University of Texas at El Paso
ScholarWorks@UTEP

Open Access Theses & Dissertations

2020-01-01

Predicting Stochastic Volatility For Extreme Fluctuations In High Frequency Time Series

Md Al Masum Bhuiyan
University of Texas at El Paso

Follow this and additional works at: https://scholarworks.utep.edu/open_etd

Part of the Applied Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation

Bhuiyan, Md Al Masum, "Predicting Stochastic Volatility For Extreme Fluctuations In High Frequency Time Series" (2020). *Open Access Theses & Dissertations*. 2934.
https://scholarworks.utep.edu/open_etd/2934

This is brought to you for free and open access by ScholarWorks@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of ScholarWorks@UTEP. For more information, please contact lweber@utep.edu.

PREDICTING STOCHASTIC VOLATILITY FOR EXTREME FLUCTUATIONS IN HIGH FREQUENCY TIME SERIES

MD AL MASUM BHUIYAN

Doctoral Program in Computational Science

APPROVED:

Maria C. Mariani, Ph.D., Chair

Natasha Sharma, Ph.D.

Thompson Sarkodie Gyan, Ph.D.

Kristine Garza, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

©Copyright

by

MD Al Masum Bhuiyan

2020

to my

Mother and Late Father

with love

PREDICTING STOCHASTIC VOLATILITY FOR EXTREME FLUCTUATIONS IN HIGH FREQUENCY TIME SERIES

by

MD AL MASUM BHUIYAN

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Doctoral Program in Computational Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2020

Acknowledgements

I would like to express my heartfelt gratitude to my doctoral advisor, Dr. Maria C. Mariani, Professor of the Computational Science Program at The University of Texas at El Paso. Her clarity of vision, her expertise in the subject, and her understanding and helpfulness are truly inspiring to me. Aside from introducing me to fascinating ideas and teaching me techniques to make them precise, she has also assisted me to become an independent researcher. Her offer to collaborate on the research of time series analysis has boosted my confidence, and her continuous support during that project has enabled me to thrive.

I also wish to thank the other members of my dissertation committee, namely, Dr. Natasha Sharma, Dr. Thompson Sarkodie Gyan, and Dr. Kristine Garza, faculties of The University of Texas at El Paso. Their generous suggestions, comments, and guidance were invaluable to the completion of this work. I also want to thank Dr. Ming Ying Leung, the Director of Computational Science program for her assistance that enabled me to successfully complete my Ph.D. dissertation.

Last but not least, I am immensely grateful to my family members for their unfailing love and support. I found in my mother, Parul Akter, my intellectual role model whom I idolize. My father Md Abdul Mannan Bhuiyan is always a pioneer to me. I found in my wife, Nusrat Sarmin, a model of love and sacrifice for me. I would also like to express my gratitude to my teachers and friends from my high-school days in Bangladesh to my doctoral years in the City of El Paso, Texas, for their love and trust in me. I admit that, in my growth as a data scientist, the credit goes more to my family, my mentors, and my friends than to myself.

NOTE: This dissertation was submitted to my Supervising Committee on the April, 2020.

Abstract

This work is devoted to the study of modeling high frequency time series including extreme fluctuations. As the high frequency data are collected at extremely fine scales, the fluctuations can capture the dynamics of data that evolve over time. A class of volatility models with time-varying parameters is used to forecast the volatility in a stationary condition at different lags. The modeling of stationary time series with consistent properties facilitates prediction with much certainty.

A large set of high frequency financial returns, closing prices of stock markets, high magnitudes of seismograms generated by the natural earthquakes, and the mining explosions is studied. The Generalized Autoregressive Conditional Heteroscedasticity (GARCH), Asymmetric Power Autoregressive Conditional Heteroscedasticity (APARCH) and Stochastic Volatility (SV) models are used to predict the data volatility. The data involving statistical noise and inaccuracies are continuously changing over time. Thus a filtering technique is performed to estimate the time-varying parameters by minimizing their variance. It is shown that the stochastic volatility (SV) is a better forecasting tool than GARCH (1, 1) and APARCH models, since it is less conditioned by autoregressive past information. We forecast one-step-ahead log volatility that is able to detect the extreme fluctuations of high frequency data.

A new approach is proposed to simulate the special case of high frequency data that do not fit always with the SV model. As the data reflect stochastic nature of most measurements over time, a stochastic differential equation with Ornstein-Uhlenbeck process has been applied in this case. This analysis helps to achieve the higher accuracy and fidelity for estimating the time-varying parameters of data volatility via Maximum Likelihood Estimation.

Table of Contents

Page

Acknowledgements v

Abstract vi

Table of Contents vii

Chapter

1 Introduction 1

1.1 Background & Research Problem 1

1.1.1 High Frequency Data and Volatility 3

1.2 Stochastic & Deterministic Modeling 4

1.2.1 White Noise 5

1.2.2 Autoregressive (AR) Models 5

1.2.3 Autocorrelation 6

1.2.4 Long Memory Technique 6

1.2.5 Filtering Approach 7

1.2.6 Kalman Filtering 7

1.2.7 Likelihood Approximation 10

2 Data Background 12

2.1 High Frequency Data with Extreme Fluctuations 12

2.1.1 Financial Time Series 12

2.1.2 Geophysical Time Series 15

2.2 Stationary Behavior of High Frequency Data 17

2.2.1 ADF test 18

2.2.2 KPSS test 20

3 Methodology 22

3.1 Volatility Models 22

3.1.1 GARCH Volatility Model 22
3.1.2 APARCH Volatility Model 23
3.1.3 Stochastic Volatility Model 24
3.2 Estimation Procedure 25
3.2.1 Parameter Estimation of SV Model 28
3.3 The Superposed Ornstein-Uhlenbeck Process 29
3.3.1 Parameter Estimation 30
4 Results & Discussion 33
4.1 Analysis of Volatility models 33
4.1.1 Results of GARCH Model 33
4.1.2 Results of APARCH Model 38
4.1.3 Results of Stochastic Volatility Model 39
4.2 A new approach for fitting SV model 42
5 Other Researches 49
5.1 Machine Learning Algorithms 49
5.1.1 Analysis of Fitted models 50
5.1.2 Model Evaluation 55
6 Concluding remarks 59
6.1 Future works 60
References 62
Curriculum Vitae 67

Chapter 1

Introduction

This chapter discusses on the literature review and research problems of this work. Some basic tools and definitions are briefly presented to facilitate the understanding of the methodologies in the application to the high frequency datasets.

1.1 Background & Research Problem

The forecasting of time series with estimation of time-varying parameters is very important in modeling the dynamic evolution of data volatility. It is assumed that a model that attracts the attention of investors can potentially be used to predict key variables, for instance, returns, volatility, and volume of stock market. It is to be noted that the development of forecasting methodologies in geophysics helps us to identify the type of source that generates a recorded seismic signal. This type of methodologies is generally applied to various fields, such as finance, geophysics, and safety of power system [1]. A reliable technique of forecasting, including the related time information, is essential to construct less risky portfolios or to make higher profits without the loss of generality and computational cost.

Financial time series manifests typical non-linear characteristics, and they involve volatility clustering where the returns indicate their dynamism. In this study, we develop a volatility forecasting method in which the logarithm of the conditional volatility follows an autoregressive time series model. R. F. Engle's paper introduced the Autoregressive Conditional Heteroskedasticity (ARCH) model to express the conditional variance of available returns as a function of previous observations [2]. Few years later, S. Bollerslev modified this concept and generalized the ARCH (GARCH) model that allows the conditional variance to depend on the previous conditional variance as well as on the squares of previous returns [3]. In other words, the system volatility in GARCH model is driven by the observed values in a pre-deterministic fashion. In fact, over the past few decades, a considerable amount of deterministic models has been suggested to forecast the observations, noise, and data volatility [4], [5], [6]. The reason is that they are simple and help to account for clustered errors and non-linearity issues. In the present study, we propose GARCH (1, 1) and asymmetric power ARCH (APARCH) models in a stationary way that is useful in the analysis of high frequency financial time series.

It is now widely believed that the measurements of a sequence of geophysics and finance are stochastically dependent on the time needed. In other words, there is a correlation among the numbers of data points at successive time intervals. In Ref. [7], [8], and [9], the authors used stochastic models to describe a unique type of measurement dependence in geophysical and financial data. It has been observed that the high frequency data may follow different behaviors over time, for instance, the mean reversion or extreme fluctuations. Such observations are unlike those of the classical modeling foundations. But the concept of time-dependent observations suggests that the current information needs to be evaluated on the basis of its past behavior [10]. This behavior of time series makes it possible to effectively forecast volatility and to obtain some stylized facts, namely, time-varying volatility, persistence, and clustering.

