# A New Regime-Switching Model for Financial Time Series

A Dissertation Presented

by

**Xiaochu Zhang**

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

**Doctor of Philosophy**

in

**Applied Math and Statistics**

Stony Brook University

December 2012

**Stony Brook University**

The Graduate School

**Xiaochu Zhang**

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Robert Frey – Dissertation Advisor
Research Professor, Department of Applied Math and Statistics

Svetlozar Rachev – Dissertation Co-Advisor
Professor, Department of Applied Math and Statistics

Andrew Mullhaupt – Chairperson of Defense
Research Professor, Department of Applied Math and Statistics

Dmytro Holod
Professor, College of Business

Noah Smith
Assistant Professor, College of Business

This dissertation is accepted by the Graduate School.

Charles Taber
Interim Dean of the Graduate
School

Abstract of the Dissertation

# A New Regime-Switching Model for Financial Time Series

by

**Xiaochu Zhang**

**Doctor of Philosophy**

in

**Applied Math and Statistics**

Stony Brook University

2012

In the first part of this dissertation research, an extension of the binomial tree model to a regime-switching volatility model in a two-state setting for volatility is derived, analyzed and tested. A dynamic programming method for mean-variance hedging is applied to price European option value. After convergence and simulation study, we demonstrate that an HMM driven stochastic volatility process will converge to a geometric brownian motion with a constant volatility.

In the second part, we further incorporate an autoregressive component into the regime switching model based on observations of the first part and derive an autoregressive regime-switching model for financial time series data. A parsimonious estimation method of autoregressive regime-switching model is developed using Gram-Schmidt orthogonalization, Frobenius norm minimization, and the EM algorithm. The positive semi-definite correlation matrix issue is considered and addressed in our estimation method. Stability and accuracy is also examined in part two.

In the third part, observations based on the analysis of real financial time series are shown. The integrated, fractional integrated and heavy tail feature of nonstationary time series is studied. Estimation, forecasting and backtesting are performed with ARMA-GARCH, FARIMA-FIGARCH and our autoregressive regime-switching model. In comparison with other models, autoregressive regime-switching model has better backtesting results for forecasting VaR models with high frequency financial time series.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

This thesis concludes my study at StonyBrook. I would like to express my gratitude to all the kind, smart, and dedicated people who have supported me to get this far, although not all of them are mentioned here.

I owe my deepest gratitude to my advisor, Dr. Robert Frey, for his time, ideas, and funding. He introduced me to mathematical finance field and he kept me up to date with the interactions between industry and academia in this field. This thesis would not have existed without his patient dedication in guiding me. I also thank Robert for being understanding and caring when I faced difficulties in my personal life.

I would like to extend my sincere gratitude to my co-advisor Dr Svetlozar Rachev, for his wise direction, patient assistance, and insightful guidance week by week during the last year of my doctorate research. Without his generous help, a large part of my Ph.D. work would not have been possible.

I would like to thank Dr. Andrew Mullhaupt who has patiently helped me with my numerous questions, including the ”Polya Eggenberger Urn Model” handout he wrote for me. He helped me to go through a tough transition period of my Ph.D. research.

I am grateful for Dr. Dmytro Holod being on my dissertation committee.

I am also thankful to Jimmie Goode, who proofread and edited this dissertation carefully.

Finally, this thesis would not have been possible without the continuous love and support

from my family. My parents always have faith in me. My husband, Tianbing, has stood by me through good times and bad times. I am especially thankful for my husbands endurance for my long working hours. His encouragements together with home cooking meals surely cheered me up when I felt frustrated.

# Chapter 1

# Introduction

There has been rising interest in modeling financial time series with regime-switching models (Tu [2010], Sims and Zha [2006], Mike et al. [1998]). This model assumes autoregressive parameters are dependent on hidden states, which is the Markov-switching process. In this chapter, we review some recent works on regime-switching model and its applications in different problems in the first section. The outline of this dissertation is given in the second section.

## 1.1 Literature on regime switching problems

An early example of regime switching model is considered in Quandt [1958], which studies consumption function. This model assumes a linear regression model between consumption variable $y_t$ and income variable $x_t$ which is a $(d \times 1)$ vector for $t = 1, ..., T$ . For any $t$, the observations are generated by one of two regimes:

$$y_t = x_t'\beta_1 + \mu_{1t},$$

$$y_t = x_t'\beta_2 + \mu_{2t},$$

where $\beta_1$ and $\beta_2$ are $(d \times 1)$ coefficient vectors, and $\mu_{1t}$ and $\mu_{2t}$ follow normal distributions $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively. There is an essential factor underline this consumption function, and it is not observable. This simple assumption renders parameter estimation feasible with limited computation capacities in 1950's. Quandt [1972] extends this work by introducing a probability model to model "that nature chooses between regimes with probability $\eta_1$ and $1 - \eta_1$".

The simple assumption that there is only one regime shift at an unknown point is not realistic and useful. A new method is described in Goldfeld and Quandt [1973] where the regime switching process is represented by a Markov chain. This method permits multiple switches, which means the system can switch back and forth between different states, which is more realistic and useful. Since 1970s, improvements on simulation estimation methodologies make it possible to solve this more complex model. A suggestion was made by Cosslett and Lee [1985] to adopt a recursive algorithm that is computationally tractable for the evaluation of the likelihood function. Furthermore, Hamilton [1989] suggests a filter algorithm. The recursive algorithms in Cosslett and Lee [1985] and Hamilton [1989] demonstrate efficient estimation and interference with likelihood function.

A well known econometric regime-switching model is presented in Hamilton [1989]. The model is a nonlinear generalization of an unobserved components trend and cycle model, and parameter estimation can be calculated as a by-product of an iterative algorithm similar in spirit to the Kalman filter. It is observed that the usual numerical maximum of the likelihood functions is subject to computational difficulties associated with the often ill-behaved likelihood surface (multiple local maxima, essential singularities, and local increases as boundary conditions are approached). Expectation-Maximization (EM) algorithm(Karlis and Xekalaki [2003], McLachlan and Peel [2000]) is used to overcome the numerical difficulties. Neftci [1984] suggests a model with transition probabilities that are duration dependent. It is clear that understanding duration dependence in business cycle is important for understand and forecast business cycle and economic nature.

A regime switching model with no autoregressive elements has been first investigated by Lindgren [1978] who proves a consistency property of maximum-likelihood estimators obtained for the model which assumes an independent sequence of hidden states from a finite mixture distribution. Lindgren's result states that, in case $y_t$ actually follows a hidden Markov model, the maximum-likelihood estimators obtained under the independence model are consistent for the stationary distribution of $y_t$. Regime-switching models that incorporate autoregressive elements can be located in the speech recognition literature Rabiner [1990] and Juang and Rabiner [1985].

Most models assume a stationary Markov transition process, and also assume only two or three regimes. Calvet and Fisher [2004] suggests a model with a much larger number of regimes. This multifractal models afford another approach for incorporating long-memory into volatility forecasting. Sims and Zha [2006] also advocates a model with a much larger number of regimes. This model parameters are estimated with prior Bayesian information.

Formal tests of the null hypothesis of no Markov switching have been proposed by Garcia [1998], Hansen [2006], Hamilton and Perez-Quiros [1996] and Carrasco et al. [2004]. The problem is to test the null hypothesis that there are $K$ regimes against the alternative of $K + 1$. When $K = 1$, it is to test whether there are any shifts in regimes at all. The parameters driving the dynamic of the underlying Markov chain are not identified under the null hypothesis. As a result, the testing problem is non-standard and the likelihood ratio test does not converge to a chi-square distribution. Garcia [1998], studies the asymptotic distribution of a sup-type Likelihood ratio test. Hansen [2006] treats the likelihood as a empirical process indexed by all the parameters (those identified and those unidentified under the null). His test relies on taking the supremum of likelihood ratio over the nuisance parameters. Both papers require estimating the model under the alternatives, which may be cumbersome. Carrasco et al. [2004] derives a class of information matrix-type tests and show that they are equivalent to the likelihood ratio test. Hence, our tests are asymptotically optimal. Moreover these tests are easy to implement as they do not require the estimation of the model under the alternative.

## 1.2 Outline

In the first part of this dissertation research, an extension of the binomial tree model to a regime-switching volatility model in a two-state setting for volatility is derived, analyzed and tested. A dynamic programming method for mean-variance hedging is applied to price European option value. After convergence and simulation study, we demonstrate that an HMM driven stochastic volatility process will converge to a geometric brownian motion with a constant volatility.

In the second part, we further incorporate an autoregressive component into the regime switching model based on observation s of the first part and derive an autoregressive regime-switching model for financial time series data. A parsimonious estimation method of autoregressive regime-switching model is developed using Gram-Schmidt orthogonalization, Frobenius norm minimization, and the EM algorithm. The positive semi-definite correlation matrix issue is considered and addressed in our estimation method. Stability and accuracy is also examined in part two.

In the third part, observations based on the analysis of real financial time series are shown. The integrated, fractional integrated and heavy tail feature of non-stationary time series is studied. Estimation, forecasting and backtesting are performed with ARMA-GARCH, FARIMA-FIGARCH (Rachev et al. [2006])and our autoregressive regime-switching model. In comparison with other models, autoregressive regime-switching model has better backtesting results for forecasting VaR models with high frequency financial time series.

# Part I

# Regime-switching stochastic volatility model

# Chapter 2

# Discrete-time Markov driven stochastic volatility model

The model extends the binomial tree model to a regime-switching volatility model in a two-state setting for volatility. It more accurately reflects the true state dependent nature of the volatility in financial markets. This model also reduces the problem to a computationally tractable form, which can be generalized to American and other forms of path-dependent options.

## 2.1 One-step model

The stock price movement can be illustrated in Figure (2.1.1). The stock price can either move up from $S_t$ to a new level $S_t u$, or down to $S_t d$. The $u$ and $d$ are functions of $\sigma_t$. The probability of an ”up” movement is denoted by $q$. The probability of a ”down” movement is $1 - q$.

The volatility follows a two-state Markov chain. The volatility is either in a high state

$$\left(\begin{array}{c|cc} & \sigma_h & \sigma_l \\ \hline \sigma_h & p_h & 1-p_h \\ \sigma_l & 1-p_l & p_l \end{array}\right)$$

Table 2.1: Markov transition matrix for the volatilities.

or low state: $\sigma(t) \in \{\sigma_h, \sigma_l\}$. The transition between two states is driven by a homogeneous transition matrix as Table (2.1). The transition matrix is called homogeneous as it remains invariant as time changes. The probability of an "up" movement from $S_t$ to $S_t u_h$ with high volatility is denoted by $q_h$. The probability of "up" movement from $S_t$ to $S_t u_l$ with low volatility is denoted by $q_l$. The payoff from an option is $H_h^u$ on the condition that stock price is $S_t u_h$. We also have $H_h^d$, $H_l^u$ and $H_l^d$ for $S_t d_h$, $S_t u_l$ and $S_t d_l$, respectively.

If $\sigma_t = \sigma_h$,

$$u_h = e^{\sigma_h \sqrt{\Delta t}}$$

$$d_h = \frac{1}{u_h} = e^{-\sigma_h \sqrt{\Delta t}}.$$

If $\sigma_t = \sigma_l$,

$$u_l = e^{\sigma_l \sqrt{\Delta t}}$$

$$d_l = \frac{1}{u_l} = e^{-\sigma_l \sqrt{\Delta t}}.$$

At time $t+1$,

$$S_{t+1} \in \{S_t u_h, S_t d_h, S_t u_l, S_t d_l\}.$$

Suppose the payoff from the option is

$$\vec{H}_{t+1} = \left(H_h^u, H_h^d, H_l^u, H_l^d\right),$$

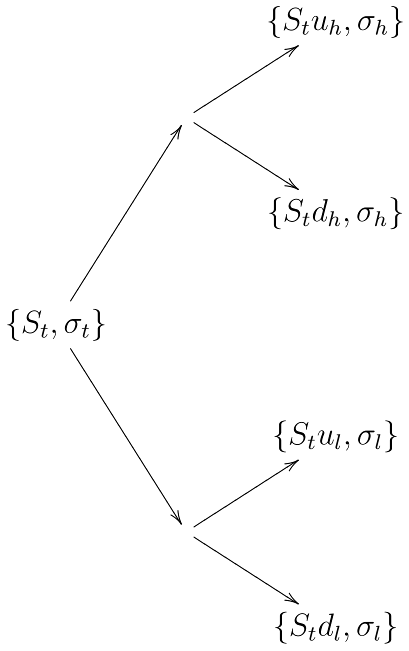


Figure 2.1.1: Illustration of stock price movement in one step.

and the payoff from option at time $t$ is denoted by a vector

$$\vec{H}_t = \left(H_h, H_l\right).$$

A method is proposed by Aingworth et al. [2006] to price the option value:

$$\vec{H}_t = e^{-r\Delta t}(W \vec{H}_{t+1}^T),$$

where the matrix $W$ is presented in Table (2.2). The $q_h$ and $q_l$ are defined as

$$q_h = \frac{e^{r\Delta t} - d_h}{u_h - d_h}$$

$$W = \left(\begin{array}{c|cccc} \sigma_t \diagdown S_{t+1} & S_t u_h & S_t d_h & S_t u_l & S_t d_l \\ \hline \sigma_h & p_h q_h & p_h(1-q_h) & (1-p_h)q_l & (1-p_h)(1-q_l) \\ \sigma_l & (1-p_l)q_h & (1-p_l)(1-q_h) & p_l q_l & p_l(1-q_l) \end{array}\right)$$

Table 2.2: The weights matrix of the model

and

$$q_l = \frac{e^{r\Delta t} - d_l}{u_l - d_l}.$$

## 2.2 Multiple-step model

It is important to notice that a node is not only determined by stock price, but also by the state of volatility. In other words, each node is dependent on two states, stock price and volatility. Thus the nodes are described as $\{(n_1, n_2, n_3, n_4), \sigma_t\}$.

We code the states of nodes as numbers. We start with coding $u_h$, $d_h$, $u_l$, $d_l$ by

$$S(1,0,0,0),\ S(0,1,0,0),\ S(0,0,1,0),\ S(0,0,0,1).$$

Then we can extend this notation to nodes at any time, for example: $u_h^2$, $u_h d_h$ are coded by $S(2,0,0,0)$, $S(1,1,0,0)$.

Figure (2.2.1) shows the stock price development constitutes a quite complicated lattice. We solve this lattice in a bottom-up fashion. We begin by creating the leaf nodes of the lattice and determining their values. We concurrently compute the moments and the European option values. We then iterate over each level of the lattice, and for each node, we push the probability weighted values to the parent node. When we arrive at the root, we have solutions for all possible initial values.

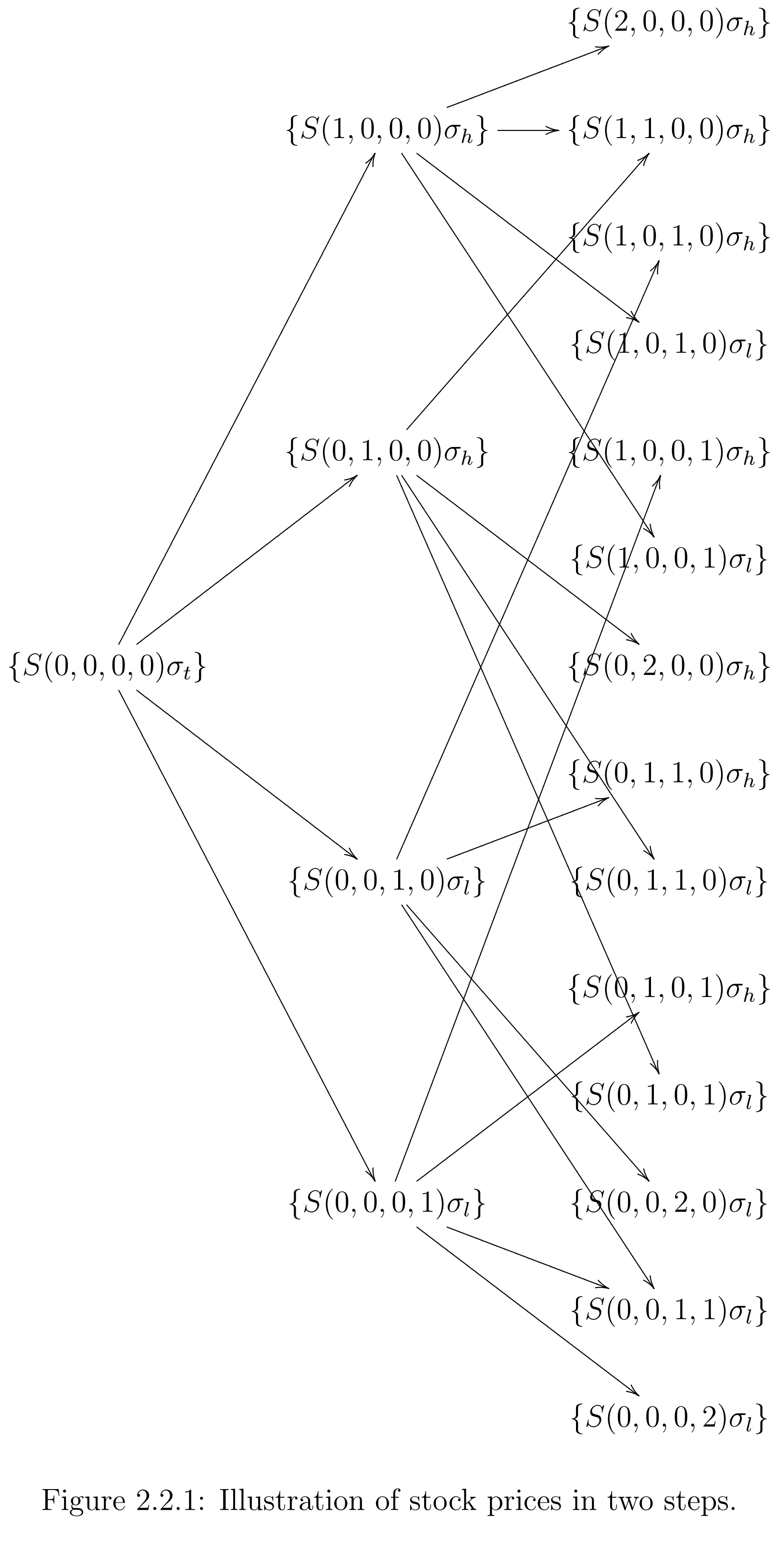


Figure 2.2.1: Illustration of stock prices in two steps.

# Chapter 3

# Mean-variance hedging

## 3.1 Dynamic programming method for mean-variance hedging

With four possible return values and only two assets to hedge with, our model as described above is incomplete. We want to choose some hedging strategy involving hedging error, which means risk. The risks and returns are measured with a utility function. We chose the simplest utility function, a quadratic function.

The *cumulative discount process* is defined as

$$S_t^0 := R^t = (1+r)^t,$$

where $r$ is the risk-free rate.

The *discounted gain process* of the basis asset $X$ is also defined as:

$$\begin{aligned}\Delta X_t &:= X_t - X_{t-1} \\ &= \frac{S_t}{S_t^0} - \frac{S_{t-1}}{S_{t-1}^0} \\ &= \frac{S_t}{R^t} - \frac{S_{t-1}}{R^{t-1}}.\end{aligned}$$

Here $V^{x,\theta}$ is the wealth gained by a self-financing strategy with an initial endowment $x$ and with shares of a risky investment $\theta = \theta_{t=0,\ldots,T-1}$.

$$V_t^{x,\theta} = V_{t-1}^{x,\theta} R + \theta_{t-1} \Delta X_t S_t^0$$

$$\begin{aligned}\frac{V_t^{x,\theta}}{S_t^0} &= \frac{V_{t-1}^{x,\theta}}{S_{t-1}^0} + \theta_{t-1} \Delta X_t \\ &= x + \sum_{i=0}^{t-1} \theta_i \Delta X_{i+1}.\end{aligned}$$

**Definition 1** (Exogenous)**.** A random variable $X : \Omega \to R$ is called exogenous if for every fixed $\omega \in \Omega$ the value $X(\omega)$ does not depend on the choice of $V_0^{x,\theta}(\omega), \theta_0(\omega), ..., \theta_{T-1}(\omega)$.

**Definition 2** (Measurable)**.** A random variable $X : \Omega \to R$ is called $F_T$ measurable if the value of $X(\omega)$ only depends on the filtration $F_T$.

**Mean Square Hedging** Given an exogenous and $F_T$ measurable payoff $H_T$, the best mean-square hedge for $H_T$ is given by the initial wealth x and portfolio weights $\theta$ which are found by minimizing the expected square replication error

$$\min_{x,\theta_0,\cdots,\theta_{T-1}} \mathbb{E}_0^P[V_T^{x,\theta} - H_T]^2.$$

**The Difficulty of Explicit Computation** It is very difficult to directly compute the mean variance hedging problem. An example will be added.

**Definition 3** (Optimal Value Process)**.** Optimal value process $U_t(x), t = 0, ..., T$. The $U_t(x)$ equals the minimum expected squared replication error at maturity, given it is now time $t$, the time $t$ wealth is $x$ and the time $t$ history is $\mathscr{F}_t$. Hence $U_t(x)$ will be an $\mathscr{F}_t$ measurable random variable.

