Duffie and K.J. Singleton, *Credit risk: Pricing, measurement, and management*, Princeton University Press, Princeton, 2003.

[41] R.J. Elliott, *New finite-dimensional filters and smothers for noisily observed markov chains*, IEEE Transactions on Information Theory **39** (1993), 265–271.

[42] ———, *Exact adaptive filters for Markov chains observed in Gaussian noise*, Automatica **30** (1994), 1399–1408.

[43] R.J. Elliott, L. Aggoun, and J.B. Moore, *Hidden Markov models: Estimation and control*, Springer Verlag, New York, 1995.

[44] R.J. Elliott, W.C. Hunter, and B.M. Jamieson, *Drift and volatility estimation in discrete time*, Journal of Economic Dynamics and Control **22** (1998), no. 2, 209–218.

[45] ———, *Financial signal processing: A self calibrating model*, International Journal of Theoretical and Applied Finance **4** (2001), no. 4, 567–584.

[46] R.J. Elliott and V. Krishnamurthy, *New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models*, IEEE Transactions on Automatic Control **44** (1999), no. 5, 938–951.

[47] R.J. Elliott and J. van der Hoek, *Stochastic flows and the forward measure*, Finance and Stochastics **5** (2001), no. 4, 511–525.

[48] Y. Ephraim and N. Merhav, *Hidden Markov processes*, IEEE Transactions on Information Theory **48** (2002), no. 6, 1518–1569.

[49] C. Erlwein, F.E. Benth, and R.S. Mamon, *HMM filtering and parameter estimation of an electricity spot price model*, Technical report, Department of Mathematics, University of Oslo, E-print No 2 (2007).

[50] C. Erlwein and R.S. Mamon, *An online estimation scheme for a Hull–White model with HMM-driven parameters*, Statistical Methods and Applications **18** (2009), no. 1, 87–107.

[51] C. Erlwein, G. Mitra, and D. Roman, *HMM based scenario generation for an investment optimization problem*, Technical Report, CTR/68/07, CARISMA, Brunel University (2007).

[52] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, Journal of Geophysical Research-All Series **99** (1994), no. C5, 143–162.

[53] ———, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean Dynamics **53** (2003), no. 4, 343–367.

[54] A. Farina, B. Ristic, D. Benvenuti, and A.M. Syst, *Tracking a ballistic target: Comparison of several nonlinear filters*, IEEE Transactions on Aerospace and Electronic Systems **38** (2002), no. 3, 854–867.

[55] D. Filipovič, *Time-inhomogeneous affine processes*, Stochastic Processes and their Applications **115** (2005), 639–659.

[56] P. Gahinet, A. Nemirovski, A.J. Laub, and M. Chilali, *LMI control toolbox*, The MathWorks, Inc., Massachusetts, 1995.

[57] A.L. Geyer and S. Pichler, *A state-space approach to estimate and test multifactor Cox-Ingersoll-Ross models of the term structure*, Journal of Financial Research **22** (1999), 107–130.

[58] I.V. Girsanov, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory of Probability and its Applications **5** (1960), no. 3, 285–301.

[59] N.J. Gordon, D.J. Salmond, and A.F. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, IEE Proceedings F: Radar and Signal Processing **140** (1993), no. 2, 107–113.

[60] G. Grimmett and D. Stirzaker, *Probability and random processes*, Oxford University Press, New York, 2001.

[61] N. Gulpinar, B. Rustem, and R. Settergren, *Simulation and optimization approaches to scenario tree generation*, Journal of Economic Dynamics and Control **28** (2004), no. 7, 1291–1316.

[62] N. Haldrup and M.Ø. Nielsen, *A regime switching long memory model for electricity prices*, Journal of Econometrics **135** (2006), no. 1-2, 349–376.

[63] J.D. Hamilton, *Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates*, Journal of Economic Dynamics and Control **12** (1988), no. 2-3, 385–423.

[64] ———, *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica **57** (1989), no. 2, 357–384.

[65] ———, *Analysis of time series subject to changes in regime*, Journal of Econometrics **45** (1990), no. 1-2, 39–70.