A distinctive feature of the high frequency time series is that the deterministic model and other volatility models do not show for a full statistical description of volatility [11]. When there are extreme fluctuations in the time series, the GARCH (1, 1) model predicts the volatility arbitrarily since it cannot capture the high volatile nature of the data [8]. It is because the volatility is not directly observable from the data. If the volatility is low at any data point, it does not imply that the risk of seismic events or financial markets is low too. The reason is that the volatility can be low while the probability distributions of data contain fat tails. So a dataset with low volatility can have much more extreme outcomes compared to another dataset with higher volatility [12]. In this work, a stochastic model with filtering technique is proposed as a way to estimate the time-varying parameters for data volatility. A continuous-time stationary sequences corresponding to finiancial returns, seismograms of natural earthquakes, and mining explosions are studied to forecast the stochastic volatility by using estimated parameters. These stationary sequences are very effective to capture the characteristic of time-varying parameters in an appropriate way [13]. The adequacy and stationarity of the data are determined by computing the estimated standard error and some powerful tests respectively, which will be discussed in chapter 3. The main difficulty of SV model is to fit it into the high frequency data with higher accuracy. The reason is that their likelihood estimations involve numerical integration over higher dimensional intractable integrals [14], whose maximization is rather complex. In this case, a recursive filtering technique with initial parameters has been used to estimate the time varying parameters via Maximum Likelihood Estimation (MLE).

Another difficulty of high frequency data is that it contains noisy observations like the values of zeroes for financial returns and seismograms of geophysical time series. The high frequency data follows log-normal distribution that allow us to use log-squared observations to forecast the volatility. However, the high frequency data does not always fit into SV methodology due to the log-squared observations and convergence of MLE is not always guaranteed [7]. In order to solve this problem, a new approach has been developed to simulate the data and then forecast the stochastic volatility, which are discussed in chapter 4. The data aspects and their related properties are now briefly discussed to understand the methodologies when they are fit to the datasets.

1.1.1 High Frequency Data and Volatility

High Frequency data is measured at extemely fine scales like monthly, weekly, daily, hourly, minutely, and second (then divide seconds into fractions) [15]. As the frequency of data appears very quickly, so the movement of such time series cover specific reasons for the fluctuations. Most of the times, the changes in time series variables with high frequency observations behave very differently than the normal time series fluctuations. For example, the stock prices which are today recorded at seconds or less frequency and investors are creating tools which determine the prediction of next movement and flow of such data to decide whether to invest in stock A vs. Stock B in Market X vs. Market Y.

The volatility of high frequency data shows how much observations move with a condition. The conditional volatility measure the uncertainty about a variable given model and the information. The volatility of high frequency data changes dramatically at a short interval and the periods of high volatility are sometimes correlated. This establishes that the volatility itself is very volatile. The fluctuations of data typically exhibit the volatility clustering [12]. That is to say, small changes in the price tend to be followed by small changes, and large changes by large ones. The volatility clustering suggests that the current information is highly correlated with past information at different levels.

1.2 Stochastic & Deterministic Modeling

Stochastic modeling is an interesting and challenging area of probability and statistics. Let us consider the daily closing prices of stock exchanges as a sequence of random variables, $S_1, S_2, S_3, \cdots$, where the random variable S_1 denotes the value taken by the series at the first time point, the variable S_2 denotes the value for the second time point, S_3 denotes the value for the third time point, and so on. In general, a collection of random variables, S_t, indexed by t is known as a stochastic process. This process describes the characteristics of the data that seemingly fluctuates in a random fashion over time [16].

Stochastic models can be contrasted with deterministic models. A deterministic model is specified by a set of equations that describe exactly how the system will evolve over time [17]. In a stochastic model, the evolution is at least partially random and if the process is run several times, it will not give identical results. Different runs of a stochastic process are often called realizations of the process.

Deterministic models are generally easier to analyze than stochastic models. However, in many cases, stochastic models are more realistic, particularly for problems that involve small observations at different time level. For example, suppose we are trying to model the management of a rare species, looking at how different strategies affect the survival of the species. In this case, the deterministic models will not hold good as the prediction will show either “definitely extinct” or “definitely survives”. In a stochastic model, there will be a probability of extinction that studies how this is affected by management practices.

1.2.1 White Noise

A time series ϵ_t is called a white noise, if ϵ_t is a sequence of independent and identically distributed (i.i.d) random variables with a mean of zero, finite variance, and no serial correlation [18]. In other words, it has the following properties:

1. Zero mean: $E(\epsilon_t) = E(\epsilon_{t-1}) = 0$
2. Constant variance: $E(\epsilon_t^2) = E(\epsilon_{t-1}^2) = \sigma^2 = Var(\epsilon_t)$
3. Uncorrelated in time: $E(\epsilon_t \epsilon_{t-i}) = E(\epsilon_{t-j}\epsilon_{t-i-j}) = \cdots = 0 = Cov(\epsilon_{t-j}\epsilon_{t-i-j}), \forall\ i \neq j.$

Here, the property 2 is called as Conditional Homoscedasticity. In particular, if the white noise follows a Normal distribution then it is known as Normal or Gaussian white noise.

1.2.2 Autoregressive (AR) Models

A time series is a sequence of measurements of the same variable(s) made over time. In other words, in an autoregressive model, a time series is regressed on previous values of the same time series [19]. For example, if the value of y_t depends on the value of y_{t-1}, the AR(1) model is defined as:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t \tag{1.1}$$

where ϵ_t is a white noise. In general, the AR(p) model is defined as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t. \tag{1.2}$$

where $\phi_1, \phi_2, \cdots, \phi_p$ are the parameters of the model and c is a constant.

1.2.3 Autocorrelation

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF) [19]. For example, the ACF for a time series y_t is given by:

$$\mathrm{Corr}(y_t, y_{t-k}).$$

where the value of k is called a lag or time gap. The lag 1 autocorrelation (*i.e.*, $k = 1$) is the correlation between values that are one time period apart. In general, a lag k autocorrelation is the correlation between values that are k time periods apart. In time series analysis, the ACF is a way to measure the linear relationship between an observation at time t and the observations at previous times. If we assume an AR (k) model, then we may wish to only measure the association between y_t and y_{t-k} and filter out the linear influence of the random variables that lie in between (i.e., $y_{t-1}, y_{t-2}, \cdots, y_{t-(k-1)}$), which requires a transformation on the time series.

1.2.4 Long Memory Technique

The long memory technique describes the higher order correlation structure of a time series. If a time series y_t follows a long-memory process, it implies that there is a persistent temporal dependence between the observations widely separated in time. This technique can also be regarded as a fractionally integrated process, i.e., an intermediate phase between stationary and unit root process [20], [21]. Considering an Auto-regressive Integrated Moving Averages (ARIMA) model, the long memory process may be represented as follows:

$$\nabla^{\lambda} y_t = (1 - L)^{\lambda} y_t = \varepsilon_t, \tag{1.3}$$

where λ is a fractional difference parameter, ε_t is a white noise with variance σ_ε and L is a lag operator. The parameter λ identifies a process as the short memory or the long memory. When the parameter is less than 0.5, the process is long memory. In this study, the estimated parameter λ showed that the datasets follow long memory approach. For details of the algorithm and estimation, the reader is referred to [22].

1.2.5 Filtering Approach

The state space model is defined by a relation between the m-dimensional observed time series, $\mathbf{y}_t$, and the n-dimensional state vector (possibly unobserved), $\mathbf{x}_t$ [23]. An observed (space) equation is driven by the stochastic process as follows:

$$y_t = Hx_t + \upsilon_t \tag{1.4}$$

where H_t is a $m \times n$ observation matrix which is stationary and noiseless connection between the state vector and the measurement vector, $\mathbf{x}_t$ is a state vector of $n \times 1$, and υ_t is a Gaussian error term ($\upsilon_t \sim (0, R)$).

The unobservable vector $\mathbf{x}_t$ is generated from the transition equation which is defined as:

$$x_t = \Phi x_{t-1} + \omega_t, \tag{1.5}$$

where Φ is a $n \times n$ transition matrix and $\omega_t \sim (0, Q)$. We assume that the process starts with a Normal vector x_0. From Eqs. (1.4) and (1.5), the estimation is made for the underlying unobserved data $\mathbf{x}_t$ from the given data $Y_m = \{y_1, \ldots, y_m\}$. When $m = t$, the process is called filtering. In this study, a recursive filtering technique is analyzed to filter out the unnecessary information (noise) for finding the "best estimate" from noisy data.

1.2.6 Kalman Filtering

The kalman filtering procedure is used for parameter estimation of a stochastic model. The kalman filtering is advantageous as it is a recursive optimal estimation tool that cleans up the measurements inaccuracies and projects these measurements onto the state estimate [24], [25]. So, the filter keeps track of the estimated state of the system and the variance or uncertainty of the estimate.