- when $t = T$, $U_t(x)$ coincides with the squared replication error, that is, $U_T(x) = (x - H_T)^2$.
- when $t < T$, the value of $U_t(x)$ satisfies the important dynamic programming functional equation:

$$U_t(x) = \min_{\theta} \mathbb{E}[U_{t+1}(S_{t+1}^0(x/S_t^0 + \theta_t \Delta X_t))].$$

**A Dynamic Programming Solution** Cernỳ [2004] presents a practical dynamic programming solution.

**Theorem 3.1.1.** *Let $k_t$ and $H_t$ be $F_t$ measurable and exogenous. The problem*

$$\min_{x,\theta_0,\cdots,\theta_{t-1}} \mathbb{E}_0^P[k_t(V_t^{x,\theta} - H_t)]^2$$

*has the same optimal controls $x$, $\theta_0, \dots, \theta_{t-1}$ as the problem*

$$\min_{x,\theta_0,\cdots,\theta_{t-2}} \mathbb{E}_0^P[k_{t-1}(V_{t-1}^{x,\theta} - H_{t-1})]^2.$$

*Here $k_t$ can be interpreted as the ratio between the value of the hedging portfolio and the option price $\frac{V_t^{x,\theta}}{H_t}$. We can calculate $k_t$, $H_t$ backward as*

$$\frac{k_{t-1}}{R^2} = \mathbb{E}_{t-1}^P[k_t] - (\mathbb{E}_{t-1}^P[k_t \Delta X_t])^2 (\mathbb{E}_{t-1}^P[k_t(\Delta X_t)^2])^{-1}, \tag{3.1.1}$$

$$H_{t-1} = \frac{\mathbb{E}_{t-1}^P[\left(k_t - \mathbb{E}_{t-1}^P[k_t \Delta X_t](\mathbb{E}_{t-1}^P[k_t \Delta X_t \Delta X_t])^{-1} k_t \Delta X_t\right) \frac{H_t}{R}]}{\frac{k_{t-1}}{R^2}}. \tag{3.1.2}$$

$$\theta_{t-1}^D = -\left(\mathbb{E}_{t-1}^P[k_t \Delta X_t \Delta X_t]\right)^{-1} \mathbb{E}_{t-1}^P[k_t \Delta X_t \left(\frac{V_{t-1}^{x,\theta}}{S_{t-1}^0} - \frac{H_t}{S_t^0}\right)] \tag{3.1.3}$$

The above theorem in presents a dynamic programming solution to the general mean-variance hedging problem in discrete time. A repeated application of the theorems starting from T with $k_T = 1$ gives us all values of $k_t$ and $H_t$ for $0 \le t \le T$. Further, at the end of the backward run we learn that the optimal value of initial wealth is $\widehat{x} = H_0$. In a forward run from time 0 we can then recover the optimal portfolio and optimal hedging wealth from (3.1.1) and (3.1.2).

## 3.2 Verification for one-step hedging

We are doing one-step least square hedging to see if the result is the same as the dynamic mean variance hedging method.

We assume the initial state is in high volatility, $\sigma_0 = \sigma_H$, then the State matrix $A$ is shown in Table (3.2), the vector of contingent claim $H$ at $t = 1$ is represented as Table (3.3), the objective measure $P$ is stated in Table (3.3).

Consider a hedging problem $A\theta = b$ with replication error $\epsilon = A\theta - b$ as a weighted least square problem. To minimize the expected squared replication error,

$$\min_x \sum_i p_i \epsilon_i^2,$$

compute new matrices $\tilde{A}$ and $\tilde{H}$ by multiplying each row of $A$, $H$ by the square root of the probability for the corresponding state,

$$\tilde{A}_{i\,\cdot} := \sqrt{p_i} A_{i\,\cdot}$$

$$\tilde{H}_i := \sqrt{p_i} H_i.$$

The problem is transformed to

$$\min_x \|\tilde{A}\theta - \tilde{H}\|_2.$$

Solving this least square problem gives

$$\theta = (\tilde{A}^T \tilde{A})^{-1} \tilde{A}^T \tilde{H}, \qquad (3.2.1)$$

which is known as optimal hedging portfolio.

We set

$$\{\sigma_H, \sigma_L, p_H, p_L, \Delta t, r, S_0, K\} = \{0.8, 0.2, 0.8, 0.9, 1, 0.01, 100, 110\},$$

$$k_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Table 3.1: The state matrix of the underlier assets.

$$A = \begin{pmatrix} u_H S_0 & 1+r \\ \frac{1}{u_H} S_0 & 1+r \\ u_L S_0 & 1+r \\ \frac{1}{u_L} S_0 & 1+r \end{pmatrix}$$

Table 3.2: The state matrix of the underlier assets.

and from equation (3.2.1), we can get the one step hedging result

$$\theta = \{0.62882, -33.5991\}$$

$$< \theta_0, \{S_0, 1\} > = 29.2828.$$

On the other side, we apply formulas in Theorem (3.1.1) with $t = 1$. $k_1$ is represented in matrix (3.1), $P$ in matrix (3.4), $H_1 = H$ in matrix (3.3) and $\Delta X_1$ in matrix (3.5).

$V_0^{x,\theta}$ in equation (3.1.3) is replaced with $H_0$, $\mathbb{E}_0^P[.]$ is computed with the inner product of $P$ and corresponding vector.

We can get the result

$$\theta_0 = 0.62882$$

$$b = H_0 - \theta_0 S_0 = -33.5991$$

$$< \{\theta_0, b\}.\{S_0, 1\} >= 29.2828,$$

which is coincident with the first method.

$$H = \begin{pmatrix} H_{u|H} \\ H_{d|H} \\ H_{u|L} \\ H_{d|L} \end{pmatrix} = \begin{pmatrix} \max(S_0 u_H, K) \\ \max(S_0/u_H, K) \\ \max(S_0 u_L, K) \\ \max(S_0/u_L, K) \end{pmatrix}$$

Table 3.3: The state matrix of contingent claims.

$$P = \begin{pmatrix} p_{u|H} \\ p_{d|H} \\ p_{u|L} \\ p_{d|L} \end{pmatrix} = \begin{pmatrix} q_H p_H \\ (1-q_H)p_H \\ q_L(1-p_H) \\ (1-q_L)(1-p_H) \end{pmatrix}$$

Table 3.4: The matrix of object measure.

$$\Delta X_1 = \begin{pmatrix} S_0 u_H \frac{1}{1+r} \\ S_0 \frac{1}{u_H} \frac{1}{1+r} \\ S_0 u_L \frac{1}{1+r} \\ S_0 \frac{1}{u_L} \frac{1}{1+r} \end{pmatrix} - \begin{pmatrix} S_0 \\ S_0 \\ S_0 \\ S_0 \end{pmatrix}$$

Table 3.5: The vector of $\Delta X$

# Chapter 4

# Numerical examples

A wide range of tests have been conducted to explore various aspects of our algorithm. In particular, we would like to study

- complexity with respect to the number of steps,
- the algorithm convergence rate,
- the impact of initial states,
- application to real market data.

## 4.1 Algorithm complexity

The major time-consuming part of our algorithm is searching backwards. The computation cost at level $i$ is $O(\binom{i+3}{3})$. As a result, the total cost estimation is

$$\sum_{i=N-1}^{0} O\left(\binom{i+3}{3}\right) = \sum_{i=N-1}^{0} O\left(\frac{(i+3)(i+2)(i+1)}{6}\right) = O(N^4). \tag{4.1.1}$$

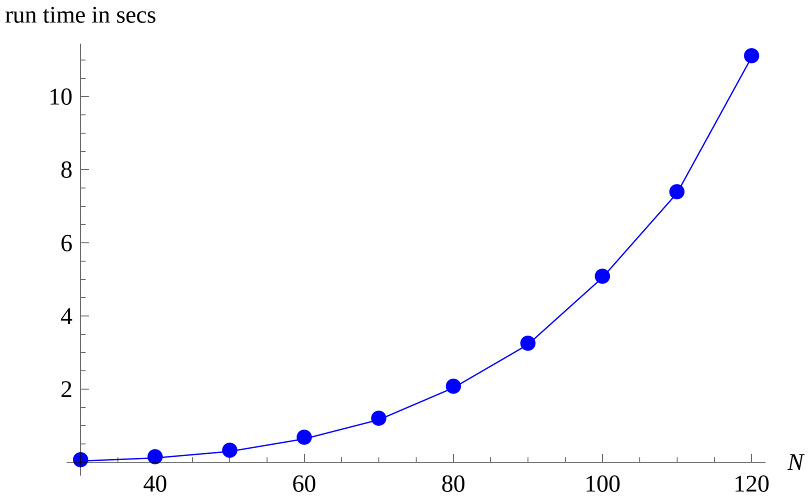


Figure 4.1.1: The running time as a function of number of steps.

To confirm that this is indeed the case, we set the model parameters as: $S_0 = 100, K = 130, T = 5, r = 0.04, N = 30, \cdots, 120, p_h = 0.5, p_l = 0.5, \sigma_h = 0.8, \sigma_l = 0.2$.

Figure (4.1.1) shows how running times (in seconds) change as $N$ increases. All tests are done on an iMac computer with 3G Intel Core 2 Duo processor. Thanks to a Hash-table approach, the pricer can be done within 6 seconds for $N = 100$.

We also plot the logarithm of the running times against $N$ in Figure (4.1.2). Least-Square fit shows a slope of 4.1, which verifies the estimation (4.1.1).

## 4.2 Convergence study

In this example, we increase the number of periods $n$ on the lattice. We examine call option values as $n$ becomes very large in order to assess when call options converge to stable values. Table (4.1) and Figure (4.2.1) show the results.

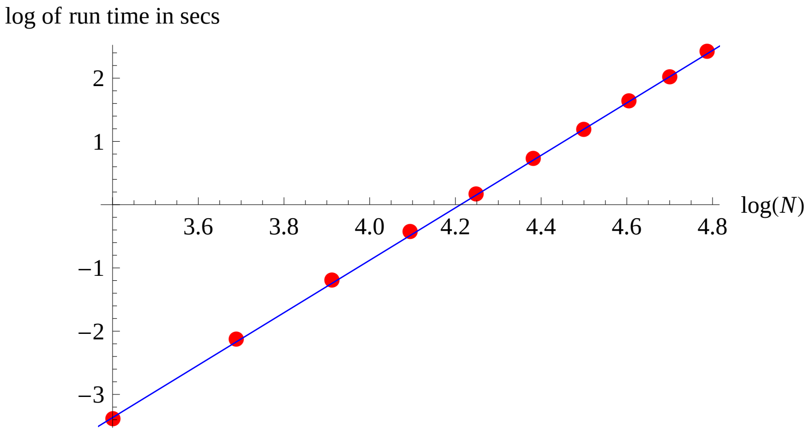


Figure 4.1.2: The log-log plot of the runtime as a function of number of steps and the least square fit.

We set the parameters as:

$$\begin{aligned} S_0 &= 100, \\ K &= 130, \\ T &= 5, \\ r &= 0.04, \\ \{\sigma_h, \sigma_l\} &= \{0.07, 0.03\}, \\ \{p_h, p_l\} &= \{0.9, 0.8\}, \\ N &= 30, \cdots, 120. \end{aligned}$$

Call option values at two different starting states are denoted by $C_h$ and $C_l$.

## 4.3 Impact of initial states

This numerical test is to examine how the model behaves with different levels of disparity between two volatilities. We set the parameters as $S_0 = 100, K = 130, T = 5, r = 0.04, N =$

Table 4.1: The convergence in terms of option values with respect to number of steps.

| $N$ | $C_h$ | $C_l$ |
|---|---|---|
| 30 | 2.701 | 2.471 |
| 40 | 2.742 | 2.572 |
| 50 | 2.767 | 2.633 |
| 60 | 2.783 | 2.673 |
| 70 | 2.794 | 2.700 |
| 80 | 2.804 | 2.721 |
| 90 | 2.811 | 2.738 |
| 100 | 2.816 | 2.750 |
| 110 | 2.821 | 2.761 |
| 120 | 2.824 | 2.770 |

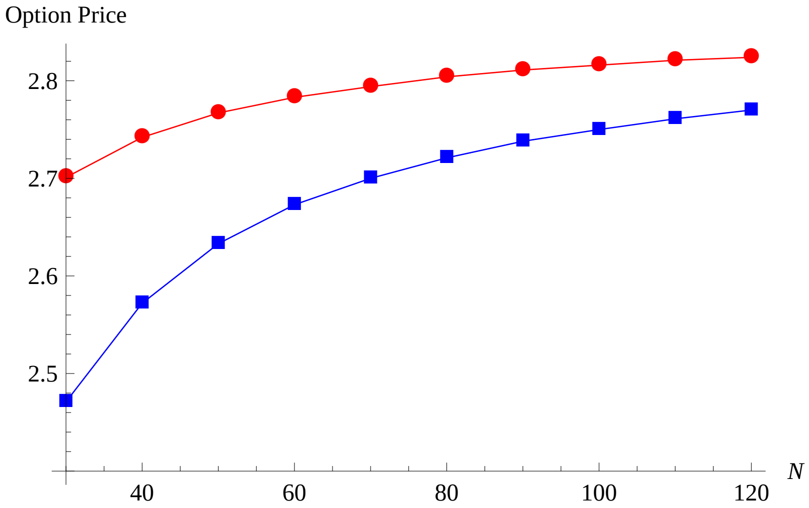


Figure 4.2.1: The convergence in terms of option values with respect to number of steps.

30.

The $C_h$ and $C_l$ still denote call option values at two different starting states.

Table (4.2) shows that with transition matrix fixed, the disparity between $C_h$ and $C_l$ declines as the disparity between two different volatility states declines.

Table 4.2: Impact of the Initial State.

| $(\sigma_h, \sigma_l)$ | $(p_h, p_l)$ | $C_h$ | $C_l$ |
|---|---|---|---|
| (0.070, 0.030) | (0.90, 0.90) | 2.310 | 1.885 |
| | (0.75, 0.75) | 2.185 | 2.080 |
| | (0.60, 0.60) | 2.156 | 2.129 |
| (0.065, 0.035) | (0.90, 0.90) | 2.140 | 1.818 |
| | (0.75, 0.75) | 2.041 | 1.961 |
| | (0.60, 0.60) | 2.017 | 1.997 |
| (0.060, 0.040) | (0.90, 0.90) | 2.005 | 1.785 |
| | (0.75, 0.75) | 1.932 | 1.877 |
| | (0.60, 0.60) | 1.914 | 1.901 |
| (0.055, 0.045) | (0.90, 0.90) | 1.897 | 1.788 |
| | (0.75, 0.75) | 1.864 | 1.837 |
| | (0.60, 0.60) | 1.857 | 1.850 |
| (0.050, 0.050) | (0.90, 0.90) | 1.849 | 1.849 |
| | (0.75, 0.75) | 1.849 | 1.849 |
| | (0.60, 0.60) | 1.849 | 1.849 |

## 4.4 European call for S&P 500

We choose S&P 500 options as our numerical example because of its well-known negative volatility skew. The spot value is 1095.95 on June 3, 2010. Call options used for analysis matured on June 18th, 2011. The LIBOR rate for the same maturity is approximated to be 0.0120406. The two volatility states and transition matrix are the scaled results of calibration:

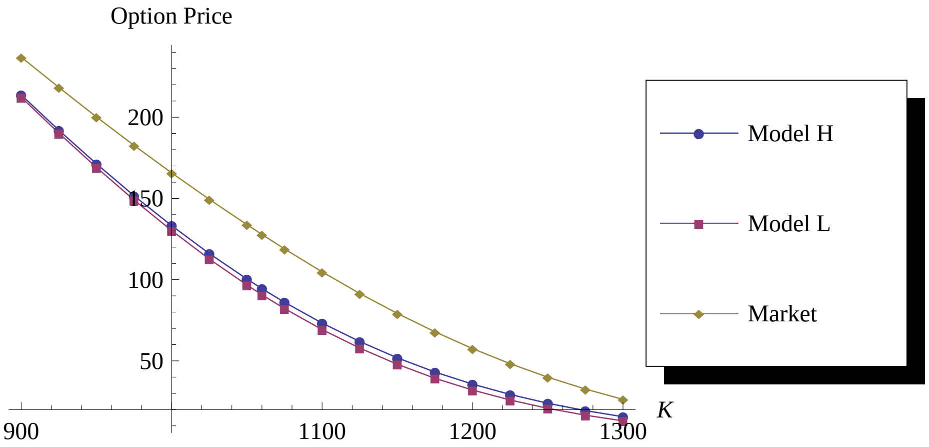


Figure 4.4.1: Option price vs. strike price for S&P500: a) model prices with starting states at high volatility; b) model prices with starting states at low volatility; c) market prices.

$$\{\sigma_h, \sigma_l\} = \{0.240145, 0.121742\},$$

$$\{p_h, p_l\} = \{0.811563, 0.956289\}.$$

The model is fitted to the monthly log return of S&P500 price from December 1, 1968 to April 5, 2010 with EM algorithm.

Even though Figure (4.4.1) shows the model can predict the market option price, the failure to generate volatility skew in figure (4.4.2) suggests deeper investigations into our model are needed.

## 4.5 Examination of parameter sets

After fitting the model to the historical data, it fails to produce a volatility smile. Is this failure due to the bad calibration of parameters, or the model itself? We try out many possible combinations of parameters in reasonable ranges to see if the smile can be generated

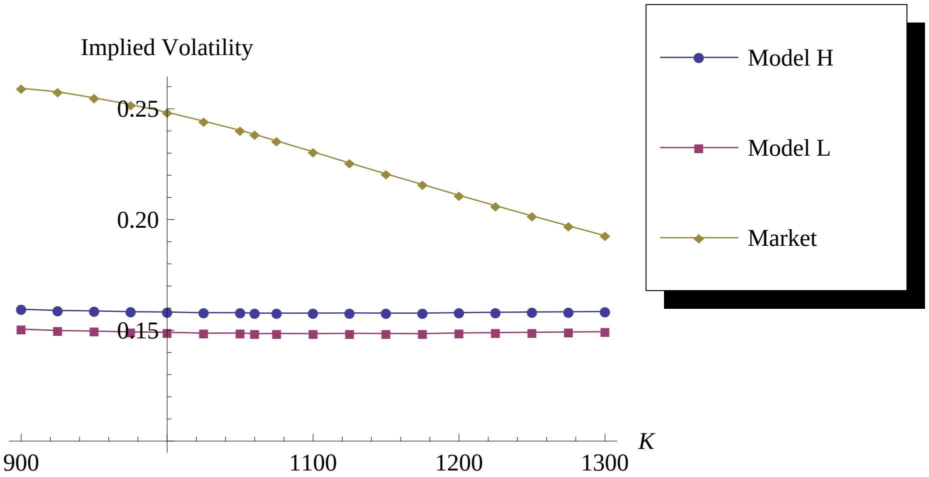


Figure 4.4.2: Implied volatility skew for S&P500: a) model implied volatilities with starting states at high volatility; b) model implied volatilities with starting states at low volatility;c) market implied volatilities.

under some parameter settings.

The input for model is two states of volatility: $\{\sigma_h, \sigma_l\}$, transition matrix: $M = \begin{pmatrix} p_h & 1-p_h \\ 1-p_l & p_l \end{pmatrix}$, stock price: $S_0$, strike price: $K$, interest rate: $r$, period: $T$.

We assume $S_0 = 100$, $r = 0$, $T = 1$, $m = K/S_0$ which is called moneyness. Then we loop over other five parameters as:

$$\sigma_l = \{0.1, \ldots, 0.5\}$$

$$\sigma_h = \{\sigma_l + \Delta, \ldots, \sigma_l + 10\Delta\}$$

$$p_h = \{0.1, \ldots, 1.0\}$$

$$p_l = \{0.1, \ldots, 1.0\}$$

$$m = \{0.5, \ldots, 1.5\},$$

in which

$$\Delta = \frac{1-\sigma_l}{5}$$

$\sigma^i_{max}$ is the highest implied volatility with different moneyness; $\sigma^i_{min}$ is the lowest implied volatility with different moneyness. We define the ratio as $\frac{\sigma^i_{max}}{\sigma^i_{min}}$, which can quantify the deepness of the volatility smile. We cut $(1.05, 2.0)$ into 8 intervals, add a special interval $(-1.0, 1.05)$ to represent two cases: bad case failing to generate implied volatility and those with ratios almost equal to 1, which represents a flat line. After filling in all the ratios into these intervals, we get a histogram-like Table (4.3):

Table 4.3: Histogram of ratios: ratio = $\frac{\sigma^i_{max}}{\sigma^i_{min}}$.

| cut | frequency |
|---|---|
| (-1.0 - 1.05) | 1451 |
| (1.05 - 1.1) | 96 |
| (1.1 - 1.2) | 74 |
| (1.2 - 1.3) | 42 |
| (1.3 - 2.0) | 21 |
| (1.4 - 1.5) | 15 |
| (1.5 - 1.6) | 29 |
| (1.6 - 1.7) | 42 |
| (1.8 - 2.0) | 10 |
| (2.0 - $\infty$) | 20 |

We pick up all the cases with ratios larger than 1.4, there is an outstanding characteristic for them:

$$p_l = 1,$$

which means state $L$ is an observing "black hole". Whenever the volatility state jumps to low state, it will stay there. We presents some of the cases with this feature as in table (4.4).