[66] ———, *Time series analysis*, Princeton University Press, Princeton, 1994.

[67] A. Hermoso-Carazo and J. Linares-Perez, *Different approaches for state filtering in nonlinear systems with uncertain observations*, Applied Mathematics and Computation **187** (2007), no. 2, 708–724.

[68] R. Hochreiter and G.C. Pflug, *Financial scenario generation for stochastic multi-stage decision processes as facility location problems*, Annals of Operations Research **152** (2007), no. 1, 257–272.

[69] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1990.

[70] P.L. Houtekamer and H.L. Mitchell, *Data assimilation using an ensemble Kalman filter technique*, Monthly Weather Review **126** (1998), no. 3, 796–811.

[71] K. Høyland, M. Kaut, and S.W. Wallace, *A heuristic for moment-matching scenario generation*, Computational Optimization and Applications **24** (2003), no. 2, 169–185.

[72] K. Høyland and S.W. Wallace, *Generating scenario trees for multistage decision problems*, Management Science (2001), 295–307.

[73] L. Jalen and R.S Mamon, *Valuation of contingent claims with mortality and interest rate risks*, Mathematical and Computer Modelling **49** (2009), no. 9-10, 1893–1904.

[74] J. James and N. Webber, *Interest Rate Modelling*, Wiley, Chichester, 2000.

[75] A.H. Jazwinski, *Stochastic processes and filtering theory*, Academic Press, New York, 1970.

[76] S.J. Julier and J.K. Uhlmann, *Unscented filtering and nonlinear estimation*, Proceedings of the IEEE **92** (2004), no. 3, 401–422.

[77] R.E. Kalman, *A new approach to linear filtering and prediction problems*, Journal of basic Engineering **82** (1960), no. 1, 35–45.

[78] R. Karlsson, T. Schon, and F. Gustafsson, *Complexity analysis of the marginalized particle filter*, IEEE Transactions on Signal Processing **53** (2005), no. 11, 4408–4411.

[79] M. Kaut and S.W. Wallace, *Evaluation of scenario-generation methods for stochastic programming*, Pacific Journal of Optimization **3** (2007), no. 2, 257–271.

[80] G. Kitagawa, *Non-Gaussian state-space modeling of nonstationary time series*, Journal of the American Statistical Association (1987), 1032–1041.

[81] ———, *Monte Carlo filter and smoother for non-Gaussian nonlinear state space models*, Journal of Computational and Graphical Statistics (1996), 1–25.

[82] H.J. Kushner and P. Dupuis, *Numerical methods for stochastic control problems in continuous time*, Springer Verlag, New York, 2001.

[83] R.D. Lee, *The Lee-Carter method of forecasting mortality with various extensions and applications*, North American Actuarial Journal **4** (2000), no. 1, 80–93.

[84] R.D. Lee and L.R. Carter, *Modelling and forecasting US mortality*, Journal of American Statistical Association **87** (1992), no. 419, 659–671.

[85] J.S. Liu and R. Chen, *Sequential Monte Carlo methods for dynamic systems*, Journal of the American Statistical Association **93** (1998), no. 443, 1032–1044.

[86] C.F. Lo and C.H. Hui, *Computing the first passage time density of a time-dependent Ornstein-Uhlenbeck process to a moving boundary*, Applied Mathematics Letters **19** (2006), no. 12, 1399–1405.

[87] E. Luciano and E. Vigna, *Non-mean reverting affine processes for stochastic mortality*, Proceedings of the 15th International AFIR Colloquium, Zurich. Available online at http://www.afir2005.ch (2005).

[88] J. Lund, *Non-linear Kalman filtering techniques for term-structure models*, Working paper, The Aarhus School of Bussiness, Department of Finance, available online at www.jesperlund.com/papers/kfnlin.pdf (1997).

[89] P.M. Lurie and M.S. Goldberg, *An approximate method for sampling correlated random variables from partially-specified distributions*, Management Science (1998), 203–218.