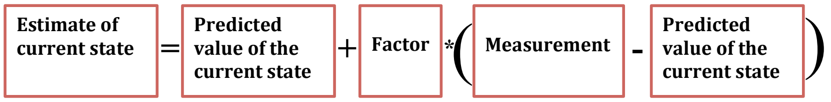
In order to estimate the system's state, an average of multiple measurement is taken into account. At initial stage, it is assumed that the estimate $\hat{x}_{t,t}$ would be the average of all previous measurements (z_t) at time t:

$$
\begin{aligned}
\hat{x}_{t,t} &= \frac{1}{t}(z_1 + z_2 + \cdots + z_t) \\
&= \frac{1}{t}\sum_{t=1}^{t} z_t \\
&= \frac{1}{t}\Big((\sum_{t=1}^{t-1} z_t) + z_t\Big) \\
&= (\frac{1}{t}\sum_{t=1}^{t-1} z_t) + \frac{1}{t} z_t \\
&= \frac{1}{t}\frac{t-1}{t-1}\sum_{t=1}^{t-1}(z_t) + \frac{1}{t} z_t \\
&= \frac{t-1}{t}\frac{1}{t-1}\sum_{t=1}^{t-1}(z_t) + \frac{1}{t} z_t \\
&= \frac{t-1}{t}\hat{x}_{t,t-1} + \frac{1}{t} z_t \\
&= \hat{x}_{t,t-1} - \frac{1}{t}\hat{x}_{t,t-1} + \frac{1}{t} z_t \\
&= \hat{x}_{t,t-1} + \frac{1}{t}(z_t - \hat{x}_{t,t-1})
\end{aligned}
$$

where x_t is the true value of the observation; z_t is the measurement value of observation at time t; $\hat{x}_{t,t}$ is the estimate of x at time t; $\hat{x}_{t,t-1}$ is the previous estimate of x that was made at time $t-1$. Thus,

$$\hat{x}_{t,t} = \hat{x}_{t,t-1} + \frac{1}{t}(z_t - \hat{x}_{t,t-1}) \tag{1.6}$$

where $(z_t - x_{t,t-1})$ is a measurement residual (innovation) that contains new information. Now the above equation can be expressed as:


Estimate of current state
=
Predicted value of the current state
+
Factor
*
(
Measurement
-
Predicted value of the current state
)

where the factor, $\frac{1}{t}$, is defined as a Kalman Gain in the filtering technique. We denote this Kalman Gain as K_t that explains how much we want to change our estimate by a given measurement. The subscript t indicates that the Kalman Gain keeps changing with every iteration based on the uncertainty in estimate and uncertainty in measurement. The equation of Kalman Gain is given as follows:

$$\begin{aligned}\text{Kalman Gain } (K_t) &= \frac{\text{uncertainty in estimate}}{\text{uncertainty in estimate} + \text{ uncertainty in measurement}} \\ &= \frac{p_{t,t-1}}{p_{t,t-1} + r_t}\end{aligned}$$

where $p_{t,t-1}$ is the extrapolated estimate uncertainty, r_t is the measurement uncertainty. From the above equation, it is clear that the Kalman Gain is a number between 0 and 1 $(0 \leq K_t \leq 1)$. Now the Eq. (1.6) can be rewritten as follows:

$$\hat{x}_{t,t} = \hat{x}_{t,t-1} + K_t(z_t - \hat{x}_{t,t-1}) \tag{1.7}$$

$$= (1 - K_t)\hat{x}_{t,t-1} + K_t z_t \tag{1.8}$$

As we can see the Kalman Gain (K_t) is the weight that is used for the measurement and $(1 - K_t)$ is the weight that is used for the estimate. When the uncertainty in the measurement is very large and the uncertainty in the estimate is very small, the Kalman Gain is close to zero. So a big weight is given to the estimate and a small weight to the measurement. On the other side, when the uncertainty in the measurement is very small and the uncertainty in the estimate is very large, the Kalman Gain is close to one. So we give a small weight to the estimate and a big weight to the measurement. If the measurement uncertainty is equal to the estimate uncertainty, then the Kalman gain equals to 0.5. We now present an algorithm of Kalman filtering as follows [26]:

1. STEP 0: Initialization

    - Initial System Estimate: $\hat{x}_{1,0}$
    - Initial System Uncertainty: $p_{1,0}$

2. STEP 1: Measurement
   - Measurement System Estimate: y_t
   - Measurement Uncertainty: r_t

3. STEP 2: State Update
   - Input: $z_{1,0}, r_t, \hat{x}_{t,t-1}, p_{t,t-1}$
   - Kalman Gain: $K_t = \frac{p_{t,t-1}}{p_{t,t-1}+r_t}$
   - State Update: $\hat{x}_{t,t} = \hat{x}_{t,t-1} + K_t(z_t - \hat{x}_{t,t-1})$
   - Covariance Update: $p_{t,t} = (1 - K_t)p_{t,t-1}$
   - Output: $\hat{x}_{t,t}, p_{t,t}$

4. STEP 3: Prediction
   - Dynamic Model (State Space Model):
     $\hat{x}_{t+1,t} = F\hat{x}_{t,t} + G\hat{u}_{t,t}$
   - Covariance Update: $p_{t+1,t} = Fp_{t,t}F^T + Q$

where, F is a state transition matrix of coefficients, G is a control matrix, u is an input variable, and Q is a process noise uncertainty. In the prediction step, we see that the Kalman filter produces estimates of the current state variables, along with their uncertainties. It is assumed that the errors terms of this algorithm are Gaussian. Once the outcome of the next measurement (necessarily involved with some amount of error and random noise) is observed, these estimates are updated using a weighted average.

1.2.7 Likelihood Approximation

Let φ denote the parameters of the state space model, which are embedded in the system matrices H_t, Φ, υ_t and ω_t. These parameters are typically unknown, but estimated from the data $Y = y_1, \ldots, y_m$.

The likelihood $L(\varphi\big|Y)$ is a function that assigns a value to each point in the parameter space Δ which suggests the likelihood of each value in generating the data [7]. However, the likelihood is proportional to the joint probability distribution of the data as a function of the unknown parameters. The maximum likelihood estimation means the estimation of the value of $\varphi \in \Delta$ that is most likely to generate the vector of the observed data y_t [27]. The likelihood can be represented as:

$$\hat{\varphi}_{MLE} = \max_{\varphi\in\Delta} L(\varphi\big|Y) = \max_{\varphi\in\Delta} L_Y(\varphi) = \max_{\varphi\in\Delta} \prod_{t=1}^{m} f(y_t\big|y_{t-1};\varphi) \tag{1.9}$$

where $\hat{\varphi}$ is the maximum likelihood estimator of φ. Since the natural logarithm function increases on $(0, \infty)$, the maximum value of the likelihood function, if it exists, occurs at the same points as the maximum value of the logarithm of the likelihood function. The log-likelihood function is defined as:

$$\hat{\varphi}_{MLE} = \max_{\varphi\in\Delta} lnL(\varphi\big|Y) = \max_{\varphi\in\Delta} lnL_Y(\varphi) = \max_{\varphi\in\Delta} \sum_{t=1}^{m} lnf(y_t\big|y_{t-1};\varphi). \tag{1.10}$$

Since this is a highly non-linear and complicated function of the unknown parameters, we first consider the initial state vector $\mathbf{x}_0$ and develop a set of recursions for the log-likelihood function with its first two derivatives [28]. We then use Newton-Raphson algorithm [29] successively until the negative of the log-likelihood is minimized to obtain the MLE.

Chapter 2

Data Background

This chapter focuses on the background of data arising in finance and geophysics. We analyzed the high frequency data that includes some extreme fluctuations. In fact, it is the dynamic behavior of the data that encourages us to apply our methodology in this paper.

2.1 High Frequency Data with Extreme Fluctuations

2.1.1 Financial Time Series

A large set of daily closing prices for both developed and emergent stock markets is studied. The emergent market indices are studied from two countries: Brazil (BOVESPA), from 04-27-1993 to 10-22- 2001; Thailand (SETI), from 07-02-1997 to 10-25-2001. For developed markets the National Association of Securities Dealers Automated Quotations (NASDAQ) and the Standard and Poor′s 500 (S&P 500) stock exchanges from USA are analyzed. A good perspective on the trending direction or risk management of the high frequency returns of stock markets as shown in Fig. 2.1. The financial crisis that occurred is evident in the large spikes in the figure.