Exclude those "black hole" cases, other cases are almost with ratios around 1. In other words, normal cases don't have volatility smile.

Table 4.4: Part of the cases with "black hole" feature.

| $\sigma_l$ | $\sigma_h$ | $p_l$ | $p_h$ | ratio |
|---|---|---|---|---|
| 0.1 | 1.0 | 1.0 | 0.3 | 2.004710 |
| 0.1 | 1.0 | 1.0 | 0.4 | 2.13712 |
| 0.1 | 1.0 | 1.0 | 0.5 | 2.232717 |
| 0.1 | 1.0 | 1.0 | 0.6 | 2.718023 |
| 0.1 | 1.0 | 1.0 | 0.7 | 2.794309 |
| 0.1 | 1.0 | 1.0 | 0.8 | 2.838305 |
| 0.1 | 1.0 | 1.0 | 0.9 | 2.796462 |

## 4.6 Richardson Extrapolation

Does the failure of generating volatility skew stem from a lack of computation accuracy? We explore this possibility by using Richardson extrapolation to improve the computation accuracy.

Let $A(0)$ be the true value and $A(h)$ be some approximation which satisfies

$$\lim_{h\to 0} A(h) = A(0).$$

Assume the error term has the format as

$$A(0) - A(h) = a_n h^n + O(h^m),$$

where $a_n$is a nonzero constant independent of $h$ and $m > n$.

For simplicity, we ignore the higher order error term. Given $A(h)$, $A(\frac{h}{2})$ and $n$, we have two equations

$$A(0) - A(h) = a_n h^n$$
$$A(0) - A\left(\frac{h}{2}\right) = a_n \left(\frac{h}{2}\right)^2 \tag{4.6.1}$$

with two unknowns $a_n$, $A(0)$. Solving these two equations yields the approximation

$$\tilde{A}(0) = A\left(\frac{h}{2}\right) + \frac{A\left(\frac{h}{2}\right) - A(h)}{2^n - 1}.$$

In many cases we don't know the order of convergence $n$. We propose the following estimation method.

Based on (4.6.1), we have

$$\log E(h) = n \log h + \log a_n,$$

where the error term $E(h) := A(0) - A(h)$.

For $\sigma_h = \sigma_l$, we simplify our model into the standard binomial tree case and the true value $A(0)$ can be calculated via the famous Black-Scholes formula. Given a series of time steps $\{h_i\}$, we may compute the corresponding error terms $\{E_i(h_i)\}$. As a result, we can estimate the order of convergence $n$ via least square fit.

We set up our testing cases as: spot = 100, expiry = 1, risk-free rate = 0.01, number of steps = 50 and 100 respectively, time step $\Delta t = 1/50$, order of convergence $n = 1$.

The last two columns in Table (4.5) show the implied volatlity is constant with respect to different strike. We tested various combinations of volatilities and transition probabilities, all of which show a "flat" volatility smile. We present a single case in the table (4.5).

Table 4.5: The implied volatilities using Richardson extrapolation: $A_h$ and $A_l$ refer to the cases where the starting states are $\sigma_h$ and $\sigma_l$ respectively.

| strike | $A_h(\Delta t)$ | $A_l(\Delta t)$ | $A_h(\frac{\Delta t}{2})$ | $A_l(\frac{\Delta t}{2})$ | $\tilde{A}_h(0)$ | $\tilde{A}_l(0)$ | $\sigma_h^i$ | $\sigma_l^i$ |
|---|---|---|---|---|---|---|---|---|
| 50 | 51.89 | 51.63 | 51.77 | 51.63 | 51.65 | 51.64 | 0.49 | 0.49 |
| 60 | 43.66 | 43.19 | 43.49 | 43.26 | 43.33 | 43.32 | 0.49 | 0.49 |
| 70 | 36.35 | 35.66 | 36.15 | 35.81 | 35.96 | 35.96 | 0.49 | 0.49 |
| 80 | 30.01 | 29.12 | 29.81 | 29.37 | 29.61 | 29.61 | 0.49 | 0.49 |
| 90 | 24.65 | 23.61 | 24.44 | 23.93 | 24.23 | 24.24 | 0.49 | 0.49 |
| 100 | 20.18 | 19.04 | 19.96 | 19.40 | 19.74 | 19.76 | 0.49 | 0.49 |
| 110 | 16.52 | 15.34 | 16.28 | 15.70 | 16.05 | 16.06 | 0.49 | 0.49 |
| 120 | 13.52 | 12.36 | 13.27 | 12.69 | 13.02 | 13.03 | 0.49 | 0.49 |
| 130 | 11.08 | 9.97 | 10.82 | 10.27 | 10.56 | 10.56 | 0.49 | 0.49 |
| 140 | 9.10 | 8.07 | 8.84 | 8.32 | 8.57 | 8.56 | 0.49 | 0.49 |
| 150 | 7.51 | 6.55 | 7.23 | 6.75 | 6.96 | 6.94 | 0.49 | 0.49 |

## 4.7 Convergence of HMM driven stochastic volatility model

In this section, we show that modifying the transition matrix into a reasonable $\Delta t$-dependent way is a potential cure for the model's failure to produce volatility smile.

### 4.7.1 Convergence of the process of volatility into an i.i.d. process

We show the Markov-driven stochastic volatility process can converge into an i.i.d. process with the help of Perron Frobenius theorem and its application on Markov matrix.

Assume the initial state is $\mathbf{x^0}$, the process evolves recursively by the rule $\mathbf{x^{k+1}} = \mathbf{Ax^k}$, or in short, $\mathbf{x^k} = \mathbf{A^kx^0}$ (MacCluer [2000]). When will such a process converge? We denote the matrix of eigenvectors of $A$ as $V$, the diagonal matrix of eigenvalues of $A$ as $\Lambda =$

$\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$, where $\lambda_{1\ldots}\lambda_n$are eigenvalues of $A$. As a result,

$$A = V\Lambda V^{-1} \Rightarrow A^k = V\Lambda^k V^{-1} = V \begin{pmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{pmatrix} V^{-1}.$$

This process will converge only if the spectral radius $r(A) \leqslant 1$, i.e., the largest absolute value of eigenvalues. If $r(A) = 1$, $A^k x^0$ converges to a nonzero value; if $r(A) < 1$, converges to 0. Perron Frobenius Theorem tells us that the Markov driven volatility process in our model will converge. In our model,

$$A = \begin{pmatrix} p_H & 1 - p_H \\ 1 - p_L & p_L \end{pmatrix}$$

and

$$\mathbf{x} = \begin{pmatrix} x_H \\ x_L \end{pmatrix}.$$

The distribution at the first step is

$$\begin{pmatrix} x_H, & x_L \end{pmatrix} \begin{pmatrix} p_H & 1 - p_H \\ 1 - p_L & p_L \end{pmatrix} = \begin{pmatrix} p_L x_L + (1 - p_H)\, x_H, & p_H x_H + (1 - p_L)\, x_L \end{pmatrix}$$

We can use Perron Frobenius theorem to show the dominant eigenvalue $\lambda_1 = 1$(Dym [2007]), then compute the other eigenvalue

$$\lambda_2 = p_L p_H - (1-p_L)(1-p_H).$$

The distribution at step $k$ is

$$\begin{aligned} &\begin{pmatrix} x_H & x_L \end{pmatrix} \begin{pmatrix} p_H & 1-p_H \\ 1-p_L & p_L \end{pmatrix}^k \\ =&\lambda_2^k \begin{pmatrix} x_H, & x_L \end{pmatrix} (Ve_2)(e_2^T V^{-1}) + \begin{pmatrix} x_H, & x_L \end{pmatrix} (Ve_1)(e_1^T V^{-1}) \end{aligned}$$

It is readily to check that the eigenvector corresponding to eigenvalue 1 is $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, in other words, $Ve_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Substituting $Ve_1$ into equation,

$$\begin{pmatrix} x_H & x_L \end{pmatrix} \begin{pmatrix} p_H & 1-p_H \\ 1-p_L & p_L \end{pmatrix}^k = (e_1^T V^{-1}) + \lambda_2^k \begin{pmatrix} x_H & x_L \end{pmatrix} (Ve_2)(e_2^T V^{-1}).$$

Since $\lambda_2 < 1$, the second term goes to zero quickly. Therefore, the process reaches the equilibrium distribution of volatility $e_1^T V^{-1}$.

We can also see the second eigenvalue $\lambda_2$ decides the speed of the convergence. For instance, $A = \begin{pmatrix} p & 1-p \\ p & 1-p \end{pmatrix}$ can converge immediately since the second eigenvalue is 0.

**Example 1.** $A = \begin{pmatrix} \frac{9}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{9}{10} \end{pmatrix}$ Estimate how many steps it takes to reach the equilibrium state.

$\lambda_1 = 1$, $\lambda_2 = \frac{9}{10} + \frac{9}{10} - \lambda_1 = \frac{4}{5}$. As $(\frac{4}{5})^{30} \approx \frac{1}{1000}$, the Markov process can converge in 30 steps.

Thus, it is easy to work out the number of lattice steps before the matrix can converge, i. e., the number of steps before the Markov driven volatility process turns into an i.i.d. process.

### 4.7.2 Geometric brownian motion with constant volatility

After showing the process converges to an i.i.d. process, we continue to point out that after the convergence, the process is actually a geometric brownian motion with a constant volatility which is a weighted sum of two original volatilities. In other words, we justify that our model won't produce volatility smile after the convergence.

**From the view of mathematical deduction**

Assume the stock price dynamics follows

$$S_{i+1} = S_i \exp^{(r - \frac{\sigma_i^2}{2})\Delta t + \sigma_i Z}$$

where

- $Z = \sqrt{\Delta t}\epsilon$
- $\epsilon$ has standard normal distribution $\mathbb{N}(0, 1)$
- $S_i$ = stock price at step $i$

- $\sigma_i$ = volatility of stock price at interval $i$
- $\Delta t = \frac{T}{N}$ length of time interval
- $r$ = riskfree rate
- $N$ = number of time intervals

If there are two states of volatility $\sigma_i \in \{\sigma_H, \sigma_L\}$, can we simplifie the dynamics of stock price into a process with constant volatility?

If the volatility is constant $\sigma_i = \bar{\sigma}$ for any $i$,

$$\begin{aligned} S_T =& S_0 \exp^{\sum_i (r - \frac{\sigma_i^2}{2})\Delta t + \sigma_i \sqrt{\Delta t}\epsilon} \\ =& S_0 \exp^{(r - \frac{\bar{\sigma}^2}{2}) N \Delta t + \bar{\sigma}\sqrt{T}\epsilon} \\ =& S_0 \exp^{(r - \frac{\bar{\sigma}^2}{2}) T + \bar{\sigma}\sqrt{T}\epsilon} \end{aligned}$$

What is a reasonable assumption for $\bar{\sigma}$? Let us regard the $\sigma_i Z$ process as a finite mixture of two Gaussian processes,

$$\begin{aligned} \sigma_H Z \sim& \mathbb{N}(0, \sigma_H^2 \Delta t) \\ \sigma_L Z \sim& \mathbb{N}(0, \sigma_L^2 \Delta t). \end{aligned}$$

From the linearity of Gaussian distribution,

$$\omega_H \sigma_H Z + \omega_L \sigma_L Z \backsim \mathbb{N}(0, \omega_H \sigma_H^2 \Delta t + \omega_L \sigma_L^2 \Delta t),$$

in other words, the mixture has a constant volatility $\omega_H\sigma_H^2 + \omega_L\sigma_L^2$. Therefore we would like to write

$$\bar{\sigma} = \sqrt{\omega_H\sigma_H^2 + \omega_L\sigma_L^2}.$$

### From the View of Simulation

The Monte Carlo simulation of a finite mixture of $\sigma_H Z$ and $\sigma_L Z$ with weights $\omega_H$,$\omega_L$ (Figure 4.7.1)shows results very close to Black-Scholes formula with constant volatility $\bar{\sigma}$.

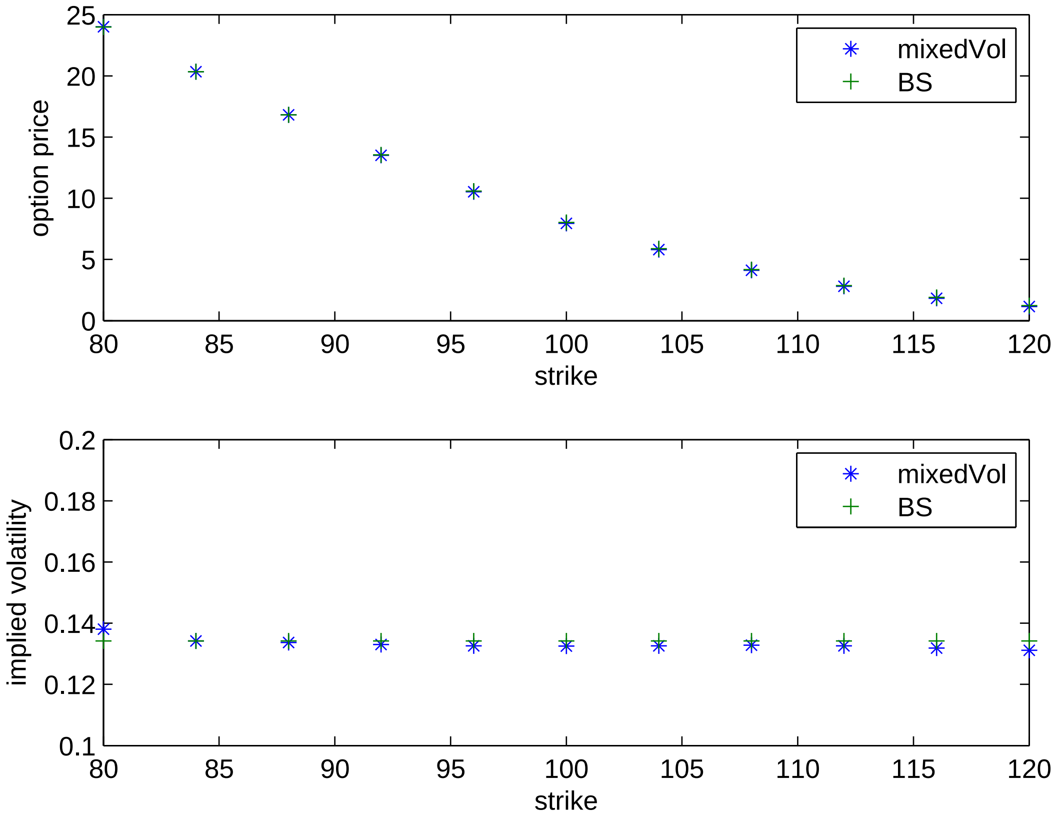


Figure 4.7.1: Comparison of Monte Carlo simulation of a finite mixture of $\sigma_H Z$ and $\sigma_L Z$ with weights $\omega_H$,$\omega_L$ with Black-Scholes formula with constant volatility $\bar{\sigma}$.

# Part II

# Autoregressive regime-switching model

# Chapter 5

# Autoregressive model with Gaussian innovation

Assume an observation window of length $K$ moves along the time series data with overlapping length $M$ (Figure (5.0.1)). In our model, $M = 1$, which is a natural case, and length $K = 5$. To be more specific, consider the observation vector $\vec{s}$ with components $(x_0, x_1, \ldots, x_{K-1})$.

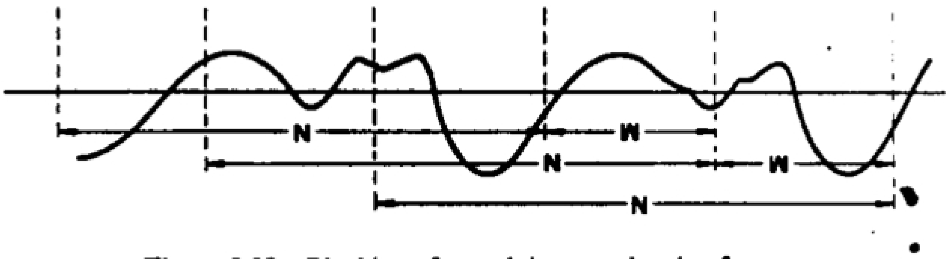


Figure 5.0.1: Illustration of overlapped observation window.

$$\begin{aligned} x_n &= -\sum_{i=1}^{p} a_i(\omega) x_{n-i} + e_n \qquad n = 0, 1, 2, \ldots, K-1 \\ s_n &= x_n \sigma(\omega) \qquad \omega = \{1, \ldots, N\} \end{aligned} \tag{5.0.1}$$

where $N$ is number of states, $e_k$ are i.i.d. Gaussian random variables with mean 0 and variance 1 , and $p$ is order of autoregression. Autoregressive coefficients $a_i$ and variance $\sigma$ follows discrete Markov process. For example, if this is a two state Markov process and order of autoregression $p = 2$, then $a_1 \in \{a_1(1), a_1(2)\}$, $a_2 \in \{a_2(1), a_2(2)\}$, $\sigma \in \{\sigma(1), \sigma(2)\}$. This process is driven by transition matrix $\left(\begin{smallmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{smallmatrix}\right)$.

Section of an autoregressive hidden Markov model, in which the distribution of the observation $\mathbf{x}_n$ depends on a subset of the previous observations as well as on the hidden state $\mathbf{z}_n$. In this example, the distribution of $\mathbf{x}_n$ depends on the two previous observations $\mathbf{x}_{n-1}$ and $\mathbf{x}_{n-2}$.

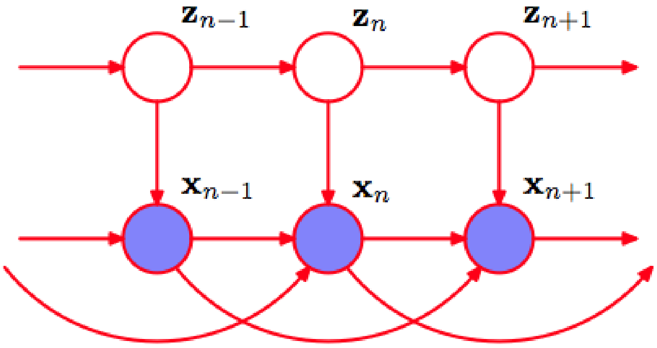


Figure 5.0.2: Illustration of autoregressive hidden Markov model.

## 5.1 Analysis of existing density function of Gaussian autoregressive source

Rabiner [1990] introduces density function of Gaussian autoregressive source in the section of autoregressive HMMS. The density function for $\vec{s}$ is

$$f(\vec{s}) = (2\pi\sigma^2)^{-\frac{K}{2}} \exp\left(-\frac{1}{2\sigma^2}\delta(\vec{s}, a)\right), \tag{5.1.1}$$

where

$$\begin{aligned}
\delta(\vec{s}, a) &= r_a(0)r(0) + 2\sum_{i=1}^{p} r_a(i)r(i)\\
a' &= [1, a_1, \ldots, a_p]\\
r_a(i) &= \sum_{n=0}^{p-i} a_n a_{n+i}\\
r(i) &= \sum_{n=0}^{K-i-1} x_n x_{n+i}.
\end{aligned}$$

After mathematical analysis and numerical tests (See section (5.2.1) for details), we discovered some approximation steps were taken in this density function which is not mentioned in Rabiner [1990]. The assumption for this approximation is that the correlation of the first $K$ observation doesn't affect the whole output. In other words, sample size $T$ is much larger than observation window size $K$. In our model, observation window $K = 5$ and the length of sample size is less than two weeks $T < 10$, thus the assumption doesn't hold for our model. We deduce a new density function for autoregressive process in following sections.