[90] A. Macdonald, D. Bartlett, C. Berman, C. Daykin, D. Grimshaw, P. Savill, and R. Willets, *Mortality improvements and cohort effect*, Continuous Mortality Investigation Working Papers 1 and 2 (2003), 1–58, Presented to the Staple Inn Actuarial Society. Available online at www.sias.org.uk.

[91] R.S. Mamon and R.J. Elliott, *Hidden Markov Models in Finance*, International Series in Operations Research & Management Science, Volume 104, Springer Verlag, New York, 2007.

[92] R.S. Mamon, C. Erlwein, and R. Bhushan Gopaluni, *Adaptive signal processing of asset price dynamics with predictability analysis*, Information Sciences **178** (2008), no. 1, 203–219.

[93] R.S. Mamon and L. Jalen, *Parameter estimation in a regime-switching model when the drift and volatility are independent*, Proceedings of the 5th International Conference on Dynamic Systems and Applications, Dynamic Publishers, Atlanta (2008), 291–298.

[94] G.J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, Wiley, New York, 1997.

[95] M.A. Milevsky and S.D. Promislow, *Mortality derivatives and the option to annuitise*, Insurance: Mathematics and Economics **29** (2001), 299–318.

[96] L.T. Nielsen, *Pricing and hedging of derivative securities*, Oxford University Press, Oxford, 1999.

[97] J.R. Norris, *Markov chains*, Cambridge University Press, Cambridge, 1997.

[98] G.C. Pflug, *Scenario tree generation for multiperiod financial optimization by optimal discretization*, Mathematical Programming **89** (2001), no. 2, 251–271.

[99] G.L. Plett, *Sigma-point Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 1: Introduction and state estimation*, Journal of Power Sources **161** (2006), no. 2, 1356–1368.

[100] A. Renshaw and S. Haberman, *Lee–Carter mortality forecasting with age-specific enhancement*, Insurance: Mathematics and Economics **33** (2003), no. 2, 255–272.

[101] ———, *Lee-Carter mortality forecasting: A parallel generalized linear modelling approach for England and Wales mortality projections*, Applied Statistics **52** (2003), 119–137.

[102] J.H. Rogers, *The currency substitution hypothesis and relative money demand in Mexico and Canada*, Journal of Money, Credit and Banking (1992), 300–318.

[103] S. Roweis and Z. Ghahramani, *A unifying review of linear Gaussian models*, Neural Computation **11** (1999), no. 2, 305–345.

[104] W. Rudin, *Real and complex analysis*, McGraw-Hill, New York, 1987.

[105] R. Saigal, *Linear programming: A modern integrated analysis*, Journal of the Operational Research Society **48** (1997), no. 7, 762.

[106] David F. Schrager, *Affine stochastic mortality*, Insurance: Mathematics and Economics **38** (2006), no. 1, 81–97.

[107] J.E. Smith, *Moment methods for decision analysis*, Management Science (1993), 340–358.

[108] H. Tanizaki and R.S. Mariano, *Prediction, filtering and smoothing in non-linear and non-normal cases using Monte Carlo integration*, Journal of Applied Econometrics (1994), 163–179.

[109] N. Topaloglou, H. Vladimirou, and S.A. Zenios, *CVaR models with selective hedging for international asset allocation*, Journal of Banking and Finance **26** (2002), no. 7, 1535–1561.

[110] O. Vasiček, *An equilibrium characterization of the term structure*, Journal of Financial Economics **5** (1977), no. 2, 177–188.

[111] G. Wang, *On the latent state estimation of nonlinear population dynamics using Bayesian and non-Bayesian state-space models*, Ecological Modelling **200** (2007), no. 3-4, 521–528.

[112] H. Wolkowicz, R. Saigal, and L. Vandenberghe, *Handbook of semidefinite programming: theory, algorithms, and applications*, Kluwer Academic Publishers, Massachusetts, 2000.

[113] C.F.Jeff Wu, *On the convergence properties of the em algorithm*, Annals of Statistics **11** (1983), no. 1, 95–103.

[114] M. Zakai, *On the optimal filtering of diffusion processes*, Probability Theory and Related Fields **11** (1969), no. 3, 230–243.