We then analyzed two financial crashes data that are useful in analyzing their effects and forecasting their stochastic volatilities. The financial crashes cause huge looses in the stock market. For example, the great recession of 2008 had considerable effects on the U.S. and global economy. The financial market around the world suffered great disruptions in asset and credit companies, massive erosions of wealth and innumerable bankruptcies. The high frequency market data (minute by minute) employed in this study are obtained from

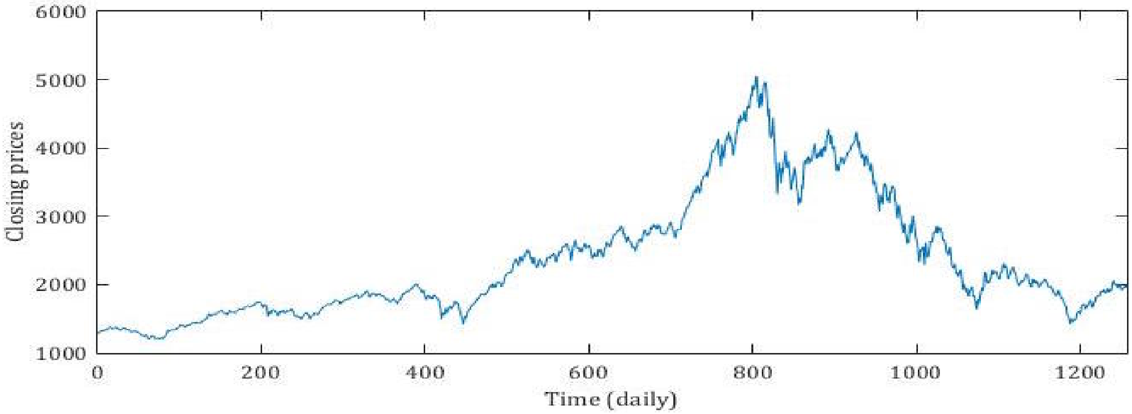


Figure 2.1: The closing prices of daily trading observations from the NASDAQ stock exchange.

the Lehman Brothers collapse and Flash Crash event, which occurred in 2008 and in 2010 respectively. So we analyze the high frequency returns from stock companies that were affected by these events. The companies are as: Citi Bank, Bank of America, ExxonMobil corporation (XOM), Walmart Retail company (WMT), Verizon Communications Inc. (VZ), United Technologies Corporation (UTX), McDonalds corporation (MCD) etc. Now the background of Lehman Brothers collapse and the Flash Crash events financial crashes are briefly discussed [30].

The Lehman Brothers collapse

Lehman Brothers, one of the biggest financial services firm in the world, was brought down by the collapse of the subprime mortgage market. The company was established by Henry Lehman with a general store that sold utensils, dry goods and groceries to cotton farmers in Montgomery, Alabama. Later, his brothers Emanuel and Mayer joined in the business and shifted it from dealing with commodities to merchant banking in New York. In the beginning, the company did pretty well in both domestic and international market, and it played an important role in the subprime market. It expanded into loan origination, gaining five mortgage lenders between 2003 and 2004, including some specializing in subprime mortgages.

These mortgages were provided to the borrowers with low credits. In 2003, the company made \$18.2 billion in loans and ranked third in lending. However, due to its significant exposure to the US subprime mortgage and real estate markets, highly-leveraged, risk- taking business strategy supported by limited equity; culture of excessive risk-taking and among others reasons, the company filed for Chapter 11 bankruptcy protection on September 15, 2008. At that time, Lehman Brothers was the fourth largest investment bank in U.S.A and employed over 25,000 people across the globe. But it had \$639 billion in assets and \$619 in debt. So it was the largest victim of the subprime mortgage crisis that made a huge financial crash in the stock market. We refer to [31] and references therein for more details.

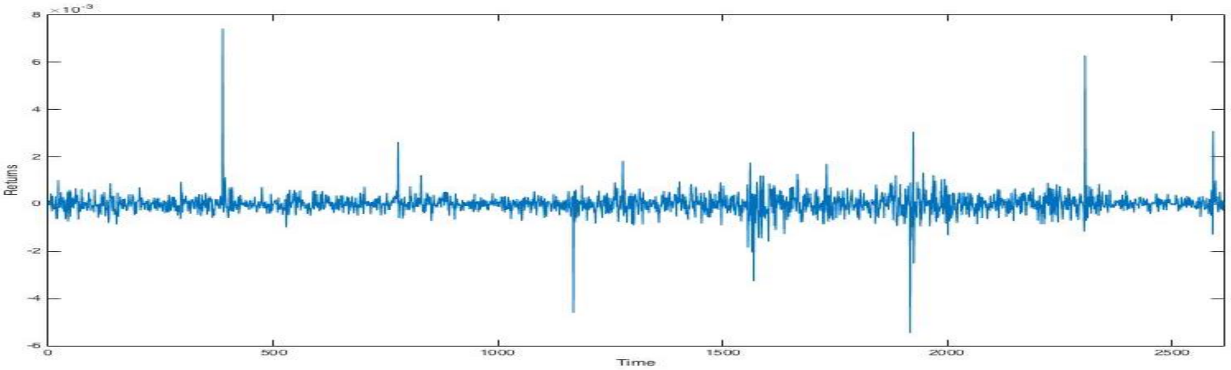


Figure 2.2: The high frequency returns sample (generated per minute) of DISCOVER after Lehman Brothers Collapse.

The Flash Crash

The Flash Crash is an event in electronic securities markets where the sudden withdrawal of stock orders quickly increases price declines. This results in the swift sell-off of securities that can happen over a few minutes, resulting in sudden declines. The Flash Crash event occurred on May 6, 2010, was a trillion-dollar financial crash which lasted for about 36 minutes. The magnitude of the crash was intensified as traders reacted to irregularities in the market, for example, the heavy selling in one or many securities and automatically begin selling large volumes at a very fast pace to avoid losses. During this period, the Dow Jones Industrial Average (DJIA) lost trillions of dollars in the value of stocks and shares.

In a report by Nanex LLC [32], high-frequency traders are said to be largely responsible for flash crashes. The rapid and aggressive sale of the E-mini contracts took out several levels of market depth and caused an explosion of quotes and traders in Exchange traded funds, equities and options with a delay of approximately 20 micro seconds (see [33]). The large amount of quotes strained the system, causing prices to sharply decline.

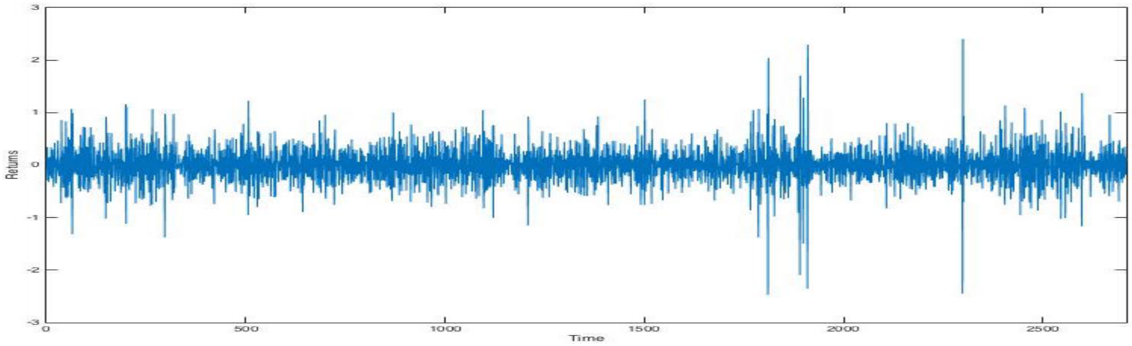


Figure 2.3: The high frequency returns sample (generated per minute) of INTEL after Flash Crash Event.

2.1.2 Geophysical Time Series

This subsection deals with the background of geophysical data including high magnitudes and extreme fluctuations. The data contains information about the date, time, longitude, latitude, the average distance to seismic events, average azimuth and the magnitude of each seismic event in the region. For details of the background, the readers are refered to [13]. In this study, we used four datasets of natural earthquakes (EA1-EA4) and four datasets of mining explosions (EX1-EX4) in order to forecast their volatilities. The dynamic behavior of these data are shown in the following subsection:

Natural Earthquakes Data

The earthquakes used in this study correspond to a set of aftershocks of the June 26, 2014, magnitude M=5.2 intraplate earthquake. It occurred far from any active tectonic boundary, located between the states of Arizona and New Mexico in the USA. We collected the seismograms containing the seismic waves from two nearby seismic stations: IU.TUC and IU.ANMO, that are located between 150 and 400 km from the seismic events. Fig. 2.4 shows examples of the seismograms recorded by the IU.TUC station from one earthquake.