## 5.2 New density function with no structure imposed on first $p$ sample data

Using the Gram-Schmidt process to orthogonalize the first $p$ samples $\{x_1 \dots x_p\}$ into $\vec{\epsilon} = \{\epsilon_1 \dots \epsilon_p\}$, we may rewrite (5.0.1) as

$$H\vec{x} = \vec{e},$$

where $\vec{x} = \{x_1, \cdots, x_K\}$, $\vec{e} = \{\epsilon_1, \cdots, \epsilon_p, e_1, \cdots, e_{K-p}\}$, $\epsilon \sim \mathbf{N}(0,1)$ and

$$H = \left(\begin{array}{cccc|cc} h_{11} & & & & & \\ h_{21} & h_{22} & & & & \\ \vdots & & & & & \\ h_{p1} & h_{p2} & \dots & h_{pp} & & \\ \hline a_p & a_{p-1} & \dots & a_1 & 1 & 0 \\ 0 & a_p & \dots & & a_1 & 1 \end{array}\right) = \left(\begin{array}{c|c} H_{11} & 0 \\ \hline H_{21} & H_{22} \end{array}\right). \tag{5.2.1}$$

To orthogonalize $\vec{x}$, without loss of generality, we assume $x_1 \sim \mathbf{N}(0, \sigma_1)$, and $x_2 \sim \mathbf{N}(0, \sigma_2)$. Since

$$\begin{aligned} &h_{11}x_1 = \epsilon_1 \\ &h_{21}x_1 + h_{22}x_2 = \epsilon_2, \end{aligned} \tag{5.2.2}$$

and $\text{var}(\epsilon_1) = 1$, $\text{var}(\epsilon_2) = 1$, $\text{cov}(\epsilon_1, \epsilon_2) = 0$, we have equations

$$\begin{aligned} &\text{Var}(h_{11}x_1) = 1 \\ &\text{Var}(h_{21}x_1 + h_{22}x_2) = 1 \\ &\text{Cov}(h_{11}x_1, h_{21}x_1 + h_{22}x_2) = 0, \end{aligned}$$

which is equivalent to

$$\begin{aligned}
&h_{11} = \frac{1}{\sigma_1} \\
&h_{21}^2\sigma_1^2 + h_{22}^2\Sigma_2^2 + 2h_{21}h_2 2\sigma_{12} = 1 \\
&h_{21}\sigma_1^2 + h_{22}\sigma_{12} = 0.
\end{aligned}$$

The solution for this equation system is

$$\begin{aligned}
h_{11} &= \frac{1}{\sigma_1^2} \\
h_{21} &= \frac{-\sigma_{12}}{\sqrt{\sigma_2^2\sigma_1^4 - \sigma_1^2\sigma_{12}^2}} \\
h_{22} &= \frac{\sigma_1}{\sqrt{\sigma_2^2\sigma_1^2 - \sigma_{12}^2}}.
\end{aligned}$$

The elements of $\vec{e}$ are uncorrelated, thereby giving

$$\begin{aligned}
\mathbf{I} &= \mathbf{E}\{\vec{e}\vec{e}^t\} \\
&= \mathbf{E}\{H\vec{x}\vec{x}^t H^t\} \\
&= H\mathbf{E}\{\vec{x}\vec{x}^t\}H^t \\
&=: H\Sigma_x H^t
\end{aligned} \tag{5.2.3}$$

Equation(5.2.3) gives

$$\Sigma_x^{-1} = H^t H. \tag{5.2.4}$$

Taking the determinant of both sides, equation(5.2.3) leads to

$$|\Sigma_x| = |H|^{-2} = |H_{11}|^{-2}.$$

Since $\vec{x} = H^{-1}\vec{e}$ and $\vec{e}$ is Gaussian white noise , $\vec{x}$ is also multi-Gaussian. Plug $|\Sigma_x|$ and $\Sigma_x^{-1}$ into the multi-variate Gaussian p.d.f. yields the p.d.f. of the autoregressive process

$$(2\pi)^{-K/2}\left|\Sigma_x\right|^{-1/2}\exp\{-\frac{1}{2}x^t\Sigma_x^{-1}x\}. \tag{5.2.5}$$

When unscaled $\vec{s}$ is used

$$\begin{aligned}\mathbf{I} &= \mathbf{E}\{\vec{e}\vec{e}^t\}\\ &= \mathbf{E}\{H\tfrac{\vec{s}}{\sigma}\tfrac{\vec{s}}{\sigma}^t H^t\}.\end{aligned} \tag{5.2.6}$$

Equation(5.2.3) gives

$$\Sigma_s^{-1} = \sigma^{-2}H^tH.$$

Taking determinant of both sides, equation(5.2.6) leads to

$$|\Sigma_s| = \sigma^{2K}\left|H_{11}\right|^{-2}$$

### 5.2.1 When sample size $T \gg$ observation window size $K$

The density function for $\vec{s}$ is approximately

$$f(\vec{s}) = (2\pi\sigma^2)^{-\frac{K}{2}}\exp\left(-\frac{1}{2\sigma^2}\delta(\vec{s},a)\right), \tag{5.2.7}$$

where

$$\begin{aligned}
\delta(\vec{s}, a) &= r_a(0)r(0) + 2\sum_{i=1}^{p} r_a(i)r(i) \\
a' &= [1, a_1, \ldots, a_p] \\
r_a(i) &= \sum_{n=0}^{p-i} a_n a_{n+i} \\
r(i) &= \sum_{n=0}^{K-i-1} x_n x_{n+i}.
\end{aligned} \tag{5.2.8}$$

We present an example with dimension equal to 3 to illustrate the difference between Equation (5.2.7) and Equation (5.2.5). We get $\Sigma_x$ from Equation (5.2.4), then compute the p.d.f function from Equation (5.2.5). This example was computed using Mathematica code HOMEPAGE/Mathematica/pdftest.nb. We only present part of computation results as

```
/*Mathematica code */
input:
H  = {{h11, 0, 0}, {h21, h22, 0}, {a2, a1, 1}};
x = {x1, x2, x3};
x.Transpose[H].H.x // Simplify
output:
a2^2 x1^2 + h11^2 x1^2 + h21^2 x1^2 + 2 h21 h22 x1 x2 + a1^2 x2^2 +
 h22^2 x2^2 + 2 a1 x2 x3 + x3^2 + 2 a2 x1 (a1 x2 + x3).
```

Comparing the expansion of $x^t\Sigma_x^{-1}x$ with Equation (5.2.8), we can see the cross terms of $a_i$ and $x_i$ are missing in Equation (5.2.7). The assumption for this approximation is that the correlation of the first K observations doesn't affect the whole output, which is true when sample size T is much larger than observation window size K.

### 5.2.2 When sample size $T \gg$ observation window size $K$ doesn't hold

Compute accurate form of density function for $\vec{s}$:

- Compute $\Sigma_{\vec{s}} = \mathbf{E}(\vec{s}\vec{s}^T)$ with $\vec{s} = \{s_1, \ldots, s_K\}$
- Use Eigenvalue Decomposition to get

$$B^T \Sigma_{\vec{s}} B = \beta = \begin{pmatrix} \beta_0 & & & \\ & \beta_1 & & \\ & & \cdots & \\ & & & \beta_{K-1} \end{pmatrix}$$

  where $B$ is an upper triangular matrix, the diagonal elements of which are all unity.

- Get the probability density

$$f(x \mid \Sigma_{\vec{s}}) = (2\pi)^{-K/2} (\sigma^2)^{-(K-p)/2} \left( \prod_{i=0}^{p-1} \frac{\beta_i}{\sigma^2} \right)^{-\frac{1}{2}} \exp\{-\vec{s}^t H^t H \vec{s}/(2\sigma^2)\} \qquad (5.2.9)$$

**Numerical example: testing integral of density funciton**

This numerical test is to examine whether

$$\int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} \cdots \int_{x_T=-\infty}^{\infty} f(\vec{x}) == 1$$

holds, where $f(\vec{x})$ are p.d.f. functions in Equation (5.1.1) and (5.2.9). We set two sets of parameters as $a = [1, 0.5, 0.3], p = 2, K = 3, T = 3, \sigma = 0.2$ and $a = [1, 0.8, 0.4], p = 2, K = 3, T = 5, \sigma = 0.1$. The results in Table (5.1) confirm our argument that density function

(5.1.1) doesn't hold when length of observation window $K$ is not significantly smaller than data size $T$. Our new density function (5.2.9) passes the integration test. This example is computed using Matlab code HOMEPAGE/Matlab/density.m.

Table 5.1: Integration of density function of observation $\vec{x}$ over $\{-\infty, \infty\}$.

| | density function (5.1.1) | density function (5.2.9) |
|---|---|---|
| parameter set 1 | $\infty$ | 1 |
| parameter set 2 | $\infty$ | 1 |

## 5.3 New density function with linear autoregressive structure imposed on first $p$ sample data

To tackle the difficulty that the relationship of (5.0.1) doesn't defined for the first $p$ samples, we may introduce some ghost variables and rewrite (5.0.1) as

$$\begin{aligned} x_1 + a_1 x_0 + a_2 x_{-1} &= \epsilon_1 \\ x_2 + a_1 x_1 + a_2 x_0 &= \epsilon_2 \\ x_3 + a_1 x_2 + a_2 x_1 &= \epsilon_3 \\ &\vdots \\ x_K + a_1 x_{K-1} + a_2 x_{K-2} &= \epsilon_K \end{aligned}$$

where $\epsilon_i$ are i.i.d. $\mathbf{N}(0,1)$.

The parametres are $\{a_1, a_2, \ldots, a_p, x_0, x_{-1}, \ldots, x_{-p}, \sigma\}$. $x_0$ and $x_{-1}$ are ghost variables, which can be treated as scalar. We compare the degree of freedom between (5.2.1) and

(5.0.1),

Table 5.2: Comparison of degree of freedom between with and without linear autoregressive structure imposed on first $p$ sample data.

| | Equation (5.2.1) | Equation (5.0.1) |
|---|---|---|
| | $\frac{1+p}{2}p+p+1$ | $2p+1$ |
| $p=2$ | 6 | 5 |
| $p=3$ | 10 | 7 |
| $p=4$ | 15 | 9 |

After normalization, we have

$$x_1 = -a_1 x_0 - a_2 x_{-1} + \epsilon_1 \tag{5.3.1}$$

$$x_2 + a_1 x_1 = -a_2 x_0 + \epsilon_2 \tag{5.3.2}$$

$$x_3 + a_1 x_2 + a_2 x_1 = \epsilon_3 \tag{5.3.3}$$

We denote $\hat{\epsilon}$ as

$$\hat{\epsilon} := \epsilon + \begin{pmatrix} -a_1 x_0 - a_2 x_{-1} \\ -a_2 x_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \epsilon + \mu,$$

therefore $\hat{\epsilon} \sim \mathbf{N}(\mu, 1)$.

We denote $H$ as

$$\begin{pmatrix} 1 & & & & \\ a_1 & 1 & & & \\ a_2 & a_1 & 1 & & \\ 0 & a_2 & a_1 & 1 & \\ \vdots & & & & \\ 0 & 0 & a_2 & a_1 & 1 \end{pmatrix},$$

so

$$Hx = \vec{e}$$

Taking the expectation of

$$\hat{e}\hat{e}^t = (e+\mu)(e+\mu)^t = ee^t + \mu e^t + e\mu^t + \mu\mu^t,$$

yields

$$\mathbf{E}(\hat{e}\hat{e}^t) = \mathbf{E}(ee^t) + \mu\mu^t = \mathbf{I} + \mu\mu^t.$$

Taking the expectation of (5.3.1) leads to

$$H\mathbf{E}(xx^t)H^t = \mathbf{E}(\hat{e}\hat{e}^t) = \mu\mu^t + \mathbf{I},$$

therefore

$$\Sigma_x = H^{-1}(\mu\mu^t + \mathbf{I})(H^t)^{-1}.$$

Since

$$\mu\mu^t + \mathbf{I} = \begin{pmatrix} \mu_1^2+1 & \mu_1\mu_2 & \dots & 0 \\ \mu_1\mu_2 & \mu_2^2+1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

we have

$$|\Sigma_x| = |H^{-1}||(\mu\mu^t + \mathbf{I})||(H^t)^{-1}| = |\mu\mu^t + \mathbf{I}| = (\mu_1^2+1)(\mu_2^2+1) - 2\mu_1\mu_2.$$

# Chapter 6

# Estimation of transition matrix

## 6.1 EM algorithm for mixture models

**Definition 4** (Complete and Incomplete Data)**.** $x = \{x_1, \ldots, x_n\}$ is the observed data, and $q = \{q_1, \ldots, q_n\}$ represents the unobserved latent data or missing data. $x$ is incomplete data, while $\{x, q\}$ is the complete data.

The log-likelihood function of the complete data is

$$\mathcal{L}(\theta \mid x, q) = \log \prod_{i=1}^{T} P(x_i, q_i \mid \theta),$$

where the $\theta$ is the set of parameters.

The EM algorithm first seeks to find the expectation of the log-likelihood with respect to the unknown data $q$ giving the observable data $x$ and the current parameter estimation $\theta^i$.

E-step:

$$\begin{aligned}\tau &= E[\mathcal{L}(\theta^{i+1} \mid x, q) \mid \theta^i, x] \\ &= \sum_{q_i} \mathcal{L}(\theta^{i+1} \mid x, q) P(q \mid \theta^i, x) \\ &= \sum_{q_i} \log(\prod_{t=1}^{T} P(x_t, q_t|\theta^{i+1})) P(q_i \mid \theta^i, x_i),\end{aligned}$$

where $P(q_i \mid \theta^i, x)$ is the marginal distribution of unobservable data $q_i$, dependent on the current estimation of parameter $\theta$ and known data $x$.

The second step of the EM algorithm is to maximize the expectation we got in the first step.

M-step:

$$\max_{\theta^{i+1}} \tau.$$

It can be proved that $\tau(\theta^{i+1}) > \tau(\theta^i)$.

**Example 2.** Gaussian mixture

We wish to model a data set by specifying a joint distribution $P(x_i, z_i) = P(x_i|z_i)P(z_i)$. Here, $z_i \sim$ multinomial($\tau$), and $x_i \sim \mathcal{N}(\mu_j, \Sigma_j)$. We let $k$ denote the number of values $z_i$ can take, $\sum_{j=1}^{k} \tau_j = 1$. Thus our model posits that randomly choose $z_i$ from $\{1 \ldots k\}$, then $x_i$ is drawn from one of $k$ Gaussian distributions depends on the value of $z_i$. For example, the $x_i \mid z_i = 1$ is generated by $\mathcal{N}(\mu_1, \Sigma_1)$, similarly $x_i \mid (z_i = 2) \sim \mathcal{N}(\mu_2, \Sigma_2)$, where $P(z_i = 1) = \tau_1$, and $P(z_i = 1) = \tau_2$.

The joint distribution is

$$\sum_{j=1}^{k} f(x_i \mid \mu_j, \Sigma_j) \tau_j 1_(z_i = j).$$

The likelihood function is

$$\mathcal{L}(x_i, z_i \mid \theta) = \prod_{i=1}^{m} \sum_{j=1}^{k} f(x_i \mid \mu_j, \Sigma_j)\tau_j 1_{(} z_i = j),$$

where $m$ is the sample size, $\theta = \{\mu, \Sigma, \tau\}$.

The conditional distribution of $Z_i$ is calculated by

$$\begin{aligned} T_{i,j} &= P(z_i = j \mid x_i) \\ &= \frac{P(z_i, x_i)}{P(x_i)} \\ &= \frac{\sum_{j=1}^{k} f(x_i \mid \mu_j, \Sigma_j)\tau_j 1_{(} z_i = j)}{\sum_{j=1}^{k} f(x_i \mid \mu_j, \Sigma_j)\tau_j} \end{aligned}$$

The expectation of likelihood function is

$$\mathbb{E}(\log \mathcal{L}(x_i, z_i \mid \theta)) = \prod_{i=1}^{m} \sum_{j=1}^{k} T_{i,j} f(x_i \mid \mu_j, \Sigma_j)\tau_j 1_{(} z_i = j).$$

## 6.2 Baum-Welch algorithm for HMM models

Given a set of observed data, we derive the EM algorithm for finding the maximum- likelihood estimation of the parameters of a hidden Markov model. This algorithm is known as the Baum-Welch algorithm.

### Discrete Hidden Markov models

Consider a system which may be described at any time by two variables: observation $O_t$ and states $S_t$. Variable $O_t$ is observable while variable $S_t$ is latent and not observable. The value of $O_t$ is discrete $O_t \in \{x_1, x_2, \ldots, x_T\}$, and $S_t \in \{s_1, s_2, \ldots, s_N\}$ with $T$ as number of steps and $N$ as number of states.

**Example 3.** Consider a three state Markov model of weather. The weather is one of three states:

state 1: rainy,

state 2: cloudy,

state 3: sunny.

The speed of wind is categorized as windy, less windy, no wind. We know wind speed of last month, however, we don't know the weather of last month except the following probability table

$$\left(\begin{array}{cccc} & \textit{windy} & \textit{less windy} & \textit{no wind} \\ \textit{rainy} & 0.6 & 0.3 & 0.1 \\ \textit{cloudy} & 0.1 & 0.5 & 0.4 \\ \textit{sunny} & 0.4 & 0.3 & 0.3 \end{array}\right)$$

The parameters for the HMM are $\{A, B, \pi\}$, where $A = \{a_{ij}\}$ is the transition matrix, $B = \{b_i(x_t)\}$ and $\pi_i$ is the initial distribution. $B = \{b_i(x_t)$ follows above table.

Two intermediate variables need to be defined first:

$$\alpha_t(i) = P(o_1 = x_1, \ldots, o_t = x_t, S_t = i)$$

and

$$\beta_t(j) = P(o_{t+1} = x_{t+1}, \ldots, o_T = x_T \mid S_t = i)$$

Variable $\alpha_t(i)$ can be calculated with the forward method (Rabiner [1990], Frey [2010]) with two steps: initialization step

$$\alpha_1(i) = \pi(i)b_i(x_1)$$

and induction step

$$\alpha_{t+1}(j) = (\sum_{i=1}^{N} \alpha_t(i)a_{ij})b_j(x_{t+1}) \tag{6.2.1}$$

$$t = 1, \ldots, T-1.$$

Variable $\beta_t(j)$ can be calculated with backward method also with two steps: initialization steps

$$\beta_T(i) = 1$$

and induction step

$$\beta_t(j) = (\sum_{i=1}^{N} \beta_{t+1}(i)a_{ij})b_j(x_{t+1}) \tag{6.2.2}$$

$$t = T-1, \ldots, 0.$$

Then with the E-step, we can get

$$\begin{aligned} \xi_t(i,j) &= P(q_t = i, q_{t+1} = j \mid X, \theta) \\ &= \frac{P(q_t = i, q_{t+1} = j, X \mid \theta)}{P(X \mid \theta)} \\ &= \frac{\alpha_t(i)\beta_{t+1}(j)a_{ij}b_j(x_{t+1})}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\beta_{t+1}(i)a_{ij}b_j(x_{t+1})} \end{aligned} \tag{6.2.3}$$

$$
\begin{aligned}
\gamma_t(i) &= P(q_t = i \mid X, \theta) \\
&= \sum_{j=1}^{N} \xi_t(i,j) \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}
\end{aligned} \tag{6.2.4}
$$

The second equation holds because

$$\beta_t(j) = (\sum_{i=1}^{N} \beta_{t+1}(i)a_{ij})b_j(x_{t+1}).$$

A set of reasonable re-estimation formulas for $\pi$, $A$ and $B$ are

$$
\begin{aligned}
\bar{\pi}_i &= \text{expected frequency in state } S_i \text{ at time } t = 1 \\
&= \gamma_{t=1}(i)
\end{aligned} \tag{6.2.5}
$$

$$
\begin{aligned}
\bar{a}_{ij} &= \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} \\
&= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}
\end{aligned} \tag{6.2.6}
$$

$$
\begin{aligned}
\bar{b}_j(k) &= \frac{\text{expected frequency in state } S_j \text{ and observing symbol } x_k}{\text{expected frequency in state } S_j} \\
&= \frac{\sum_{t=1}^{T-1} \gamma_t(j)b_j(x_k)}{\sum_{t=1}^{T-1} \gamma_t(j)}
\end{aligned} \tag{6.2.7}
$$

### 6.2.1 Continuous time Hidden Markov models

Our discussion in the last section has considered only the case when observations were characterized as discrete symbols. A continuous observation density will be used in this section. The pdf function is a mixture of the form

$$b_j(O) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(O, \mu_{jm}, U_{jm})$$

where $c_{jm}$ is the coefficient for $m-th$ mixture component in state $j$, which satisfy the constraint

$$\sum_{m=1}^{M} c_{jm} = 1$$

$$c_{jm} \geq 0$$

When $M = 1$, the continuous observation follows a Gaussian distribution.

The parameters for the HMM are $\{A, B, \pi\}$, where $A = \{a_{ij}\}$ is the transition matrix, $B = \{b_i(x_t) = \sum_{m=1}^{M} c_{im} \mathcal{N}(\mu_{im},\, U_{im})(x_t)\}$ and $\pi_i$ is the initial distribution.

$$\bar{\mu}_{ik} = \frac{\sum_{t=1}^{T} x_t \gamma_t(i,k)}{\sum_{t=1}^{T} \gamma_t(i,k)} \tag{6.2.8}$$

$$\bar{U}_{ik} = \frac{\sum_{t=1}^{T} (x_t - \mu_{ik})(x_t - \mu_{ik})' \gamma_t(i,k)}{\sum_{t=1}^{T} \gamma_t(i,k)} \tag{6.2.9}$$

$$b_t(i,k) = \frac{c_{jk} \mathcal{N}(O, \mu_{jk}, U_{jk})}{\sum_{m=1}^{M} c_{jm} \mathcal{N}(O, \mu_{jm}, U_{jm})}, \tag{6.2.10}$$

where $b_t(i,k)$ denotes the probability of being at state $i$ at time $t$ for $k$-th component accounting for observation $s_t$.

$$\gamma_t(i,k) = \left(\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}\right)\left(\frac{c_{jk}\mathcal{N}(O,\mu_{jk},U_{jk})}{\sum_{m=1}^{M}c_{jm}\mathcal{N}(O,\mu_{jm},U_{jm})}\right)$$

### 6.2.2 Hidden Markov driven autoregressive model

We estimate mean vector $\mu_i$ and covariance matrix $\Sigma_i$ from sample data as

$$\mu_i = \frac{\sum_{t=1}^{T} x_t\gamma_t(i)}{\sum_{t=1}^{T}\gamma_t(i)}$$

and

$$\Sigma_i = \frac{\sum_{t=1}^{T}(x_t-\mu_i)(x_t-\mu_i)'\gamma_t(i)}{\sum_{t=1}^{T}\gamma_t(i)}, \tag{6.2.11}$$

then use algorithm (4) to estimate $H_i$ from $\Sigma_i$.