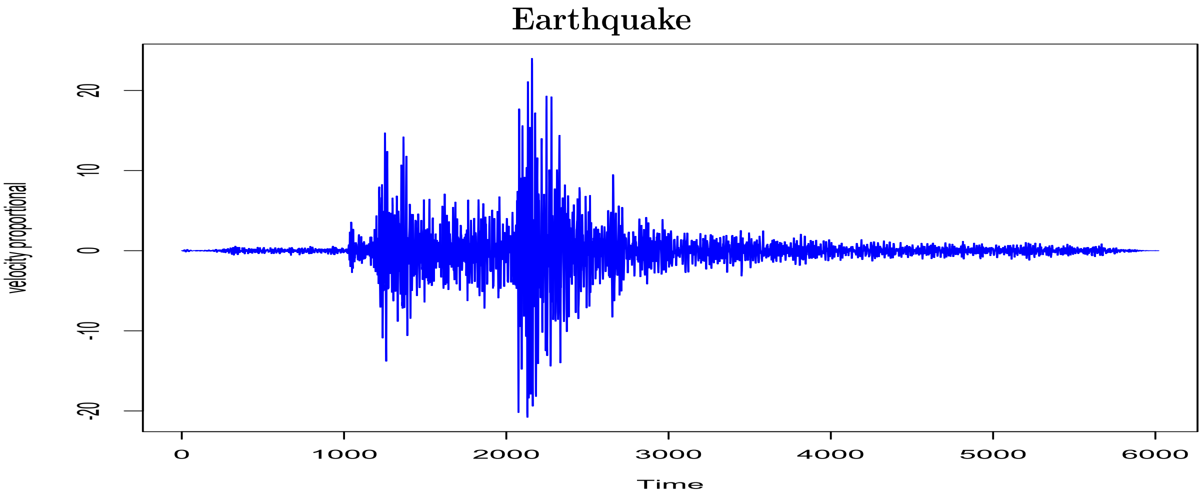


Figure 2.4: The arrival phases from an earthquake as recorded by IU.TUC seismic station.

Mining Explosions Data

The human made mining explosions cataloged with similar magnitudes as the earthquakes (M=3.0-3.3) are studied. These seismic events are located in the same region within a radius of 10 km where a large surface copper mine triggered off several explosions forming part of quarry blasts activities. The broadband seismograms are downloaded from the Incorporated Research Institutions for Seismology Data Management Centers (IRIS DMC). The Seismic Analysis Code (SAC) software was employed to preprocess the data. The displacement of the ground (in nm) in the vertical (Z-component) direction of sesimograms are used in this analysis. Figure 2.5 shows the arival phases of explosion seismograms recorded by IU.TUC station [34].

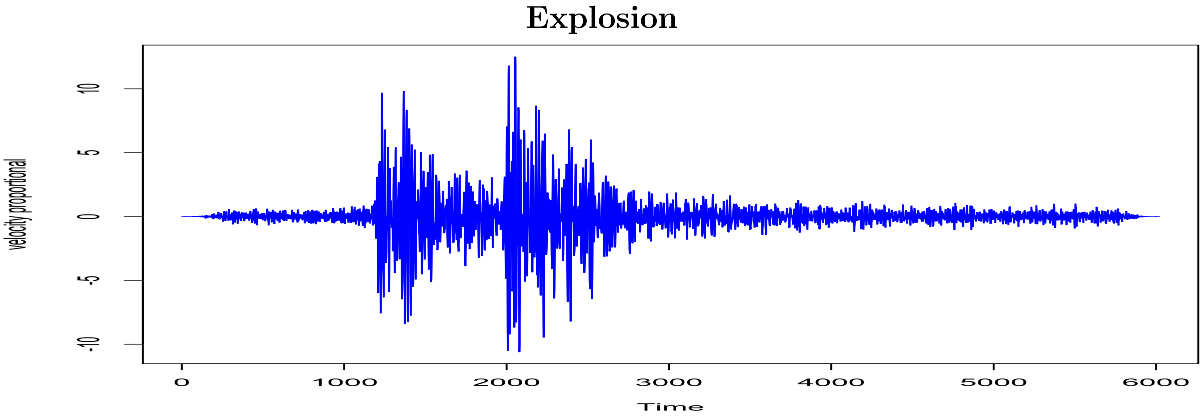


Figure 2.5: The arrival phases from an explosion as recorded by IU.TUC seismic station.

In Fig. 2.4 & 2.5, it is observed that the frequency components change from one interval to another in earthquake or explosion as long as they last. The mean of the series appears to be stable with an average magnitude of approximately zero. This dynamic behavior illustrates the time evolution of the magnitude with its volatility. The volatility changes with time and it is high at a certain intervals, but low at another intervals. The volatility clustering reflects its time varying nature, as well as the mean reversion characteristics of the data.

2.2 Stationary Behavior of High Frequency Data

The stationary time series follow some statistical properties that are independent of time. If a time series x_t is stationary, then its second-order behavior, namely, mean, variance and covariance do not change with time t. A stationary series is relatively easy to predict and defined as follows:

1. $E(x_t^2) < \infty$,

2. $E[x(t)] = \mu$,

3. Cov $x(s,t) =$ Cov $x(s+h, t+h), \forall s,t,h \in \mathbb{Z}$

So x_t must have three features to be a stationary form: finite variation, constant first moment, and the constant variance. In this study, the stationarity of financial, earthquake and mining explosion time series are analyzed by using unit root test. A unit root test provides a way to test whether an autoregressive process is a random walk as opposed to a stationary process. The test-statistics of data are computed by using two unit root tests, namely, the Augmented Dickey Fuller (ADF) [35] and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests [36]. These two tests are powerful and capable of handling complex models.

2.2.1 ADF test

The ADF test is an augmented version of the Dicky Fuller test, that is used for large and complicated set of time series models. The ADF test statistic is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some levels of confidence. It takes into account the basic autoregressive unit root test to accommodate general Autoregressive Moving Average (ARMA) (p,q) models with unknown orders and non-zero lagged values. The summary statistics of this test are given below when the tests are applied to the high frequency data:

Test interpretation:

H_0 : There is a unit root for the time series.

H_a : There is no unit root for the time series, i.e., the series is stationary.

Since the computed p-values in Tables 2.1 and 2.2 are lower than the significance level $\alpha = 0.05$, the null hypothesis H_0 is rejected that has a unit root. So the alternative hypothesis H_a is accepted that the data are stationary time series.

Table 2.1: **ADF t-statistics test for financial data**

Events	Daily data		Minute Data	
	Test statistics	p-value	Test statistics	p-value
Discover	-11.92	0.01	-12.85	0.01
Microsoft	-12.95	0.01	-14.90	0.01
Walmart	-12.87	0.01	-13.56	0.01
JPM Chase	-13.38	0.01	-12.80	0.01

Table 2.2: **ADF t-statistics test for geophysical data**

Events	Test statistics	p-value	Events	Test statistics	p-value
EA1	-42.01	0.01	EX1	-40.30	0.01
EA2	-41.40	0.01	EX2	-36.35	0.01
EA3	-37.39	0.01	EX3	-40.83	0.01
EA4	-39.09	0.01	EX4	-41.94	0.01

2.2.2 KPSS test

The KPSS test are used for testing a null hypothesis that an observable time series is stationary against the alternative of no unit root. The major difference between KPSS and ADF tests is that the the KPSS test is able to check stationarity around a deterministic trend. The deterministic trend suggests that the slope of the trend in the series does not change permanently. Therefore, if the series goes through a shock at any time point, the series tends to regain its original path over time. The summary statistics for the results of the KPSS test are displayed in Tables 2.3 and 2.4 respectively.

Test interpretation:

H_0 : The series is trend stationary.

H_a : The series is non-stationary

Table 2.3: **KPSS t-statistics test for financial data**

Events	Daily data		Minute Data	
	Test statistics	p-value	Test statistics	p-value
Discover	0.1633	0.1	0.1731	0.1
Microsoft	0.1980	0.1	0.1297	0.1
Walmart	0.1037	0.1	0.0599	0.1
JPM Chase	0.1719	0.1	0.1059	0.1

Table 2.4: **KPSS t-statistics test for geophysical data**

Events	Test statistics	p-value	Events	Test statistics	p-value
EA1	0.0017	0.1	EX1	0.0013	0.1
EA2	0.012	0.1	EX2	0.0014	0.1
EA3	0.0013	0.1	EX3	0.0012	0.1
EA4	0.0012	0.1	EX4	0.0017	0.1

As the computed p-values are greater than the significance level $\alpha = 0.05$, the null hypothesis H_0 is accepted for all data sets. Thus the time series used in this paper are stationary time series.