The density function for observation $\vec{x}$ is defined by Equation (5.2.5), so the emission probability is

$$b_t(i) = (2\pi)^{-K/2}\left|\Sigma_i\right|^{-1/2}\exp\{-\frac{1}{2}x^t\Sigma_i^{-1}x\},$$

where

$$\Sigma_i^{-1} = H_i^tH_i$$

and

$$\left|\Sigma_i\right| = \left|H_i\right|^{-2}.$$

Three pseudocodes below summarize algorithm of applying Balm Welch to three HMM cases.

**Algorithm 1** HMM Forward.

```
Initialize: $t \leftarrow 0$, $a_{ij}$, $b_j$, visible sequence $\vec{s}$, $\alpha_j(0)$
repeat
    $t \leftarrow t + 1$
    $\alpha_j(t) \leftarrow b_j(s_t) \sum_{i=1}^{M} \alpha_i(t-1) a_{ij}$
until $t = T$
return $\alpha_j(T)$ for the final state
```

**Algorithm 2** HMM Backward.

```
Initialize: $t \leftarrow T$, $a_{ij}$, $b_j$, visible sequence $\vec{s}$, $\beta_j(T)$
repeat
    $t \leftarrow t - 1$
    $\beta_i(t) \leftarrow \sum_{j=1}^{M} \beta_j(t+1) a_{ij} b_j s_{t+1}$
until $t = 1$
return $\beta_i(0)$ for the known initial state.
```

### 6.2.3 Numerical example: performance of Balm-Welch estimator

Assume the transition matrix is

$$\begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}.$$

We set up our testing cases as:

- $p_1 = (0.95, 0.8, 0.5)$,

- $p_2 = (0.8, 0.5, 0.3)$,

- $\mu = (0.5, -0.3)$,

- $\sigma = (0.5, 0.8)$,

- sample size = 10000,

**Algorithm 3** HMM EM algorithm.

```
Initialize: $a_{ij}$, $b_j$, training sequence $\vec{x}$, convergence criterion $\theta$, $z \leftarrow 0$
repeat
    $z \leftarrow z + 1$
    Compute $\alpha_i(t)$ by forward algorithm (1).
    Compute $\beta_i(t)$ by backward algorithm (1).
    Compute sufficient statistics $\xi_{i,j}(t)$ from $\alpha_i(t)$, $\beta_i(t)$ and $b(z-1)$ by Eq. (6.2.3)
    Compute sufficient statistics $\gamma_i(t)$ from $\alpha_i(t)$, $\beta_i(t)$ and $b(z-1)$ by Eq. (6.2.4)
    Update transition matrix $a(z)$ from $a(z-1)$, $\xi_{i,j}(t)$ and $\gamma_i(t)$ by Eq. (6.2.6)
    if Discrete case model then
        Compute $b(z)$ by Eq. (6.2.7)
    end if
    if Gaussian mixture model then
        Compute $\bar{\mu}_{ik}$ by Eq. (6.2.8)
        Compute $\bar{U}_{ik}$ by Eq. (6.2.9)
        Compute $b(z)$ by Eq. (6.2.10)
    end if
    if Autoregressive model then
        Estimate covariance matrix $\Sigma_i$ for each state $i$ by Eq. (6.2.11)
        Compute $H_i$ for each state $i$ by Algorithm (4)
        Compute emission probabilities $b(z)$ by Eq (6.2.2)
    end if
until $z = T$
return
```

Figure (6.2.1) shows a boxplot of Balm-Welch estimates for the hidden Markov driven Gaussian mixture model from replications 100 with sample size 10000 with parameters $\sigma_1 = 0.5$, $\sigma_2 = 0.8$, $\mu_1 = 0.5$, $\mu_2 = -0.3$, $\{p_1, p_2\} \in P$. $P$ is set of combinations of $\{p_1, p_2\}$. Let

$$\begin{aligned} P :=& \{0.95, 0.8\}, \{0.95, 0.5\}, \{0.95, 0.3\}, \{0.8, 0.8\}, \{0.8, 0.5\}, \\ & \{0.8, 0.3\}, \{0.5, 0.8\}, \{0.5, 0.5\}, \{0.5, 0.3\}. \end{aligned}$$

We see that the estimator performs well for $p_1$, $p_2$, $\sigma_1$ and $\sigma_2$, but tends to have some bias for $\mu_1$ and $\mu_2$. Although this bias exists, it still delivers reasonable estimates for number of observation as small as 100. We can also see cases with symmetric parameters, for example, $p1 = 0.5$, $p2 = 0.5$, has better estimation. Figures (6.2.4a) to (6.2.4c) show number of observations $= (25, 50, 200, 3000)$ for each combination. It also demonstrates that models with symmetric parameters converge better under our algorithm.

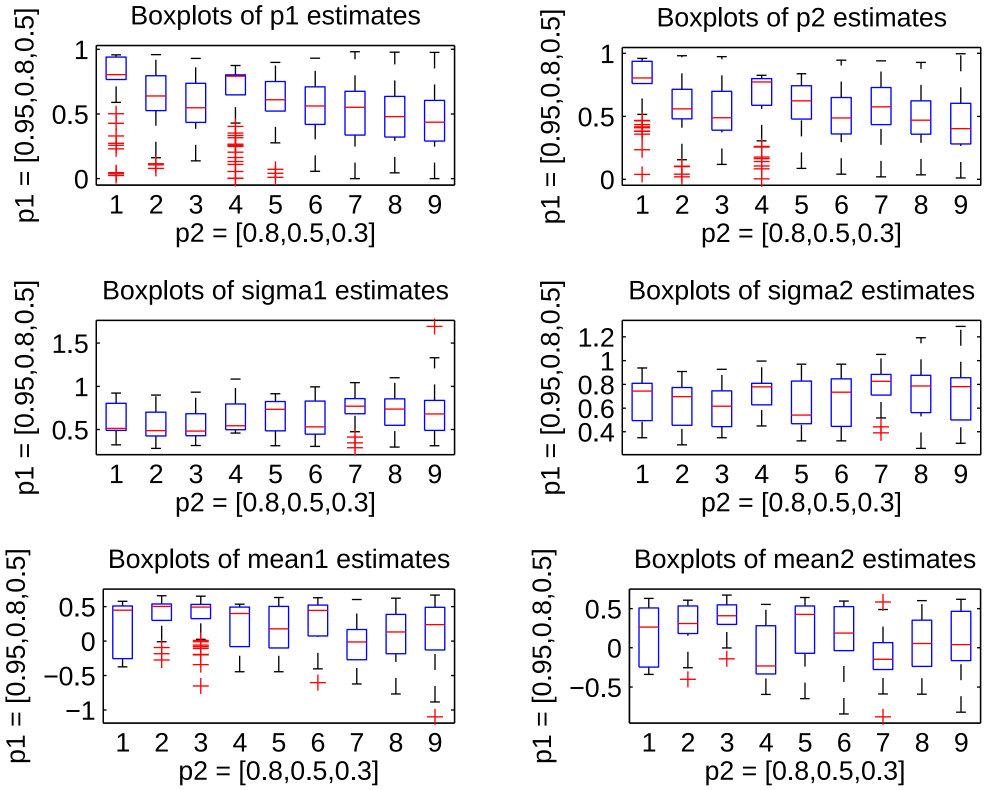
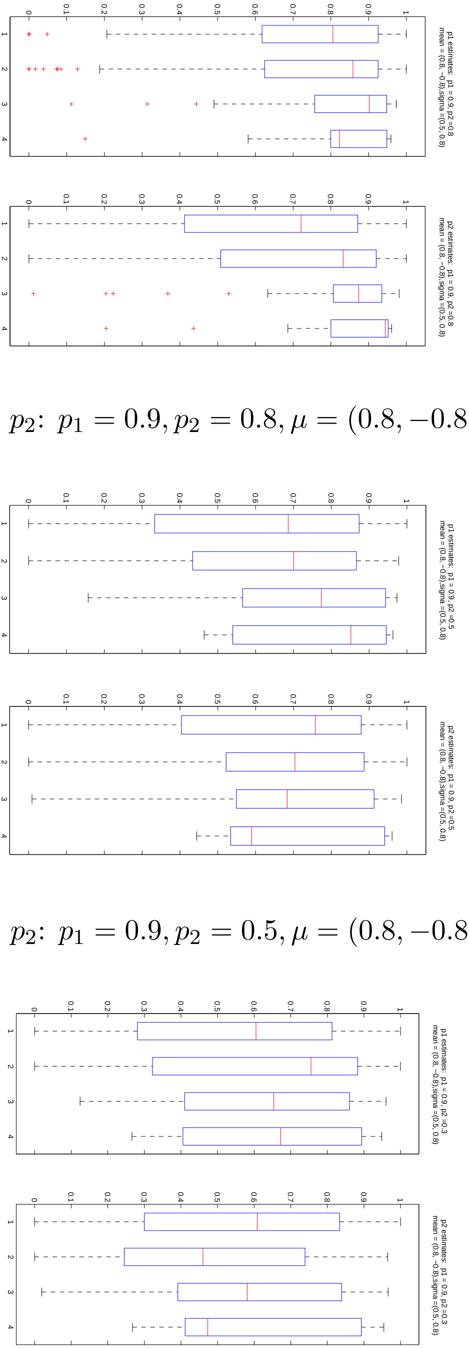


Figure 6.2.1: Boxplots of estimated $p_1$, $p_2$, $\sigma_1$, $\sigma_2$, $\mu_1$, $\mu_2$ for number of observations = 100 with parameters $\sigma_1 = 0.5$, $\sigma_2 = 0.8$, $\mu_1 = 0.5$, $\mu_2 = -0.3$, $\{p_1, p_2\} \in P$.

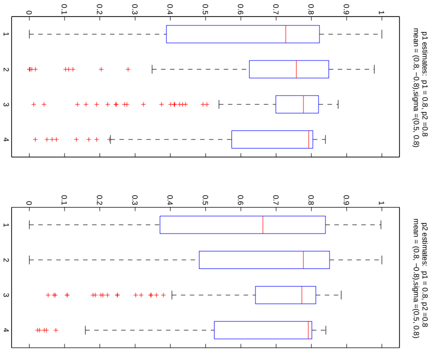


(a) Estimated $p_1$ and $p_2$: $p_1 = 0.9, p_2 = 0.8, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

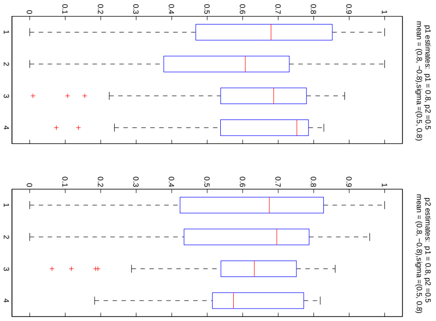
(b) Estimated $p_1$ and $p_2$: $p_1 = 0.9, p_2 = 0.5, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

(c) Estimated $p_1$ and $p_2$: $p_1 = 0.9, p_2 = 0.3, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

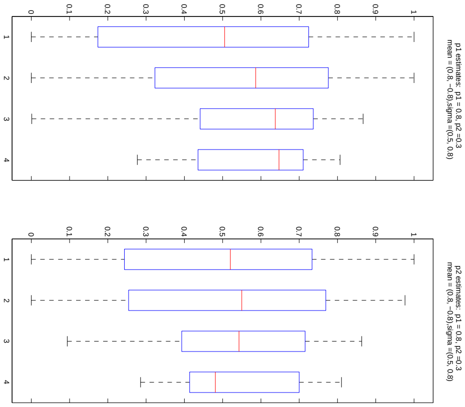
Figure 6.2.2: Boxplots of estimated $p_1$ and $p_2$ with number of observations $=$ $(25, 50, 200, 3000)$.



(a) Estimated $p_1$ and $p_2$: $p_1 = 0.8, p_2 = 0.8, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

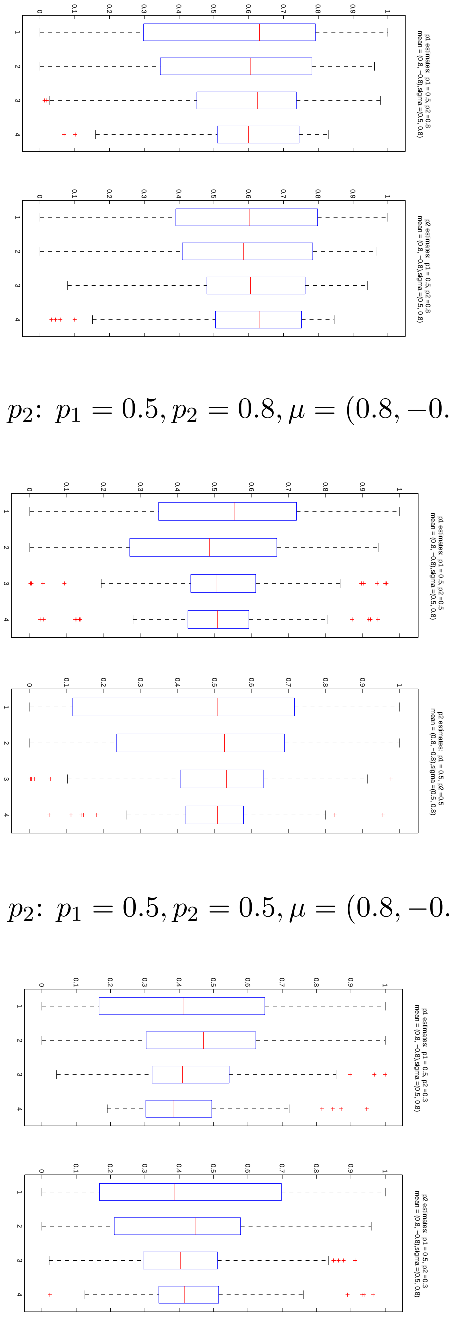


(b) Estimated $p_1$ and $p_2$: $p_1 = 0.8, p_2 = 0.5, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.



(c) Estimated $p_1$ and $p_2$: $p_1 = 0.8, p_2 = 0.3, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

Figure 6.2.3: Boxplots of estimated $p_1$ and $p_2$ with number of observations = $(25, 50, 200, 3000)$.



(a) Estimated $p_1$ and $p_2$: $p_1 = 0.5, p_2 = 0.8, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

(b) Estimated $p_1$ and $p_2$: $p_1 = 0.5, p_2 = 0.5, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

(c) Estimated $p_1$ and $p_2$: $p_1 = 0.5, p_2 = 0.3, \mu = (0.8, -0.8), \sigma = (0.5, 0.8)$.

Figure 6.2.4: Boxplots of estimated $p_1$ and $p_2$ with number of observations $= (25, 50, 200, 3000)$.

# Chapter 7

# Estimation of autoregressive coefficients

We develop three estimation methods for autoregressive coefficients. The MLE method needs approximation of the p.d.f. to some extent, thus we apply the MLE method to problems with large sample size $T$. The ordinary least square (OLS) method doesn't give information of distribution of data, therefore we use it to generate an initial guess. The Frobenius norm minimization method is parsimonious, stable, and shows high resolution precision according to our tests. We choose Frobenius norm minimization method as estimation method for our model. Due to data errors, the correlation matrix we get from real data is not always positive definite, thus we further provide correlation matrix fixup method and numerical examples.

## 7.1 MLE estimation with approximate expression of p.d.f.

**One method to calibrate $a$ and $\sigma$** The p.d.f. function (5.1.1) is defined by parameter $\sigma$ and $a$. Given a data observation vector $\vec{s} = (x_0, x_1, \ldots, x_{K-1})$, we can determine the maximum likelihood estimate of $\sigma$ and $a$ that best characterizes the observed $\vec{s}$. The log likelihood function is

$$\log f(\vec{s} \mid \sigma, a) = -\frac{K}{2}\log(2\pi\sigma^2) - \frac{\delta(\vec{s}, a)}{2\sigma^2}.$$

Instead of searching for optimal values in two dimensions, we search in one dimension first, then search in the other dimension.

$$\begin{aligned} h(\sigma, a) &:= \log f(\vec{s} \mid \sigma, a) \\ g_\sigma(a) &:= h(\sigma, a) \\ \hat{a}(\sigma) &= \arg\max_a g_\sigma(a) \\ \hat{\sigma} &= \arg\max_\sigma h(\sigma, \hat{a}). \end{aligned}$$

Therefore, the ML estimate is

$$\begin{aligned} \hat{a}_{ML} &= \arg\max_a \log f(\vec{s} \mid \sigma, a) \\ &= \arg\min_a \delta(\vec{s}, a) \\ &= \arg\min_a (a' R a) \end{aligned}$$

where $R = [r_{ij}]$ with $r_{ij} = r(|i-j|)$.

If there is no constraint for $a$, since $R$ is symmetric positive definite, we have optimal result $\hat{a}'R\hat{a} = 0$ with $\hat{a} = 0$, however, with constraint $a_0 = 1$, $\min_{\vec{a}} f(\vec{a}) = \min_{a_1, a_2, \ldots, a_{p-1}} f(\vec{a})$, which can be solved with Lagrangian multipliers.

Furthermore, optimize object function $\log f(\vec{s} \mid \sigma, \hat{a})$ over $\sigma$

$$\min_{\sigma} \log f(\vec{s} \mid \sigma, \hat{a}) \Leftrightarrow \min_{\sigma} \left( -\frac{K}{2} \log(2\pi\sigma^2) - \frac{\delta(\vec{s}, \hat{a})}{2\sigma^2} \right)$$

Take derivative with respect to $\delta$, and let $x = \sigma^2, x \geq 0$

$$-\frac{K}{2x} + \frac{\delta}{2x^2} = 0 \Leftrightarrow -\frac{1}{2x^2}(Kx - \delta) = 0 \Leftrightarrow \hat{\sigma}^2 = \frac{\delta}{K}.$$

Estimated $\sigma$ is

$$\begin{aligned} \sigma_{ML} &= \arg\min_{\sigma} \log f(\vec{s} \mid \sigma, \hat{a}) \\ &= \delta(\vec{s}, \hat{a})/K. \end{aligned}$$

**Alternative method to calibrate the model** When $p = 2$, the object function is

$$\begin{aligned} &\max_{a_1, a_2} \log \prod_{i}^{K} p(e_i) \\ \Leftrightarrow &\max_{a_1, a_2} \sum_{i}^{K} \log p(e_i) \\ \Leftrightarrow &\max_{a_1, a_2} \left( K \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i}^{K} e_i^2 \right) \\ \Leftrightarrow &\min_{a_1, a_2} \sum_{i}^{K} e_i^2. \end{aligned}$$

This equals to maximize the probability of error term $e_i$ appearing around mean 0. When $K = 10$,

$$\min_{a_1,a_2} \begin{pmatrix} e_1^2 \\ +e_2^2 \\ \vdots \\ +e_{10}^2 \end{pmatrix} = \begin{pmatrix} (x_2 + a_1x_1 + a_2x_0)^2 \\ +(x_3 + a_1x_2 + a_2x_1)^2 \\ \vdots \\ +(x_{10} + a_1x_9 + a_2x_8)^2 \end{pmatrix} =: V(a_1, a_2, x_0, \ldots, x_10)$$

Take $\frac{\partial V}{\partial a_1} = 0$ and $\frac{\partial V}{\partial a_2} = 0$, we can get

$$A \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 0$$

with

$$A := \begin{pmatrix} m_1 & m_2 \\ n_1 & n_2 \end{pmatrix},$$

$$\begin{aligned} m_1 &= x_1^2 + x_2^2 + \ldots + x_9^2 \\ &= \sum_{i=1}^{9} x_i^2, \end{aligned}$$

$$\begin{aligned} m_2 &= x_0x_1 + x_1x_2 + \ldots + x_8x_9 \\ &= \sum_{i=1}^{9} x_{i-1}x_i, \end{aligned}$$

and

$$\begin{aligned} b_1 &= x_1x_2 + x_2x_3 + \ldots + x_9x_{10} \\ &= \sum_{i=1}^{9} x_i x_{i+1} \end{aligned}$$

In similar way, we can get

$$\begin{aligned} n_1 &= \sum_{i=1}^{9} x_{i-1}x_i \\ n_2 &= \sum_{i=0}^{8} x_i^2 \\ b_2 &= \sum_{i=0}^{8} x_i x_{i+2} \end{aligned}$$

Then, then we have estimated

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = -A^{-1} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

With estimated $\{\hat{a}_1, \hat{a}_2\}$, we can get $\hat{e}$. We find the optimal $\sigma$ by searching in the other direction:

$$\begin{aligned} &\max_{\sigma} \left( K \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i}^{K} \hat{e}_i^2 \right) \\ \Leftrightarrow &\min_{\sigma} \left( K \log \sqrt{2\pi}\sigma + \frac{1}{2\sigma^2} \sum_{i}^{K} \hat{e}_i^2 \right) \\ \Leftrightarrow &\hat{\sigma} = \left(\frac{\sum_i \hat{e}_i^2}{K \log \sqrt{2\pi}}\right)^{1/3} \end{aligned}$$

## 7.2 Estimation of coefficients matrix with ordinary least squares

This section discusses the application of ordinary least squares (OLS) to estimate autoregressive coefficients. For example, suppose order of regression $p = 2$, observation window $K = 5$, sample size $T = 10$, observations of stock returns $x_1, x_2, \ldots, x_{10}$. We wish to predict tomorrow's stock return based on today's and yesterday's return. Thus we write this problem as

$$\begin{aligned}
&x_3 = -a_1 x_2 - a_2 x_1 + e_3 \\
&x_4 = -a_1 x_3 - a_2 x_2 + e_4 \\
&\vdots \\
&x_{10} = -a_1 x_9 - a_2 x_8 + e_{10}.
\end{aligned}$$

We want to minimize

$$e := \begin{pmatrix} e_3 \\ e_4 \\ \vdots \\ e_{10} \end{pmatrix} = \begin{pmatrix} x_2 & x_1 \\ x_3 & x_2 \\ \vdots & \\ x_9 & x_8 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} x_3 \\ x_4 \\ \vdots \\ x_{10} \end{pmatrix} := Xa - x.$$

Let $\hat{a}$ be the solution of the least square problem, which is also the estimated autoregressive coefficients. We still need to estimate $\sigma$ for our model. We get estimated $\hat{\sigma} = \frac{\sigma_{imp}}{\sqrt{T}}$, where $\sigma_{imp}$ is implied volatility from one year European option. The next example justifies this estimation method.