Chapter 3

Methodology

3.1 Volatility Models

This chapter describes the methodologies that are used to forecast the volatility of high frequency financial and geophysical data. Three types of volatility models and filtering techniques are analyzed to estimate the time-varying parameters of data. It is also shown that how the connection is established between SV model and superposed OU-type models that help to forecast the volatility for special case of high frequency data.

3.1.1 GARCH Volatility Model

The Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model ([2], [3]), was introduced in order to model the fluctuations of the variances of financial data. It is conditional, because the nature of subsequent volatility is conditioned by the information of current period. Heteroscedasticity refers to non-constant volatility. The observations y_t of high frequency financial time series used in this paper may be represented as:

$$y_t = \sigma_t \eta_t, \tag{3.1}$$

where σ_t is volatility of the observations and $\{\eta_t\}_{t\in\mathbb{N}}$ is a Gaussian white noise sequence, independent of $\{\sigma_t\}_{t\in\mathbb{N}}$ and $\{y_t\}_{t\in\mathbb{N}}$. This equation can be interpreted as the observation equation of a state space model (see subsection 1.2.5), whereby the state equation is a recursive formula for the state σ_t:

$$\sigma_t^2 = a_0 + a_1 y_{t-1}^2 + b_1 \sigma_{t-1}^2, \tag{3.2}$$

where $a_0, a_1, b_1 \geq 0$, so that $\sigma_t^2 > 0$ for any values of y_t. Eqs. (3.1) and (3.2) admit a non-Gaussian ARMA (1,1) model [37] for the squared process as:

$$y_t^2 = a_0 + (a_1 + b_1)y_{t-1}^2 + \phi_t - b_1\phi_{t-1}, \tag{3.3}$$

where $\phi_t = \sigma_t^2(\eta_t^2 - 1)$. In order to compute the variance at time t, we follow the standard GARCH (m, n) model which is of the form:

$$\sigma_t^2 = a_0 + \sum_{j=1}^{m} a_j y_{t-j}^2 + \sum_{j=1}^{n} b_j \sigma_{t-j}^2. \tag{3.4}$$

If $n = 0$ then the GARCH model changes into an ARCH (m) model.

The parameters a_0, a_i and b_j are estimated by MLE (subsection 1.2.7) using the likelihood function. Taking into account the Normal probability density function, the conditional likelihood in Eq. (1.10) is obtained from the product of Normal $(N(0, \sigma_t^2))$ densities with σ_t^2. Using the estimated parameters, we obtain one-step-ahead prediction of the volatility $(\widehat{\sigma}_t^2)$, that is,

$$\widehat{\sigma}_t^2 = \widehat{a}_0 + \sum_{j=1}^{m} \widehat{a}_j y_{t+1-j}^2 + \sum_{j=1}^{n} \widehat{b_j} \widehat{\sigma}_{t+1-j}^2. \tag{3.5}$$

We can analyze the residuals and squared residuals to test the Normality using some statistical tests, for instance, Jarqua-Bera test [38], Shapiro-Wilk test [39], Ljung-Box test [40] and LM-Arch test [41].

3.1.2 APARCH Volatility Model

The asymmetric power ARCH (APARCH) model is also used for estimating the data volatility. We see that the above standard GARCH model (Eq. 3.4) estimate volatility (σ_t) based on the past values of y_t^2, for example, Eq. (3.4) is a function of y_{t-1}^2. But the model does not take into account whether the past values of y_t are positive or negative. It is known that the function of y_{t-1}^2 is symmetric in y_{t-1}. In this case, APARCH model uses a flexible class of non-negative functions instead of square function. It also offers more flexibility compared to GARCH model by modeling σ_t^δ, where $\delta > 0$ is another parameter. The APARCH (p, q) model for the conditional standard deviation is as follows:

$$\sigma_t^\delta = \alpha_0 + \sum_{j=1}^{p} \alpha_j (|y_{t-j}| - \gamma_j y_{t-j})^\delta + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^\delta \tag{3.6}$$

where $\delta > 0$ and $-1 < \gamma_j < 1$, $j = 1, \cdots, p$. Here, if $\delta = 2$ and $\gamma_1 = \cdots = \gamma_p = 0$, it will be a form of standard GARCH model.

3.1.3 Stochastic Volatility Model

This subsection focuses on the stochastic volatility (SV) model used in this paper. The SV technique implies that the volatility is driven by an innovation sequence, that is, independent of observations [42]. It causes the volatility through an unobservable process that allows it (volatility) to vary stochastically. From Eq. (3.1), the observations y_t can be expressed as:

$$y_t = \sigma_t \eta_t, \tag{3.7}$$

where σ_t is the data volatility and $\{\eta_t\}_{t \in \mathbb{N}}$ is a Gaussian white noise sequence.
To develop the SV model, we use the log-squared observations of the time series in Eq.(3.7):

$$log y_t^2 = log \sigma_t^2 + log \eta_t^2$$

which can be rewritten as:

$$m_t = h_t + log \eta_t^2, \tag{3.8}$$

where $m_t = log y_t^2$ and $h_t = log \sigma_t^2$. Thus the observations m_t are generated by two components namely, the unobserved volatility h_t and the unobserved noise $log \eta_t^2$. Considering the autoregression, the first term on the right hand side of Eq.(3.8) i.e. h_t can be expressed as:

$$h_t =_0 + \alpha_1 h_{t-1} + \omega_t, \tag{3.9}$$

where ω_t is a white Gaussian noise with the variance σ_ω^2. Eqs. (3.8) and (3.9) constitute the stochastic volatility model by Taylor [43]. To compute the observation noise, we take into account the mixtures of two Normal distributions with one centered at zero. Thus we have:

$$y_t = \beta + h_t + \gamma_t, \tag{3.10}$$

where β is the mean of log-squared observations and $\gamma_t = B_t z_{t0} - (B_t - 1)z_{t1}$, which fulfills the following conditions:

$$\begin{aligned} z_{t0} &\sim \text{i.i.d N}(0, \sigma_0^2), \\ z_{t1} &\sim \text{i.i.d N}(\mu_1, \sigma_1^2), \\ \text{and } B_t &\sim \text{i.i.d Bernoulli } (p), \end{aligned}$$

where p is an unknown mixing probability and i.i.d implies independently and identically distributed. The time-varying probabilities are defined as $\Pr\{B_t = 0\} = p_0$ and $\Pr\{B_t = 1\} = p_1$, where $p_0 + p_1 = 1$. The SV model has a characteristic function that describes the probability density function of the model. In particular, the high frequecy data in our study seems to have a Normal distribution but not exactly, sometimes there is a little skew on one tail (see the density plot in Figs. 4.4 and 4.5). So, we typically keep the ARCH Normality assumption on ϵ_t [2]. In this study, our approach is to estimate the parameters $\alpha_0, \alpha_1, \sigma_\omega, \sigma_0$ and σ_1 and then predict future observations y_{n+m} from n observations.

3.2 Estimation Procedure

In this subsection, a general estimation procedure for estimating time-varying parameters of SV model are discussed. We will estimate x_t with its error term: $e_t = \hat{x_t} - x_t$ and begin the estimation procedure with state space model described in chapter 1, where the assumptions [44] are as:

$$y_t = Hx_t + \upsilon_t \tag{3.11}$$

$$x_{t+1} = \Phi x_t + \omega_t \tag{3.12}$$

where $\Phi_{n\times m}$ is a stationary transition matrix from the state at t to the state at $t+1$; ω_t is the associated white noise process with known covariance; H is the noiseless connection between the state vector and the measurement vector; υ_t is the associated measurement error. The covariances of the two noise terms are assumed as stationary over time and computed as: $Q = E[\omega_t \omega_t^T]$ and $R = E[\upsilon_t \upsilon_t^T]$. The aim is to find the optimal filter that minimizes the mean squared error, $E(e_t^2)$, which is equivalent to P_t (the error covariance matrix at time t):

$$P_t = E[e_t e_t^T] = E[(x_t - \hat{x}_t)(x_t - \hat{x}_t)^T] \tag{3.13}$$

The $\hat{x}_t{}'$ is assumed as the prior estimate of $\hat{x}_t$ which is obtained from the system. Now we can write the state update equation for the new estimate, combining the old estimate with measurement. Eq. (1.7 indicates,

$$\hat{x}_t = \hat{x}_t{}' + K_t(y_t - H\hat{x}_t{}') \tag{3.14}$$

where K_t is the Kalman gain and $(y_t - H\hat{x}_t{}')$ is the innovation or measurement residual. We now substitute y_t from Eq. (3.11) into Eq. (3.14) as:

$$\hat{x}_t = H\hat{x}_t{}' + K_t(Hx_t + \upsilon_t - H\hat{x}_t{}') \tag{3.15}$$

At this point, we compute the the error-covariance matrix using Eq. (3.13) as:

$$P_t = E[[(I - K_t H)(x_t - \hat{x}_t{}') - K_t \upsilon_t][(I - K_t H)(x_t - \hat{x}_t{}') - K_t \upsilon_t]^T] \tag{3.16}$$

Here, $x_t - \hat{x}_t{}'$ is the error of the prior estimate which is uncorrelated with the measurement noise, thus the above equation can be written as:

$$P_t = (I - K_t H)E[(x_t - \hat{x}_t{}')(x_t - \hat{x}_t{}')^T](I - K_t H) + K_t E[\upsilon_t \upsilon_t']K_t^T \tag{3.17}$$

$$\Rightarrow P_t = (I - K_t H)P_t^T(I - K_t H)^T + K_t R K_t^T \tag{3.18}$$

which is a error-covariance update equation and P_t' is the prior estimate of P_t. Now the mean squared error (MSE) can be obtained by computing the trace of following matrix (P_t):

$$P_t = \begin{bmatrix} E[e_{t-1}e_{t-1}^T] & E[e_t e_{t-1}^T] & E[e_{t+1}e_{t-1}^T] \\ E[e_{t-1}e_t^T] & E[e_t e_t^T] & E[e_{t+1}e_t^T] \\ E[e_{t-1}e_{t+1}^T] & E[e_t e_{t+1}^T] & E[e_{t+1}e_{t+1}^T] \end{bmatrix} \tag{3.19}$$

So the MSE can be minimized by minimizing the trace of P_{tt}. The Eq. (3.17) can be furnished as follows:

$$P_t = P'_t - K_t H P'_t - P'_t H^T K_t^T + K_t(HP'_t H^T + R)K_t^T \tag{3.20}$$

The trace of P_t is given as:

$$tr([P_t]) = tr([P'_t]) - 2tr([K_t H P'_t]) + tr([K_t(HP'_t H^T + R)K_t^T]) \tag{3.21}$$

To minimize the tr($[P_t]$), we differentiate it with respect to K_t and set it to zero as:

$$\frac{dtr[P_t]}{dK_t} = -2(HP'_t)^T + 2K_t(HP'_t H^T + R) = 0 \tag{3.22}$$

Solving for K_t gives as:

$$K_t = P'_t H^T (HP'_t H^T + R)^{-1} \tag{3.23}$$

which is a Kalman gain equation. The Eq. (3.20) is now updated for the error covariance matrix with this optimal gain. So the updated P_t is as follows:

$$\begin{aligned} P_t &= P'_t - P'_t H^T (HP'_t H^T + R)^{-1} HP'_t \\ &= P'_t - K_t H P'_t \\ &= (I - K_t H)P'_t \end{aligned} \tag{3.24}$$

which is the update equation for the error-covariance matrix with optimal gain. The three Eqs. (3.15), (3.23) and (3.25) develop an estimate of the variable x_t. The state projection can be obtained as:

$$\hat{x'}_{t+1} = \phi \hat{x}_t$$

At the same time, we also project the error covariance matrix into the next time interval $t+1$ as:

$$e'_{t+1} = x_{t+1} - \hat{x'_{t+1}} = (\phi x_t + \omega_t) - \phi \hat{x}_t = \phi e_t + \omega_t \tag{3.25}$$

So the corresponding error-covariance matrix (using Eq.(3.13)) is computed as:

$$P'_{t+1} = E[e'_{t+1} e'^T_{t+1}] = E[(\phi e_t + \omega_t)(\phi e_t + \omega_t)^T]$$

We know that e_t and ω_t have zero cross-correlation because the noise ω_t actually accumulates between t and $t+1$. So P'_{t+1} can be written as:

$$P'_{t+1} = \phi P_t \phi^T + Q$$

which shows that it is a recursive filter that gives minimum MSE.

3.2.1 Parameter Estimation of SV Model

This subsection describes the estimation of time-varying parameters of SV models. In this case, we used the above filtering technique by three steps namely, forecasting, updating, and parameter estimation [8]. In the first step, we forecast the unobserved state vector h_t on time series observations as follows:

$$h_{t+1}^{t} = \alpha_0 + \alpha_1 h_t^{t-1} + \sum_{j=0}^{1} p_{tj} K_{tj} \eta_{tj}, \tag{3.26}$$

where the predicted state estimators $h_t^{t-1} = E(h_t|y_1, \ldots, y_{t-1})$. The corresponding error covariance matrix is defined as:

$$P_{t+1}^{t} = \alpha_1^2 P_t^{t-1} + \sigma_\omega^2 - \sum_{j=0}^{1} p_{tj} K_{tj}^2 \sum\nolimits_{tj}. \tag{3.27}$$

At this point, the innovation covariances are given as $\sum_{t0} = P_t^{t-1} + \sigma_0^2$ and $\sum_{t1} = P_t^{t-1} + \sigma_1^2$, where $P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi^t + V$, $P_0^0 = \sum_0$, $\sum_t = \text{var}(\eta_t)$, and $V = \text{var}(w_t)$. Furthermore, we use Kalman filter to measure the estimates precision, which may be shown as:

$$K_{t0} = \alpha_1 P_t^{t-1}/(P_t^{t-1} + \sigma_0^2) \text{ and } K_{t1} = \alpha_1 P_t^{t-1}/(P_t^{t-1} + \sigma_1^2). \tag{3.28}$$

The second step deals with updating results while we have a new observation of y_t at time t. The prediction errors of the likelihood function are computed using the following relations:

$$\eta_{t0} = y_t - \beta - h_t^{t-1} \text{ and } \eta_{t1} = y_t - \beta - h_t^{t-1} - \mu_1. \tag{3.29}$$

For estimating the parameters, we complete the updating step by assessing the time-varying probabilities (for $t = 1, \ldots, m$):

$$p_{t1} = \frac{p_1 d_1(t|t-1)}{p_0 d_0(t|t-1) + p_1 d_1(t|t-1)}$$
$$\text{and } p_{t0} = 1 - p_{t1},$$

where $d_j(t|t-1)$ is considered to be the conditional density of y_t, given the previous observations $y_1, \ldots, y_{t-1}$.

Since the observation noise of this model is not fully Gaussian, it is computationally difficult to obtain the exact values of $d_j(t|t-1)$. Hence, we use a good approximation of $d_j(t|t-1)$ that provides Normal density which is: $N(h_t^{t-1} + \mu_j, \sum_{tj})$, for $j = 0, 1$ and $\mu_0 = 0$.

Finally, we estimate the parameters ($\Theta = (\alpha_0, \alpha_1, \sigma_w, \beta, \sigma_0, \mu_1, \sigma_1)'$) by maximizing the expected likelihood, where the MLE is represented as:

$$lnL_Y(\Theta) = \sum_{t=1}^{m} ln\big(\sum_{j=0}^{1} p_j d_j(t|t-1)\big). \tag{3.30}$$

3.3 The Superposed Ornstein-Uhlenbeck Process

This subsection deals with a stochastic differential equation arising from the superposition of independent Ornstein–Uhlenbeck processes driven by a $\Gamma(a, b)$ process. Superposition of independent $\Gamma(a, b)$ Ornstein-Uhlenbeck processes offers analytic flexibility and provides a class of continuous time processes capable of exhibiting long memory behavior. The methodology is applied to the high frequency data in order to simulate them [7].

A process is an Ornstein-Uhlenbeck process if it is *cádlág* (i.e. it is right continuous and has a left limit at every point) and satisfies the stochastic differential equation,

$$dS_t = -\lambda S_t dt + dZ_{\lambda t}, S_0 > 0, \quad \lambda \in \mathbb{R}^+. \tag{3.31}$$

where S_t is a continuous and non-negative stochastic process, $Z = \{Z_{\lambda t}\}_{t \geq 0}$ is a Lévy process and the rate parameter λ is a positive number. The process $Z = \{Z_{\lambda t}\}_{t \geq 0}$ is termed the background driving Lévy process (BDLP).

If we consider the unusual timing in the BDLP, the solution to Equation (3.31) is

$$S_t = \mathrm{e}^{-\lambda t} S_0 + \int_0^t \mathrm{e}^{-\lambda(t-q)} dZ(\lambda q). \tag{3.32}$$

In order to obtain the analytic flexibility and to ensure correlation structures for the process S_t, we consider the sum of two independent Ornstein-Uhlenbeck processes in Eq. (3.32), where each component process is an independent Ornstein–Uhlenbeck process with rate parameters λ_1 and λ_2:

$$\begin{aligned} S_t = w_1 S_1 \mathrm{e}^{-\lambda_1 t} &+ \int_0^t w_1 \mathrm{e}^{-\lambda_1 (t-q)} dZ(\lambda_1 q) \\ + w_2 S_2 \mathrm{e}^{-\lambda_2 t} &+ \int_0^t w_2 \mathrm{e}^{-\lambda_2 (t-q)} dZ(\lambda_2 q), \quad t \geq 0. \end{aligned} \tag{3.33}$$

The approach adopted in this work specified a parametric form for the marginal distribution of Eq. (3.33) and then work out the corresponding distribution of BDLP. We will do this by specifying a distribution for the Lévy process and simulate the data via BDLP.