For example, $s$ is daily return of stock, $\sigma$ is standard deviation of daily returns, $\sigma_{imp}$ is implied volatility of one year option, $T = 250$, $\Delta t = 1/250$. Since

$$\frac{ds}{s} = \mu dt + \sigma dW_t,$$

therefore

$$\frac{s_t - s_{t-1}}{s_{t-1}} = \mu \Delta t + \sigma_{imp} \sqrt{\Delta t} z_t,$$

mean $\mu$ is not stochastic but determinant, $\Delta t = 1/250 = 0.004$, $\sqrt{\Delta} t = 0.0632$, so the first term on the right hand side can be ignored. In other words, return of stock $r_t = \frac{s_t - s_{t-1}}{s_{t-1}} \sim z_t$ with standard deviation $\sigma_{imp}\sqrt{\Delta t}$.

There occurred a shortcoming for this estimation method that it doesn't give information of the distribution, therefore it is used to generate initial guess.

## 7.3 Estimation of coefficients matrix $H$ by minimizing Frobenius norm

The assumption for this method is that $\vec{x}$ are correlated multivariate normal variables and $\vec{e}$ are i.i.d. $\mathbf{N}(0, 1)$ random variables. We have

$$H\vec{x} = \vec{e}$$

with

$$H = \left(\begin{array}{cc|ccc} h_{11} & 0 & & 0 & \\ h_{21} & h_{22} & & & \\ \hline \vdots & \vdots & \ddots & & 0 \\ 0 & a_2 & a_1 & 1 & \\ 0 & 0 & a_2 & a_1 & 1 \end{array}\right).$$

First, we estimate covariance matrix $\Sigma_x$ from sample data $x$.

Then, based on equation(5.2.3), we know

$$\Sigma_x^{-1} = H^t H,$$

so we use Cholesky decomposition to get $\Sigma_x$

$$U^t U = \mathrm{Chol}(\Sigma_x)$$
$$\tilde{H}^{-1} = U^t.$$

Last, we minimize the Frobenius norm of $\tilde{H} - H$

$$\epsilon_{ij} = \tilde{H}_{ij} - H_{ij}$$
$$a^* = \min_{a_1,a_2}(\sum\nolimits_{1<i,\,j<K} \epsilon_{ij}^2)$$

**Algorithm 4** Estimation of autocorrelation coefficients.

```
Given $\vec{x} = x_1^{(i)}, \cdots, , \; x_K^{(i)}$.
Estimate covariance matrix $\vec{\Sigma}_x$ from $\vec{x}$.
$U \leftarrow \mathrm{Chol}(\vec{\Sigma}_x)$
$\tilde{H} = (U^{-1})^t$
$\epsilon_{ij} = \tilde{H}_{ij} - H_{ij}$
$a^* \leftarrow \min_{a_1, a_2}(\sum_{1<i,\, j<K} \epsilon_{ij}^2)$
```

### 7.3.1 Numerical example: accuracy test

We begin with

$$H = \left(\begin{array}{cc|ccc} h_{11} & 0 & & 0 & \\ h_{21} & h_{22} & & & \\ \hline \vdots & \vdots & \ddots & & 0 \\ 0 & a_2 & a_1 & 1 & \\ 0 & 0 & a_2 & a_1 & 1 \end{array}\right).$$

After taking

$$\Sigma_x = H^{-1} H^{-t},$$

we generate scenarios $\vec{x}$ with covariance $\Sigma_x$ and expectation $\mathbf{E}(\vec{x}) = 0$. Having $\vec{x}$, we use our method Algorithm (4) to find $\tilde{H}$. Then make a comparison of Frobenius norms between $\tilde{H}_{ij}$ and $H_{ij}$ to see if

$$\|\tilde{H} - H\|_F^2 \simeq 0,$$

where

$$\Delta_F := \|\tilde{H} - H\|_F^2 = \sum_{i,\, j} (\tilde{H}_{i,\, j} - H_{i,\, j})^2.$$

Table (7.1) shows estimation errors measured by $\Delta_F$ with respect to 6 sets of parameters. We can see when $\vec{a} = \{a_0, a_1, a_2\}$ with $a_0 = 1$, $|a_1| < 1$ and $|a_2| < 1$, errors $\Delta_F$ is less than

0.02. When $a_0 = 1$, $|a_1| > 1$ and $|a_2| > 1$, errors $\Delta_F$ is larger than 0.03. From the model definition, we know $a_0 = 1$, $|a_1| < 1$ and $|a_2| < 1$ is a reasonable assumption, which means yesterday's price has less impact than today's price, the day before yesterday's price has less impact than both yesterday's and today's price. Table (7.2) shows when $|a_1| < 1$ and $|a_2| < 1$, this method can get errors less than 0.02 with simulation number 5000.

Table 7.1: Errors $\Delta_F$ with respect to different parameters, with 10000 simulations.

|  | $\{h_{11}, h_{21}, h_{22}\}$ |  |
|---|---|---|
| $\vec{a}$ | $\{0.02, 0.01, 0.05\}$ | $\{0.2, 0.1, 0.5\}$ |
| $\{1, 5, 3\}$ | 0.0548 | 0.0345 |
| $\{1, 0.5, 0.3\}$ | 0.0015 | 0.0046 |
| $\{1, 0.5, -0.3\}$ | 0.0030 | 0.0065 |

Table 7.2: Errors $\Delta_F$ with respect to different simulation numbers.

|  | $\{h_{11}, h_{21}, h_{22}\} = \{0.2, 0.1, 0.5\}$ |  |
|---|---|---|
| numer of simuation | $\vec{a} = \{1, 5, 3\}$ | $\vec{a} = \{1, 0.5, 0.3\}$ |
| 100 | 0.4146 | 0.0871 |
| 1000 | 0.1050 | 0.0381 |
| 5000 | 0.0845 | 0.0197 |
| 10000 | 0.0319 | 0.0056 |

### 7.3.2 Numerical example: stability test

We add a small value $\iota$ to zero entries in $H$ with $|\iota_i| \leq \frac{|h_{ij}|}{100}$ to make it more realistic

$$H = \left(\begin{array}{cc|ccc} h_{11} & \iota & & \iota & \\ h_{21} & h_{22} & & & \\ \hline \vdots & \vdots & \ddots & & \\ \iota & a_2 & a_1 & 1 & \\ \iota & \iota & a_2 & a_1 & 1 \end{array}\right).$$

With new $H$, we apply the same algorithm as in Section (7.3.1). Compared with Table (7.2), results in Table (7.3) show that with perturbation, this method can also get errors less than 0.02 with simulation number 5000 when $|a_1| < 1$ and $|a_2| < 1$. To summarize, this is a stable method.

Table 7.3: Errors $\Delta_F$ with respect to different simulation numbers with perturbations.

| | $\{h_{11}, h_{21}, h_{22}\} = \{0.2, 0.1, 0.5\}$ | |
|---|---|---|
| numer of simuation | $\vec{a} = \{1, 5, 3\}$ | $\vec{a} = \{1, 0.5, 0.3\}$ |
| 100 | 0.7477 | 0.1595 |
| 1000 | 0.1657 | 0.0334 |
| 5000 | 0.0759 | 0.0178 |
| 10000 | 0.0278 | 0.0083 |

## 7.4 Correlation matrix fixup method

Given a symmetric matrix $A$, we construct a positive semi-definite matrix $\hat{A}$ by using eigen-decomposition of A. We begin with normalizing $A$ so that it has unit diagonal elements.

Let

$$A = DA^{(0)}D,$$

where $D$ is diagonal matrix, $A^{(0)}$ is symmetric and has unit diagonal elements. Let

$$A^{(0)} = Q\Lambda Q^t,$$

be the symmetric eigendecomposition of $A^{(0)}$ into orthogonal matrix $Q$ of eigenvectors and diagonal matrix $\Lambda$. Let $\Lambda^+$ be the diagonal matrix consisting of the elements $\max(\lambda_i, 0)$. Let

$$A^{(1)} = Q\Lambda^+ Q^t,$$

This matrix $A^1$ is positive semi-definite, but we normalize it so it has unit diagonal elements. Define a diagonal matrix $S$ to have diagonal elements $s_{ii} = {a_{ii}^{(1)}}^{-1/2}$, where $a_{ii}^{(1)}$ are the diagonal elements of $A^{(1)}$. Then the matrix

$$A^{(2)} = SA^{(1)}S$$

is a positive semi-definite symmetric matrix with unit diagonal. Let

$$\hat{A} = DA^{(2)}D$$

is the desired semi-definite symmetric matrix close to our original $A$.

### 7.4.1 Numerical example: speed and accuracy test

We perform some tests where we measured the time taken for each method along with the distance from the original $A$ and the fixed-up matrix $\hat{A}$. To measure the distance we use the $L^2-$norm, i.e. we want to minimize the quantity

$$\chi^2 = \|A - \hat{A}\|^2_{L^2} = \sum_{i,j}(a_{i,j} - \hat{a}_{i,j})^2.$$

For each matrix size $N$ we computed around $1000/N$ random symmetric matrices with unit diagonal and non-diagonal elements between $-1$ and 1. For the ones we fixed up, we looked at the average distance $\chi^2$. We also recorded the time taken for different methods. Table (7.4) also shows normalized distance $\chi^2_N$, i.e. $\chi^2$ divided by $N^2$ where $N$ denotes the size of matrix. The times are in milliseconds.

Table 7.4: Speed and accuracy test for correlation matrix fixup.

| Size | Time | $\chi^2$ | $\chi^2_N$ |
|---|---|---|---|
| 5 | 0.1 | 0.67 | 0.027 |
| 10 | 0.455 | 7.01 | 0.070 |
| 15 | 1.25 | 22.17 | 0.098 |
| 20 | 2.55 | 47.61 | 0.119 |
| 25 | 4.6 | 84.23 | 0.134 |
| 30 | 8.0 | 132.5 | 0.147 |
| 60 | 50 | 685.5 | 0.190 |
| 100 | 225 | 2174.1 | 0.217 |

## 7.5 A numerical example

To clarify our new estimation method for our model, consider a two-state autoregressive HMM model with autoregressive order $p = 2$, length of observation window $K = 5$. Coefficients matrices for each state are

$$H_1 = \begin{pmatrix} 0.1 & & & & \\ 0.9 & 0.2 & & & \\ 0.1 & 0.6 & 1 & & \\ & 0.1 & 0.6 & 1 & \\ & & 0.1 & 0.6 & 1 \end{pmatrix}$$

and

$$H_2 = \begin{pmatrix} 0.8 & & & & \\ 0.2 & 0.8 & & & \\ 0.3 & 0.5 & & & \\ & 0.3 & 0.5 & & \\ & & 0.3 & 0.5 & \end{pmatrix}$$

for each state. The transition probability matrix between two states is

$$\begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}.$$

The observation window with length 5 moves along time axis as in Figure (7.5.1). The autoregressive coefficients matrix $H_t$ describes dependence within each observation window

$$H_t \hat{x}_t = e_t$$

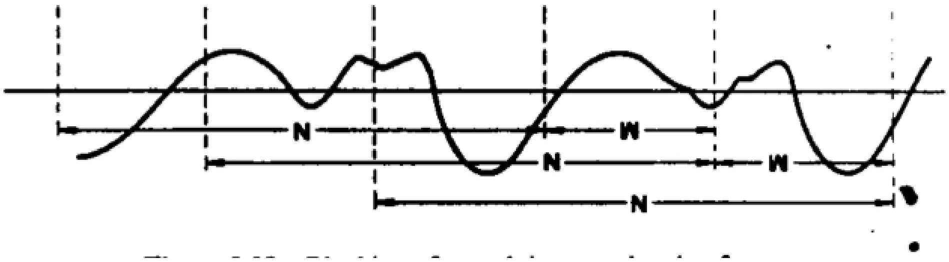


Figure 7.5.1: Illustration of overlapped observation window.

with $\hat{x}_t = \{x_t, x_{t+1}, x_{t+2}, x_{t+3}, x_{t+4}\}$. Since two states of $H$ occurs, $H_t \in \{H_1, H_2\}$, so there are two sets of dependence relationships within each observation window. Matrix $H_t$ is driven by hidden state variable $z_t$. Thus distribution of observation $x_t$ not only depends on previous observations, in this example, $x_{t-1}$ and $x_{t-2}$, but also depends on hidden state $z_t$. We study this example through simulation.

1. Initialize the process at $t = 0$ with initial state i drawn from the distribution $\pi$;

2. Call the current state $i$, simulate the new state $j$: simulate a discrete random variable with probability distribution given by the $i$-th row of the transition matrix, i.e., $q_{ij}/q_i, j \neq i$;

3. Given current state $i$, simulate a multi-gaussian random variable with mean $H^{-1}e$, variance $H^{-1}(ee^t + \mathbf{I})H^{-t}$

4. If t is less than a preassigned maximum time $T_{max}$, return to step 2.

The following program implements this algorithm in Matlab.

```
function M = sampleDiscrete(prob, r, c)

n = length(prob);R = rand(r, c);M = ones(r, c);

cumprob = cumsum(prob(:));
```

```
if n < r*c
    for i = 1:n-1
        M = M + (R > cumprob(i));
    end
else
    cumprob2 = cumprob(1:end-1);
    for i=1:r
        for j=1:c
            M(i,j) = sum(R(i,j) > cumprob2)+1;
        end
    end
end
end
function [S] = autoSample(H, m)
    [k,k] = size(H);
    mu = inv(H)*m;
    exx = inv(H)*(m*m.’ + eye(k))*inv(H.’);
    sigma = exx - mu*mu.’;
    S = mvnrnd(mu,sigma,1);
end
```

We estimate parameters of this example with Algorithm (5), and get estimated transition matrix

$$\hat{A} = \begin{pmatrix} 0.2023 & 0.7977 \\ 0.8026 & 0.1973 \end{pmatrix}.$$

**Algorithm 5** Estimate parameters of Autoregressive HMM model

1. **Estimate of transition matrix** Estimate $A$ with Balm-Welch algorithm.
2. **Fix correlation matrix** If correlation matrix is not positive semi-definite.
3. **Estimate autoregressive coefficients matrix** Estimate $H$ with our Frobenius norm minimization method.
4. **implementation issues**
   - Initialization: Randomly initialize the parameters, use multiple restarts, and pick the best solution.
   - Termination: Set maximum iteration number = 100, Tolerance of convergence $= 1e - 6$.

and two autoregressive coefficients matrix

$$\hat{H}_1 = \begin{pmatrix} 0.0994 & 0 & 0 & 0 & 0 \\ 0.9187 & 0.2040 & 0 & 0 & 0 \\ 0.1018 & 0.5978 & 1.0000 & 0 & 0 \\ 0 & 0.1018 & 0.5978 & 1.0000 & 0 \\ 0 & 0 & 0.1018 & 0.5978 & 1.0000 \end{pmatrix}$$

and

$$\hat{H}_2 = \begin{pmatrix} 0.8053 & 0 & 0 & 0 & 0 \\ 0.2099 & 0.7954 & 0 & 0 & 0 \\ 0.2955 & 0.5031 & 1.0000 & 0 & 0 \\ 0 & 0.2955 & 0.5031 & 1.0000 & 0 \\ 0 & 0 & 0.2955 & 0.5031 & 1.0000 \end{pmatrix}.$$

Errors are still quantified by $\Delta_F = \|H - \hat{H}\|_F^2$. We have

$$\Delta_F^{(1)} = \|H_1 - \hat{H}_1\|_F^2 = 0.0154$$

and

$$\Delta_F^{(2)} = \|H_2 - \hat{H}_2\|_F^2 = 0.0198,$$

both of it are less than 2%. We consider these to be good estimation results.

# Part III

# Real data analysis

# Chapter 8

# Market estimation

In this chapter, we analyze all models using stock index returns. In our study, we use the historical closing index values of the S&P 500 index until October 24, 2012, obtained from Bloomberg L.P. The daily, 1 hour, 30 minute, 5 minute, and 1 minute log-returns are calculated. The size of this data set, 3600 observations, is large enough for ARMA-GARCH model fitting. Smaller sizes of 1200 and 2400 observations are also tested.

## 8.1 Global MLE and quasi MLE estimation

The ARMA(p,q)-GARCH(m,n) model for$\{x_t\}$ is given by

$$\begin{aligned}
&x_t = c + \sum_i^p a_i x_{t-i} + \sum_i^q b_i \epsilon_{t-i} + \epsilon_t \\
&\epsilon_t = \sigma_t u_t, u_t \sim \mathbf{N}(0,1) \\
&\sigma_t^2 = \gamma + \sum_i^m \alpha_i \sigma_{t-i}^2 + \sum_i^m \beta_i \epsilon_{t-i}^2.
\end{aligned} \tag{8.1.1}$$

We estimate parameters using the classical maximum likelihood estimation (MLE) procedure. The log-likelihood of an ARMA-GARCH model in the form of (8.1.1) is given by

$$L(\theta \mid u_1, \ldots, u_T) = \Sigma_{t=1}^{T} \log \left( \frac{u_t \mid \theta}{\sigma_t} \right),$$

where $\theta = (c, a_1, \ldots, a_p, b_1, \ldots, b_q, \gamma, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n)$ is the vector of parameters to be estimated and $f(x)$ is the probability density function of the distribution assumed for $u_t$ with $t = 1, \ldots, T$ . There are two types of MLE we use

- Quasi-MLE: Given an observed univariate time series, estimate the parameters of a conditional mean specification of ARMA form first. The estimation process also infers the residuals $\epsilon_t$ from the input series. then fits the conditional variance specification of GARCH via maximum likelihood to residuals $\epsilon_t$. Thus QMLE requires two maximizations of two different likelihood functions, one for the mean process fit and another for the conditional variance process.

- Global-MLE: only one maximization of the global likelihood function is performed.

- The Matlab garchfit function does global-MLE estimation, our HOMEPAGE/Matlab/armaxfilter.m and HOMEPAGE/Matlab/tarch.m do quasi-MLE estimation.

Table (8.1) shows estimation results of the ARMA(1,1)-GARCH(1,1) for S&P 500 data with sample size 1200. Based on our study, global-MLE has better performance for large sample sizes, while quasi-MLE estimates are more reliable for small data sizes. We use quasi-MLE for our later studies.

Table 8.1: Parameter estimates of ARMA(1,1)-GARCH(1,1) for S&P 500 data with sample size 1200.

| | **c** | **a(AR)** | **b(MA)** | $\gamma$ | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|
| | | | **global MLE** | | | |
| **daily return** | 0.02 | 1.00 | -0.09 | 0.00 | 0.87 | 0.12 |
| **5mins return** | 0.18 | 0.99 | 0.01 | 0.00 | 0.00 | 1.00 |
| **1min return** | 1.86 | 0.87 | 0.57 | 0.00 | 0.40 | 0.57 |
| | | | **quasi MLE** | | | |
| **daily return** | 0.05 | 1.00 | -0.14 | 0.00 | 0.89 | 0.10 |
| **5mins return** | 0.11 | 0.99 | 0.03 | 0.00 | 0.75 | 0.25 |
| **1min return** | 0.03 | 1.00 | 0.10 | 0.00 | 0.91 | 0.09 |

## 8.2 GARCH models

The parameter estimates of ARMA(1,1)-GARCH(1,1) models for S&P 500 data are reported in Table (8.2). Standard deviations are given in parentheses. This table shows different patterns between daily and hourly returns and high frequency returns. We can see daily returns, hourly returns, and 30 minutes returns have unit root phenomena with almost $\alpha = 1$, which means this time series is not stationary. (A time series $x_1, x_2, x_3, \ldots$ is said to be covariance stationary if $\mathbf{E}(x_t)$ and $\mathbf{Cov}(x_t, x_{t+k})$ do not depend on $t$.) The unit root phenomena is especially obvious for 30 minute returns, where $\alpha = 1$. We don't observe unit root phenomena in 5 minute and 1 minute return series for the ARMA(1,1)-GARCH(1,1) model.

We choose three time series with 1200, 2400 and 3600 historical data points before October 24, 2012. Table (8.2) also shows that parameter estimates give similar results for different sample sizes, which means the last 1200 data points of this time series are essential to determine model parameters.

The estimated parameters $\alpha$ and $\beta$ observed in Table (8.1) and Table (8.2) sum up to

Table 8.2: Parameter estimates of ARMA(1,1)-GARCH(1,1) for S&P 500 data.

| | c | a(AR) | b(MA) | $\gamma$ | $\beta$ | $\alpha$ |
|---|---|---|---|---|---|---|
| | Sample size = 1200 | | | | | |
| daily return | -0.0001 | 0.7818 | -0.8186 | 0.0000 | 0.0963 | 0.8363 |
| | (0.0000) | (0.0160) | (0.0127) | (0.0000) | (0.0008) | (0.0026) |
| 1 hour return | 0.0001 | 0.1868 | -0.2959 | 0.0000 | 0.0380 | 0.9507 |
| | (0.0000) | (0.0683) | (0.0577) | (0.0000) | (1.5269e-04) | (1.4315e-04) |
| 30 min return | -0.0001 | -0.2496 | 0.2494 | 0.0000 | 0.0000 | 0.9998 |
| | (0.0000) | (0.7116) | (0.6898) | (0.0000) | (0.0000) | (1.9878e-06) |
| 5 mins return | -0.0001 | -0.4004 | 0.4345 | 0.0000 | 0.3357 | 0.6641 |
| | (0.0000) | (0.5252) | (0.5335) | (0.0000) | (0.0109) | (0.0042) |
| 1 min return | 0.0000 | 0.2312 | -0.1691 | 0.0000 | 0.5261 | 0.0581 |
| | (0.0000) | (0.0426) | (0.0445) | (0.0000) | (0.1102) | (0.6504) |
| | Sample size = 2400 | | | | | |
| daily return | -0.0001 | 0.7103 | -0.7518 | 0.0000 | 0.0762 | 0.9181 |
| | (0.0000) | (0.0222) | (0.0209) | (0.0000) | (1.7174e-04) | (1.9106e-04) |
| 1 hour return | -0.0000 | -0.10 | -0.1013 | 0.0000 | 0.0225 | 0.9746 |
| | (0.000) | (6.755e-04) | (5.510e-04) | (0.0000) | (2.6664e-05) | (2.8123e-05) |
| 30 min return | -0.0000 | 0.3225 | -0.3017 | 0.0000 | 0.0079 | 0.9915 |
| | (0.0000) | (0.0186) | (0.0185) | (0.0000) | (4.4229e-06) | (3.1188e-06) |
| 5 mins return | -0.0000 | -0.8436 | 0.8556 | 0.0000 | 0.3482 | 0.6516 |
| | (0.0000) | (0.0924) | (0.0802) | (0.0000) | (0.0093) | (0.0029) |
| 1 min return | -0.0000 | 0.2008 | -0.1249 | 0.0000 | 0.4982 | 0.1858 |
| | (0.0000) | (0.3282) | (0.3420) | (0.0000) | (0.0491) | (0.1046) |
| | Sample size = 3600 | | | | | |
| daily return | -0.0001 | 0.5223 | -0.6015 | 0.0000 | 0.0872 | 0.9056 |
| | (0.0000) | (0.0506) | (0.0446) | (0.0000) | (9.3159e-05) | (1.0331e-04) |
| 1 hour return | -0.0001 | -0.9433 | 0.9256 | 0.0000 | 0.0186 | 0.9772 |
| | (0.0000) | (0.0019) | (0.0027) | (0.0000) | (1.517e-05) | (2.011e-05) |
| 30 min return | -0.0000 | 0.3124 | -0.2842 | 0.0000 | 0.0070 | 0.9922 |
| | (0.0000) | (0.1177) | (0.1195) | (0.0000) | (2.6991e-06) | (2.8575e-06) |
| 5 mins return | -0.0000 | -0.7744 | 0.7938 | 0.0000 | 0.3054 | 0.6287 |
| | (0.0000) | (0.0339) | (0.0298) | (0.0000) | (0.0105) | (0.0091) |
| 1 min return | 0.0000 | 0.3954 | -0.3163 | 0.0000 | 0.5387 | 0.1716 |
| | (0.0000) | (0.0166) | (0.0134) | (0.0000) | (0.0322) | (0.0440) |

values close to 1. Based on this observation, we apply the integrated GARCH(1,1), which is named IGARCH(1,1). IGARCH is a process for which

$$\gamma > 0$$

and

$$\sum_{i=1}^{m} \alpha_i + \sum_{i=1}^{n} \beta_i = 1.$$

Table (8.3) shows that the IGARCH(1,1) model captures the integrated feature.

Table 8.3: Parameter estimates of ARMA(1,1)-IGARCH(1,1) for S&P 500 data

| | c | a(AR) | b(MA) | $\gamma$ | $\beta$ | $\alpha$ |
|---|---|---|---|---|---|---|
| | Sample size = 1200 | | | | | |
| daily return | 0.01 | 1.00 | -0.01 | 0.00 | 0.12 | 0.89 |
| 1 hour return | 0.07 | 0.99 | -0.11 | 0.00 | 0.05 | 0.95 |
| 30 min return | 0.06 | 1.00 | 0.01 | 0.00 | 0.00 | 1.00 |
| 5 mins return | -0.02 | 1.00 | 0.03 | 0.00 | 0.33 | 0.67 |
| 1 min return | 0.06 | 1.00 | 0.06 | 0.00 | 0.66 | 0.34 |
| | Sample size = 2400 | | | | | |
| daily return | 0.02 | 1.00 | -0.03 | 0.00 | 0.08 | 0.92 |
| 1 hour return | 0.06 | 1.00 | -0.04 | 0.00 | 0.03 | 0.97 |
| 30 min return | 0.01 | 1.00 | 0.02 | 0.00 | 0.01 | 0.99 |
| 5 mins return | -0.00 | 1.00 | 0.01 | 0.00 | 0.35 | 0.65 |
| 1 min return | 0.01 | 1.00 | 0.07 | 0.00 | 0.57 | 0.43 |
| | Sample size = 3600 | | | | | |
| daily return | 0.03 | 1.00 | -0.08 | 0.00 | 0.09 | 0.91 |
| 1 hour return | 0.04 | 1.00 | -0.02 | 0.00 | 0.02 | 0.98 |
| 30 min return | 0.01 | 1.00 | 0.03 | 0.00 | 0.01 | 0.99 |
| 5 mins return | 0.03 | 1.00 | 0.02 | 0.00 | 0.36 | 0.64 |
| 1 min return | 0.02 | 1.00 | 0.08 | 0.00 | 0.60 | 0.40 |

Since the clear distinction between GARCH and IGARCH models has been criticized, we consider the generalized fractional integrated GARCH(FIGARCH) model and corresponding mean process FARIMA to capture fractional features of a time-series of index returns. FARIMA processes are more specifically ARIMA(p, d, q) process with $0 < |d| < 0.5$, that satisfy difference equations of the form

$$(1-L)^d a(L)x_t = b(L)\epsilon_t$$

where $a(L)$ and $b(L)$ are polynomials of degree $p$ and $q$, respectively, satisfying, $a(z) \neq 0$, $b(z) \neq 0$, for all $z$ such that $|z| \leq 1$. L is the lag operator, and $\epsilon_t$ is a white noise sequence with mean 0 and finite variance $\sigma^2$.

The conditional variance, $h_t$, of a FIGARCH(p, d, q) process is modeled as follows:

$$\sigma_t^2 = \omega + [1 - \beta(L) - \phi(L)(1-L)^d]\epsilon_t^2 + \beta(L)\sigma_t.$$

The parameter estimates of FARIMA(1,1)-FIGARCH(1,1) for S&P 500 data are reported in Table (8.4). Standard deviations are given in parentheses. This table shows that for a FARIMA(1,1) mean process, every time series has fractional $d$. For a FIGARCH(1,1) variance process, daily, hourly, 30 minute, and 1 minute returns have fractional $d$, and only 5 minute returns have $d = 1$.

We also observe negative $d$ in FARIMA, which can be explained by short memory. The long memory can be empirically observed, e.g. by a slowly decaying auto-covariance function (ACF) (Beran [1994]). The classic example of a long-range dependent process is the fractional autoregressive integrated moving average (FARIMA) model with a power-law ACF. It appears that the values of FARIMA with Gaussian noise, for the memory parameter $d$ greater than 0, have such a slowly decaying ACF that it is not absolutely summable. This behavior serves as a classical definition of the long-range dependence (Beran [1994]). When

$d < 0$, the ACF still follows a power law, hence exhibiting more significant dependence than any other process with exponentially decaying ACF, such as, e.g. an autoregressive moving average (ARMA) time series, but the rate of decay is slower than for the $d$-positive case making the ACF absolutely summable. This negative memory phenomenon can be described as follows: increases in the values of the time series are likely to be followed by decreases and, conversely, decreases are more likely to be followed by increases (negative correlation). Such a time series is said to have short memory.

After fitting GARCH style models to time series, we examine the innovations. We fit classic tempered stable (CTS) distributions to the innovations inferred from ARMA(1,1)-GARCH(1,1).

Let $\alpha \in (0,1) \bigcup (1,2)$, $C$, $\lambda_+$, $\lambda_- > 0$, and $m \in R$. X is said to have the classic tempered stable (CTS) distribution if the characteristic function of X is given by

$$\begin{aligned}\phi_x(u) &= \phi_{CTS}(u; \alpha, C, \lambda_+, \lambda_-, m)\\ &= \exp(ium - iuC\Gamma(1-\alpha)(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1})\\ &+ C\Gamma(-\alpha)((\lambda_+ - iu)^\alpha - \lambda_+^\alpha + (\lambda_- + iu)^\alpha - \lambda_-^\alpha)),\end{aligned}$$

and we denote $X \sim CTS(\alpha, C, \lambda_+, \lambda_-, m)$. Table (8.5) presents parameter estimates.

Table 8.4: Parameter estimates of FARIMA(1,d,1)-FIGARCH(1,d,1) for S&P 500 data

| | d | a(AR) | b(MA) | $\omega$ | $\phi$ | d | $\beta$ |
|---|---|---|---|---|---|---|---|
| | Sample size = 1200 | | | | | | |
| daily return | -0.6611 | 0.9629 | -0.3293 | 0.0000 | 0.0486 | 0.3465 | 0.3075 |
| | (0.0289) | (0.0289) | (0.0289) | (0.0000) | (0.1258) | (0.0258) | (0.2019) |
| 1 hour return | -0.5174 | 0.9299 | -0.5010 | 0.0000 | 0.0965 | 0.8070 | 0.9035 |
| | (0.0283) | (0.0283) | (0.0283) | (0.0000) | (0.0020) | (0.0038) | (0.0006) |
| 30 min return | -0.6483 | 0.9101 | -0.2695 | 0.0000 | 0.4327 | 0.1346 | 0.5673 |
| | (0.0289) | (0.0289) | (0.0289) | (0.0000) | (0.0194) | (0.0009) | (0.0250) |
| 5 mins return | -0.5945 | 0.9661 | -0.3741 | 0.0000 | 0.0000 | 1.0000 | 0.5939 |
| | (0.0289) | (0.0289) | (0.0289) | (0.0000) | (0.0000) | (0.0125) | (0.0036) |
| 1 min return | -0.6784 | 0.9304 | -0.1943 | 0.0000 | 0.4605 | 0.0791 | 0.0000 |
| | (0.0289) | (0.0289) | (0.0289) | (0.0000) | (0.0183) | (0.0436) | (0.0000) |
| | Sample size = 2400 | | | | | | |
| daily return | -0.4549 | 0.9049 | -0.5049 | 0.0000 | 0.1424 | 0.4175 | 0.5218 |
| | (0.0204) | (0.0204) | (0.0204) | (0.0000) | (0.0040) | (0.0048) | (0.0075) |
| 1 hour return | -0.3679 | 0.8837 | -0.5807 | 0.0000 | 0.0429 | 0.9141 | 0.9571 |
| | (0.0204) | (0.0204) | (0.0204) | (0.0000) | (0.0010) | (0.0015) | (0.0002) |
| 30 min return | -0.4329 | 0.8715 | -0.4305 | 0.0000 | 0.4163 | 0.1674 | 0.5837 |
| | (0.0204) | (0.0204) | (0.0204) | (0.0000) | (0.0025) | (0.0055) | (0.0153) |
| 5 mins return | -0.3534 | 0.9063 | -0.5919 | 0.0000 | 0.0000 | 1.0000 | 0.6787 |
| | (0.0218) | (0.0218) | (0.0218) | (0.0000) | (0.0000) | (0.0208) | (0.0023) |
| 1 min return | -0.4061 | 0.8387 | -0.3554 | 0.0000 | 0.1933 | 0.2852 | 0.0000 |
| | (0.0204) | (0.0204) | (0.0204) | (0.0000) | (0.0135) | (0.0107) | (0.0000) |
| | Sample size = 3600 | | | | | | |
| daily return | -0.3313 | 0.9104 | -0.7107 | 0.0000 | 0.0430 | 0.7667 | 0.7805 |
| | (0.0167) | (0.0167) | (0.0167) | (0.0000) | (0.0006) | (0.0030) | (0.0014) |
| 1 hour return | -0.3260 | 0.8655 | -0.5962 | 0.0000 | 0.0454 | 0.9091 | 0.9546 |
| | (0.0182) | (0.0182) | (0.0182) | (0.0000) | (0.0007) | (0.0018) | (0.0003) |
| 30 min return | -0.3085 | 0.7959 | -0.4744 | 0.0000 | 0.4174 | 0.1653 | 0.5826 |
| | (0.0167) | (0.0167) | (0.0167) | (0.0000) | (0.0046) | (0.0044) | (0.0195) |
| 5 mins return | -0.3227 | 0.8689 | -0.5581 | 0.0000 | 0.0000 | 1.0000 | 0.6517 |
| | (0.0188) | (0.0188) | (0.0188) | (0.0000) | (0.0000) | (0.0293) | (0.0126) |
| 1 min return | -0.3186 | 0.8058 | -0.4072 | 0.0000 | 0.1584 | 0.2438 | 0.0000 |
| | (0.0167) | (0.0167) | (0.0167) | (0.0000) | (0.0104) | (0.0050) | (0.0000) |

Table 8.5: Fit standard CTS to innovations of ARMA-GARCH with sample size 3600

| | $\alpha$ | $\lambda_+$ | $\lambda_-$ |
|---|---|---|---|
| daily return | 0.0001 | 2.3900 | 2.1617 |
| 1 hour return | 1.1908 | 0.2620 | 0.2640 |
| 30 min return | 1.1057 | 0.2642 | 0.2750 |
| 5 mins return | 1.3754 | 0.2851 | 0.2123 |
| 1 min return | 0.7627 | 0.7431 | 0.6944 |

# Chapter 9

# Out-of-Sample performance

One of the essential objectives of financial modeling is forecasting. The ARMA-GARCH style models studied in Chapter (8) are some of the most popular for univariate forecasting. Based on our observations in Chapter (8), the white noise series $u_t$ are not i.i.d. $\mathbf{N}(0, 1)$ in high frequency data. Correlations occur within the white noise series, thus we improve ARMA-GARCH models by introducing our new autoregressive regime-switching model into white noise series, which is named ARMA-GARCH with autoregressive regime-switching noise model. We forecast VaR models with both both original ARMA-GARCH and our new ARMA-GARCH with autoregressive regime-switching noises, then compare performance with Bernoulli and Berkowitz tests.

## 9.1 Forecasting

The steps for forecasting via ARMA-GARCH models of the form:

$$
\begin{aligned}
& x_t = c + \sum_i^p a_i x_{t-i} + \sum_i^q b_i \epsilon_{t-i} + \epsilon_t \\
& \epsilon_t = \sigma_t u_t, u_t \sim \mathbf{N}(0,1) \\
& \sigma_t^2 = \gamma + \sum_i^m \alpha_i \sigma_{t-i}^2 + \sum_i^m \beta_i \epsilon_{t-i}^2,
\end{aligned}
$$

are as follows:

1. Estimate parameters and corresponding $\epsilon_t$, $\sigma_t$ and $u_t$, where $t = 1, \ldots, h$.
2. Get $\sigma_{h+1}^2 = \hat{\gamma} + \hat{\alpha}\sigma_h^2 + \hat{\beta}\epsilon_h^2$.
3. Generate 100 random variable $u_{h+1} \sim \mathbf{N}(0,1)$.
4. Get $\epsilon_{h+1} = \sigma_{h+1} u_{h+1}$.
5. Get $x_{h+1} = \hat{c} + \hat{a} x_h + \hat{b}\epsilon_h + \epsilon_{h+1}$.
6. Move estimation window one step forward, estimate parameters and corresponding $\epsilon_t$, $\sigma_t$ and $u_t$, where $t = 2, \ldots, h+1$.
7. Loop is closed.

The steps for forecasting via ARMA-GARCH models with autoregressive regime-switching noises model are:

1. Estimate parameters and corresponding $\epsilon_t$, $\sigma_t$ and $u_t$, where $t = 1, \ldots, h$.
2. Fit autoregressive regime-switching model to noise series $u_t$, get autoregressive coefficients matrix $\{\hat{H}_1, \hat{H}_2\}$ and state transition matrix $\hat{A}$.

3. Simulate noise process $\tilde{u}_t$ with estimated autoregressive coefficients matrix $\{\hat{H}_1, \hat{H}_2\}$ and state transition matrix $\hat{A}$.

4. Get $\sigma_{h+1}^2 = \hat{\gamma} + \hat{\alpha}\sigma_h^2 + \hat{\beta}\epsilon_h^2$.

5. Get $\epsilon_{h+1} = \sigma_{h+1}\tilde{u}_{h+1}$.

6. Get $x_{h+1} = \hat{c} + \hat{a}x_h + \hat{b}\epsilon_h + \epsilon_{h+1}$.

7. Move estimation window one step forward, estimate parameters and corresponding $\epsilon_t$, $\sigma_t$ and $u_t$, where $t = 2, \dots, h+1$.

8. Loop is closed.

## 9.2 Backtesting

Backtesting aims to take ex ante value-at-risk (VaR) forecasts from a particular model and compare them with ex post realized returns (i.e., historical observations). Whenever losses exceed VaR, a VaR violation is said to have occurred. There are several methods to backtest models. We discuss the binomial and Berkowitz tests here.

### 9.2.1 Bernoulli test

The specific notation used in this section is:

- $W_E$ is estimation window size;
- $T$ is number of observations in a sample;
- $\eta_t$ indicates whether a VaR violation occurs (i.e. $\eta = 1$);

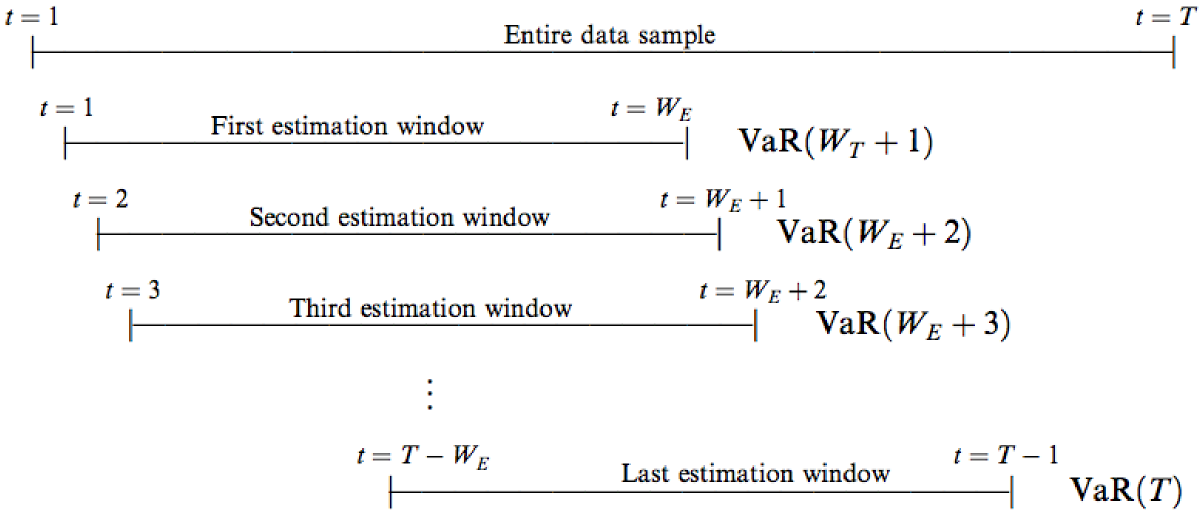


Figure 9.2.1: Compute VaR.

- $\nu_i, i = 0, 1$ is number of violations ( $i = 1$ ) and number of no violation ( $i = 0$ ) observed.

We estimate parameters of the model from first estimation window $W_E$, then forecast VaR for day $E + 1$, The estimation window is then moved forward by one step to get the risk forecast for day $E + 2$. The estimation window is moved forward by one day until $T - 1$ (Figure 9.2.1), then we have $T - E$ VaR forecasts. As the data from day $E + 1$ to day $T$ are already known, VaR forecasts can be compared with the actual outcome. If the actual return on a particular day exceeds the VaR forecast percentile limit, then the VaR limit is said to have been violated. We denote the violations as $\eta_t$, which has the value 1 when a violation occurs and 0 when a violation doesn't occur. The number of violations are stored in the variable $\nu_1$ and $\nu_0$, where $\nu_1$ is the number of days with violations and $\nu_0$ is the number of days without violations(Danielsson [2011]).

We then use the Bernoulli coverage test to find out the proportion of violations. The null hypothesis for VaR violations is:

$$H0 : \eta \sim \mathbf{B}(p)$$

where B stands for the Bernoulli distribution.

Table 9.1: Comparison of ARMA-GARCH(model 1) and HMM-autoregressive noise model(model 2) via Bernoulli test.

| | Bernoulli test | |
|---|---|---|
| | Test statistics | p-value |
| | daily return | |
| Model 1 | 7.3524 | 0.0715 |
| Model 2 | 5.2749 | 0.2527 |
| | 1 hour return | |
| Model 1 | 2.7773 | 0.5273 |
| Model 2 | 6.5951 | 0.0960 |
| | 5 minute return | |
| Model 1 | 4.3962 | 0.3217 |
| Model 2 | 3.8844 | 0.3742 |
| | 1 minute return | |
| Model 1 | 10.2653 | 0.0264 |
| Model 2 | 0.7875 | 0.7525 |

The Bernoulli tests for standard ARMA(1,1)-GARCH(1,1) and ARMA(1,1)-GARCH(1,1) with regime-switching noise forecasting VaR for S&P 500 data are reported in Table (9.1). The length of estimation window is 1000, the forecasting horizon is 1000 steps, and number of paths is 200. We can see for daily, hourly, and 5 minute returns, both models have $p-$values larger than 5 percent . For 1 minute returns, the $p-$value for the standard ARMA(1,1)-GARCH(1,1) model is 0.0264, which means that the null hypothesis is rejected, while the $p-$value of our HMM-autoregressive noise model is 0.7525, which is not rejected.

### 9.2.2 Berkowitz test

A test would be needed to verify the $i-$th observed return $r_i$ follows the predicted distribution $P_i$ from the model. The problem is we have only one observed return $r_i$ for each sliding-window. We carry out a probability integral transform PIT transformation and map each $r_i$ to its percentile point on its forecasted density function. For example, some observed return is equal to the $30th$ percentile on the forecasted density, then its value maps to 0.3. Under the null hypothesis that the model is adequate, the mapped value $\hat{r}_i = P_i^{-1}(r_i)$ follows a uniform distribution $U(0,1)$. So we can evaluate the model by using Kolmogorov's test to test $\hat{r}_i$ is uniform distributed. However, we use the Berkowitz test to perform the transformation $\hat{\hat{r}}_i \phi^{-1}(\hat{r}_i) \sim N(0,1)$. Then we can use any test for normality over $\hat{\hat{r}}_i$ to evaluate our model(Lobato et al. [2007] Christodoulakis et al. [2007]).

In practice, daily forecasts can be obtained with the following procedure:

1. Estimate parameters of the model from first estimation window $W_E$, then forecast return for day $E+1$.

2. Carry out a probability integral transform (PIT) transformation of the forecasted return for day $E+1$.

3. Map observed return $r_{E+1}$ for day $E+1$ to its percentile point on its forecasted density function $P_{E+1}$.

4. The estimation window is then moved up by one day to obtain the forecasted density function $P_{E+2}$.

5. The estimation window is moving forward by a step of one day until $T-1$ Figure (9.2.1), then we have $T-E$ forecasted density function.

6. Carry out Berkowitz test for paired observed returns $r_i$ and $P_i$ with $i = \{E+1, \dots, T\}$ to see if $r_i$ follows $P_i$.

Table 9.2: Comparison of ARMA-GARCH(model 1) and HMM-autoregressive noise model(model 2) via Berkowitz test.

| | Berkowitz test | |
|---|---|---|
| | Test statistics | p-value |
| | daily return | |
| Model 1 | 7.3524 | 0.0615 |
| Model 2 | 5.2749 | 0.1527 |
| | 1 hour return | |
| Model 1 | 2.7773 | 0.4273 |
| Model 2 | 6.5951 | 0.0860 |
| | 5 minute return | |
| Model 1 | 4.3962 | 0.2217 |
| Model 2 | 3.8844 | 0.2742 |
| | 1 minute return | |
| Model 1 | 10.2653 | 0.0164 |
| Model 2 | 0.7875 | 0.8525 |

The Berkowitz test for standard ARMA(1,1)-GARCH(1,1) and ARMA(1,1)-GARCH(1,1) with regime-switching noise forecasting VaR for S&P 500 data are reported in Table (9.2). The length of estimation window is 1000, the forecasting horizon is 100 steps, and number of paths is 200. We can see for daily return, hourly return and 5 minutes return, both models have $p-$values larger than 5 percent, for high frequency data 1 minute return, the $p-$value of standard ARMA(1,1)-GARCH(1,1) is 0.0164, which means null hypothesis is rejected, while the $p-$value of our HMM-autoregressive noise model is 0.8525, which is not rejected.

# Chapter 10

# Conclusion

Our research starts with implementing a two-state stochastic volatility model. This model can be used to produce general option pricing tools which reflect the true volatility structure of financial markets more accurately. We apply the dynamic programming method for general discrete-time mean-variance hedging problem to this case. It works in a manner which is statistically parsimonious and computationally efficient. It can also price American options. Dynamic programming allows us to go backward from the expiry date and decide the value at each node. An improved hash table has also been used to improve computation speed. After we get the pricer, more numerical tests are performed to prove the limiting process has a constant volatility which is a weighted-sum of two original volatilities from math deduction and simulation results.

To improve this two-state stochastic volatility model, we add an autoregression component, and extend the HMM driven model to an autoregressive HMM driven model. To estimate autoregressive HMM driven model, we start from examining the existing autoregressive HMM model and ascertain that an essential assumption is that autoregressive order is greatly less than the length of observation. Without this assumption, the approximation for p.d.f. function doesn't hold. In our case, the autoregressive order is about one week (5 days), and length of observation is about two weeks. As a result, the assumption for approximation doesn't hold in our case. So we develop our own estimation method, test it's stability and precision, and ensure it is parsimonious and won't fail when the correlation matrix is positive definite.

We then attempt to incorporate autoregressive HMM driven models into GARCH style models to generate better backtesting results than existing models. We estimate and test ARMA-GARCH, ARIMA-IGARCH, and FARIMA-FIGARCH first and find that the white noise series is not iid Gaussian. Thus we apply the autoregressive HMM driven model to the white noise series and test it's forecasting effect. The results indicate that standard ARMA-GARCH and our autoregressive- HMM-noises model can both performs good in daily S&P 500 log returns, while autoregressive- HMM-noise model can do better in high frequency data.

# Bibliography

A.B. Abel. Exact solutions for expected rates of return under markov regime switching: Implications for the equity premium puzzle. Technical report, National Bureau of Economic Research, 1992.

A.B. Abel. Risk premia and term premia in general equilibrium. *Journal of Monetary Economics*, 43(1):3–33, 1999.

D.D. Aingworth, S.R. Das, and R. Motwani. A simple approach for pricing equity options with Markov switching state variables. *Quantitative Finance*, 6(2):95–105, 2006.

A. Ang and G. Bekaert. International asset allocation with regime shifts. *Review of Financial studies*, 15(4):1137–1187, 2002.

O.E. Barndorff-Nielsen and N. Shephard. Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241, 2001.

G. Bekaert and R.J. Hodrick. Characterizing predictable components in excess returns on equity and foreign exchange markets. *The Journal of Finance*, 47(2):467–509, 2012.

J. Beran. *Statistics for long-memory processes*, volume 61. Chapman & Hall/CRC, 1994.

N.P.B. Bollen, S.F. Gray, and R.E. Whaley. Regime switching in foreign exchange rates::: Evidence from currency option prices. *Journal of Econometrics*, 94(1-2):239–276, 2000.

RL Brown and J. Durbin. Methods of investigating whether a regression relationship is constant over time. In *European Statistical Meeting, Amsterdam*, 1968.

J. Cai. A markov model of unconditional variance in arch. *Journal of Business and Economic Statistics*, 12(3):309–316, 1994.

L.E. Calvet and A.J. Fisher. How to forecast long-run volatility: regime switching and the estimation of multifractal processes. *Journal of Financial Econometrics*, 2(1):49–83, 2004.

M. Carrasco, L. Hu, and W. Ploberger. Optimal test for markov switching. 2004.

S.G. Cecchetti, P.S. Lam, and N. Mark. Mean reversion in equilibrium asset prices, 1990.

S.G. Cecchetti, P. Lam, and N.C. Mark. The equity premium and the risk-free rate: Matching the moments. *Journal of Monetary Economics*, 31(1):21–45, 1993.

A. Cernỳ. Dynamic programming and mean-variance hedging in discrete time. *Applied Mathematical Finance*, 11(1):1–25, 2004.

A. Černỳ. *Mathematical techniques in finance: tools for incomplete markets*. Princeton Univ Pr, 2009.

T. Chartier and A. Greenbaum. Richardson's extrapolation. http://www.math.washington.edu/~greenbau/Math_498/lecture04_richardson.pdf, 2008.

G. Christodoulakis, G.A. Christodoulakis, and S. Satchell. *The analytics of risk model validation*. Academic Press, 2007.

R. Cont and P. Tankov. *Financial modelling with jump processes*, volume 2. Chapman & Hall/CRC, 2003.

S.R. Cosslett and L.F. Lee. Serial correlation in latent discrete variable models. *Journal of Econometrics*, 27(1):79–97, 1985.

Q. Dai, K.J. Singleton, and W. Yang. Regime shifts in a dynamic term structure model of us treasury bond yields. *Review of Financial Studies*, 20(5):1669–1706, 2007.

J. Danielsson. Financial risk forecasting, 2011.

D. Duffie, D. Filipovic, and W. Schachermayer. Affine processes and applications in finance. *Ann. Appl. Probab*, 13(3):984–1053, 2003.

H. Dym. *Linear algebra in action*. Amer Mathematical Society, 2007. ISBN 082183813X.

R.J. Elliott, L. Chan, and T.K. Siu. Option pricing and esscher transform under regime switching. *Annals of Finance*, 1(4):423–432, 2005.

C. Engel. Can the markov switching model forecast exchange rates? Technical report, National Bureau of Economic Research, 1994.

C. Engel and J.D. Hamilton. Long swings in the exchange rate: Are they in the data and do markets know it? Technical report, National Bureau of Economic Research, 1989.

J.U. Farley and M.J. Hinich. A test for a shifting slope coefficient in a linear model. *Journal of the American Statistical Association*, 65(331):1320–1329, 1970.

N. Fiess and R. Shankar. Determinants of exchange rate regime switching. *Journal of International Money and Finance*, 28(1):68–98, 2009.

J.P. Fouque, G. Papanicolaou, and K.R. Sircar. *Derivatives in financial markets with stochastic volatility*. Cambridge Univ Pr, 2000.

R.J. Frey. Hidden Markov Models with Univariate Gaussian Outcomes, 2010.

M. Frömmel, R. MacDonald, and L. Menkhoff. Markov switching regimes in a monetary exchange rate model. *Economic Modelling*, 22(3):485–502, 2005.

R. Garcia. Asymptotic null distribution of the likelihood ratio test in markov switching models. *International Economic Review*, pages 763–788, 1998.

R. Garcia and P. Perron. An analysis of the real interest rate under regime shifts. *The Review of Economics and Statistics*, pages 111–125, 1996.

S.M. Goldfeld and R.E. Quandt. A markov model for switching regressions. *J. Econom.*, 1: 3–16, 1973.

S.M. Goldfeld and R.E. Quandt. A markov model for switching regressions. *Journal of econometrics*, 1(1):3–15, 2010.

S.F. Gray. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27–62, 1996.

M. Guidolin and A. Timmermann. Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, 31(11):3503–3544, 2007.

M. Guidolin and A. Timmermann. International asset allocation under regime switching, skew, and kurtosis preferences. *Review of Financial Studies*, 21(2):889–935, 2008.

J.D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384, 1989.

J.D. Hamilton. Analysis of time series subject to changes in regime. *Journal of econometrics*, 45(1):39–70, 1990.

J.D. Hamilton and G. Perez-Quiros. What do the leading indicators lead? *Journal of Business*, pages 27–49, 1996.

J.D. Hamilton and R. Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1):307–333, 1994.

B.E. Hansen. The likelihood ratio test under nonstandard conditions: testing the markov switching model of gnp. *Journal of applied Econometrics*, 7(S1):S61–S82, 2006.

S.L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of financial studies*, 6(2):327–343, 1993.

A. Hordijk, D.L. Iglehart, and R. Schassberger. Discrete time methods for simulating continuous time Markov chains. *Advances in Applied Probability*, 8(4):772–788, 1976. ISSN 0001-8678.

J. Hull. *Options, futures and other derivatives.* Pearson Prentice Hall, 2009.

J. HULL and A. WHITE. The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2):281–300, 1987.

K.R. Jackson, S. Jaimungal, and V. Surkov. Option pricing with regime switching Lévy processes using Fourier space time stepping. In *Proceeding of the Fourth IASTED International Conference on Financial Engineering and Applications*, pages 92–97. Citeseer, 2007.

H. Johnson and D. Shanno. Option pricing when the variance is changing. *Journal of Financial and Quantitative Analysis*, 22(2):143–151, 1987.

B.H. Juang and L. Rabiner. Mixture autoregressive hidden markov models for speech signals. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(6):1404–1413, 1985.

A. Kabašinskas, S.T. Rachev, L. Sakalauskas, W. Sun, and I. Belovas. Alpha-stable paradigm in financial markets. 2008.

D. Karlis and E. Xekalaki. Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590, 2003.

Y.S. Kim. Option pricing and hedging under a stochastic volatility Lévy process model. 2011.

Y.S. Kim, S.T. Rachev, M.L. Bianchi, and F.J. Fabozzi. Financial market models with lévy processes and time-varying volatility. *Journal of Banking & Finance*, 32(7):1363–1378, 2008.

Y.S. Kim, S.T. Rachev, D.M. Chung, and M.L. Bianchi. The modified tempered stable distribution, garch models and option pricing. *Probability and Mathematical Statistics, to appear*, 2009.

M. Lettau, S.C. Ludvigson, and J.A. Wachter. The declining equity premium: What role does macroeconomic risk play? Technical report, National Bureau of Economic Research, 2004.

Z. Lin, Y.S. Kim, S.T. Rachev, M.L. Bianchi, and F.J. Fabozzi. Option pricing with regime-switching tempered stable processes, 2010.

G. Lindgren. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, pages 81–91, 1978.

RH Liu, Q. Zhang, and G. Yin. Option pricing in a regime-switching model using the fast Fourier transform. *Journal of Applied Mathematics and Stochastic Analysis*, 2006(6), 2006. ISSN 1048-9533.

J.M.H. Lobato, D.H. Lobato, and A. Suárez. Time series models for measuring market risk technical report. 2007.

D.G. Luenberger. *Investment science*. Oxford University Press New York, 1998.

CR MacCluer. The many proofs and applications of Perron's theorem. *SIAM review*, 42(3): 487–498, 2000. ISSN 0036-1445.

G.J. McLachlan and D. Peel. *Finite mixture models*, volume 299. Wiley-Interscience, 2000.

C. Menn and S.T. Rachev. A garch option pricing model with [alpha]-stable innovations. *European journal of operational research*, 163(1):201–209, 2005a.

C. Menn and S.T. Rachev. Smoothly truncated stable distributions, garch-models, and option pricing. *University of Karlsruhe and UCSB. Retrieved on March*, 6:2009, 2005b.

EC Mike, P. So, K. Lam, and WK Li. A stochastic volatility model with markov switching. *Journal of Business & Economic Statistics*, 16(2):244–253, 1998.

S. Mittnik and S.T. Rachev. Stable Paretian models in finance, 2001.

S.N. Neftci. Are economic time series asymmetric over the business cycle? *The Journal of Political Economy*, pages 307–328, 1984.

D.B. Nelson. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370, 1991.

E. Nicolato and E. Venardos. Option pricing in stochastic volatility models of the ornstein-uhlenbeck type. *Mathematical finance*, 13(4):445–466, 2003.

J.R. Norris. *Markov chains*. Number 2008. Cambridge Univ Pr, 1998.

R.E. Quandt. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 53(284):873–880, 1958.

R.E. Quandt. A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67(338):306–310, 1972.

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, 53(3):267–296, 1990.

S. Rachev and S. Mittnik. Stable paretian models in finance. 2000.

S.T. Rachev, S. Mittnik, F.J. Fabozzi, S.M. Focardi, T. Jašić, et al. *Financial econometrics: from basics to advanced modeling techniques*, volume 150. Wiley, 2006.

S.T. Rachev, Y.S. Kim, M.L. Bianchi, F.J. Fabozzi, et al. *Financial models with Lévy processes and volatility clustering*, volume 187. Wiley, 2011.

E. Renault and N. Touzi. Option hedging and implied volatilities in a stochastic volatility model. *Mathematical Finance*, 6(3):279–302, 1996.

A.K.M.E. Saleh. *Theory of preliminary test and Stein-type estimation with applications*, volume 517. Wiley-Interscience, 2006.

M. Schmelzle. Option pricing formulae using fourier transform: Theory and application. *Available at pfadintegral. com/articles*, 2010.

G.W. Schwert. Why does stock market volatility change over time?, 1990.

C.A. Sims and T. Zha. Were there regime switches in us monetary policy? *The American Economic Review*, 96(1):54–81, 2006.

E.M. Stein and J.C. Stein. Stock price distributions with stochastic volatility: an analytic approach. *Review of financial Studies*, 4(4):727–752, 1991.

R.S. Tsay. *Analysis of financial time series*, volume 543. Wiley-Interscience, 2005.

J. Tu. Is regime switching in stock returns important in portfolio decisions? *Management Science*, 56(7):1198–1215, 2010.

C.M. Turner, R. Startz, and C.R. Nelson. A markov model of heteroskedasticity, risk, and learning in the stock market. *Journal of Financial Economics*, 25(1):3–22, 1989.

P. Veronesi. Stock market overreactions to bad news in good times: a rational expectations equilibrium model. *Review of Financial Studies*, 12(5):975–1007, 1999.

R.F. Whitelaw. Stock market risk and return: An equilibrium approach. *Review of Financial Studies*, 13(3):521–547, 2000.

J.B. Wiggins. Option values under stochastic volatility: theory and empirical estimates. *Journal of financial economics*, 19(2):351–372, 1987.

# Appendix A

# Mathematical notations

- $Z$ is a random variable with standard normal distribution $\mathbb{N}(0, 1)$.
- $S$ is stock price.
- $r$ is risk-free interest rate.
- $N$ is number of time steps.
- $T$ is expiry.
- $\Delta t = \frac{T}{N}$ is time step.
- $\{\sigma_1,\, \sigma_2\}$ are two states of stock price volatility.
- $\Omega$ is a set of parameters.
- $X_t$ is continuous time hidden Markov chain process.
- $M$ is number of states.
- $A = \begin{bmatrix} p_1 & 1 - p_1 \\ 1 - p_2 & p_2 \end{bmatrix}$ is the Markov matrix which drives volatility process.
- $\{\pi_1,\, \pi_2\}$ is equilibrium distribution for matrix $A$.
- $\{q_1,\, q_2\}$ are the probabilities for stock price to move up with respect to $\{\sigma_1,\, \sigma_2\}$ seperately.
- $\mu$ is the expected rate of return for stock.
- $dW$ is standard Wiener process with a drift rate of zero and variance of 1.
- $\sigma_{imp}$ is implied volatility.

- $T$ is sample size of observations.
- $K$ is length of observation window.
- $p$ is order of autoregression.
- $\vec{s} = \{s_1, s_2, \dots, s_K\}$ is sequence of observations.
- $\vec{x} = \{x_1, x_2, \dots, x_K\}$ is sequence of normalized observations with $s_i = x_i\delta$.
- $\Sigma_x$ is the covariance matrix of $\vec{x}$.