3.3.1 Parameter Estimation

A stochastic process $S = \{S_t, t \geq 0\}$ with parameters u and v is a Gamma process if it fulfills the following conditions:

1. $X_0 = 0$.
2. The process has independent increments.
3. The process has stationary increments.
4. For $m < t$, the random variable $S_m - S_t$ has $\Gamma(a(t-m), b)$ distribution.

Recall that a random variable S has a Gamma distribution $\Gamma(u, v)$ with rate and shape parameters $u > 0$ and $v > 0$ respectively, if its density function is given by:

$$f_S(s; u, v) = \frac{v^u}{\Gamma(u)} s^{u-1} \mathrm{e}^{-vs}, \quad \forall \quad s > 0, \tag{3.34}$$

where Γ denotes the Gamma function. The $\Gamma(a,b)$ process has a Lévy density given by $u(x) = ax^{-1}e^{-bx}, x \neq 0$. The $\Gamma(a,b)$ distribution is known to be self-decomposable. Therefore, taking into account the BDLP, Z has a Lévy density given by

$$w(x) = uve^{-vx}, x \neq 0$$

and the associated cumulative function is given as:

$$k(m) = \frac{uv}{v+m} \tag{3.35}$$

So it is concluded that the BDLP for the $\Gamma(a,b)$ process is a compound Poisson process which jumps a finite number of times in every compact time interval. Now a proposition is followed as:

Suppose that $\{Z_t\}_{t\geq 0}$ *is a Lévy process such that* $E(Z(1)) = \mu < 1$ *and Var* $(Z(1)) = \sigma^2 < 1$. *Assume that* $\lambda_1, \lambda_2 > 0$*, then the followings are true*:

1. $E(X_0) = \mu$
2. Var $(X_0) = \frac{\sigma^2}{2}$

From this proposition, the parameters μ and σ^2 can be expressed as related to a and b:

$$a = \frac{2\mu^2}{\sigma^2} \text{ and } b = \frac{2\mu}{\sigma^2}$$

Now the intensity parameter λ_1 is estimated by taking into account the autocorrelation function of (3.33) i.e.,

$$\rho(k) = w_1\mathrm{e}^{-\lambda_1|k|} + w_2\mathrm{e}^{-\lambda_2|k|}, \tag{3.36}$$

where $w_1 + w_2 = 1$ for any $k \in \mathbb{R}^+$.

From (3.36), λ_1 can be computed by assuming $\lambda_1 = \lambda_2$ and $k = 1$. This results in

$$\lambda_1 = -log(\hat{\rho}(1)) \tag{3.37}$$

where $\hat{\rho}(1)$ denotes the empirical autocorrelation function of lag 1 based on the time series data sets. Once λ_1 is estimated, λ_1 is adjusted to obtain λ_2 in order to fit the superposed $\Gamma(u, v)$ Ornstein-Uhlenbeck model. Then the data is simulated the based on the parameters $u, v, \lambda_1, \lambda_2$ that are estimated from the original data. For the details of the methodology, the readers are referred to [8] and references therein.

Chapter 4

Results & Discussion

This chapter focuses on the analysis of volatility models and a simulation technique that are applied to the high frequency datasets contained extreme fluctuations. The estimates of time-varying parameters and one-step ahead predicted volatility are presented to detect the extreme fluctuations. The analyses were performed by R and Python programs.

4.1 Analysis of Volatility models

This section deals with the results of volatility models that are applied to the daily returns of financial market data, emergent market data, developed market data, minute by minute returns data, and some geophysical data. In chapter 2, Figs. 2.2-2.5 provide a good perspective on the trending direction or risk management of the high frequency returns of stock markets or extreme fluctuations of seismic events. The extreme fluctuations that occurred is evident in the large spikes in the figures. We see that the volatility of data changes dramatically at a short interval and that the periods of high volatility are sometimes correlated. This establishes that the volatility itself is very volatile. The fluctuations of stock returns or seismic events typically exhibit the volatility clustering that suggests that the current information is highly correlated with past information at different levels.

4.1.1 Results of GARCH Model

The GARCH volatility model is used to predcit the volatiltiy of high frequency financial market retruns in U.S.A. such as the Citi Bank, the Microsoft company, the Bank of America, the Discover Financial Services, the INTEL semiconductor manufacturing company, the IBM hardware company, the Walmart retail company, the IAG stock exchanges stock markets and among others. In this dissertation, the estimated parameters, standard errors, and forecasting behaviors of two stock companies: Citi Bank and Microsoft stock exchanges are presented (see Tables 4.1-4.4 and Figs. 4.1-4.2). The results of other stock markets data can be found in [22].

The estimates of the parameters a_0, a_1, and b_1 of the GARCH (1, 1) model are stable, as the GARCH-statistic shows (see Tables 4.1-4.4). Also, the estimated standard errors of the parameters in most cases are small. The smaller p-values ($<$ significance level) provide strong evidence that the GARCH $(1, 1)$ model with the specified parameters is a good fit for our data. The volatility level of persistence can be determined by non-negative parameters a_1 and b_1 from these tables. It is observed that the constraint $(a_1 + b_1)$ is less than 1, which allows for the existence of a stationary solution. This also supports the results of stationary tests in subsection 2.2.

The standardized residuals (R) tests for the Citi Bank and the Microsoft stock exchanges are summarized in Tables 4.3-4.4. The Jarque-Bera and Shapiro-Wilk tests of Normality strongly reject the null hypothesis that the white noise innovation process η_t is Gaussian. The p-values (> 0.05) of Ljung-Box test for squared residuals (at lag $10, 15, 20$) and LM-Arch test suggest that the model fits the data well, with the exception of the non-normality of η_t. It is because the null hypothesis cannot be rejected at any reasonable level of significance.

To facilitate the understanding of forecasting concepts, we superimpose the plot of one-step-ahead predicted volatility with high frequency stock market data in Figs. 4.1-4.2. The predicted log volatility is displayed as grey color across the original time series (blue color). It visually shows how the values of predicted volatility differ over time and able to detect the extreme fluctuations.

Limitations

There are some limitations of GARCH (1,1) model itself in the application to the high frequency financial datasets. It is evident that the positive and negative high frequency returns have the same effect, because volatility depends on squared returns. The model does not help to understand the source of variations of a financial time series. That is to say, the cause of the variation in volatility is unknown. The model provides only a mechanical way to describe the behavior of conditional variance. The grey lines during financial crashes (in Figs. 4.1-4.2) indicate that the model tends to over predict the volatility, because it is often restrictive due to the tight constraints on parameters ($\alpha_0, \alpha_1, \beta_1 \geq 0$).

Table 4.1: **GARCH statistics for CITI Bank stock exchange**

Parameter	Estimate	Error	t-statistic	p-value
a_0	0.221	0.0216	10.216	< 2E-06
a_1	1.104	0.0418	26.384	< 3.0E-05
b_1	0.115	0.0289	3.989	6.6E-05

Table 4.2: **GARCH statistics for the Microsoft stock exchange**

Parameter	Estimate	Error	t-value	p-value
a_0	0.019	0.0617	3.165	0.0015
a_1	0.165	0.0372	4.439	9.04E-06
b_1	0.845	0.0265	32.212	<2E-16

Table 4.3: **Standardized Residuals Tests for CITI Bank stock exchange**

	Residuals	Tests	Statistics	p-value
Jarque-Bera Test	R	χ^2	1404.85	0
Shapiro-Wilk Test	R	W	0.92670	0
Ljung-Box Test	R^2	Q(10)	83.5710	9.9E-14
Ljung-Box Test	R^2	Q(15)	121.130	0
Ljung-Box Test	R^2	Q(20)	134.161	0
LM-Arch Test	R	TR^2	72.9724	8.8E-11

Table 4.4: **Standardized Residuals Tests for Microsoft stock exchange**

			Statistics	P-value
Jarque-Bera Test	R	χ^2	2915.06	0
Shapiro-Wilk Test	R	W	0.86025	0
Ljung-Box Test	R^2	Q(10)	9.13460	0.51937
Ljung-Box Test	R^2	Q(15)	12.0977	0.6716
Ljung-Box Test	R^2	Q(20)	17.4958	0.6205
LM Arch Test	R	TR^2	10.27118	0.59218

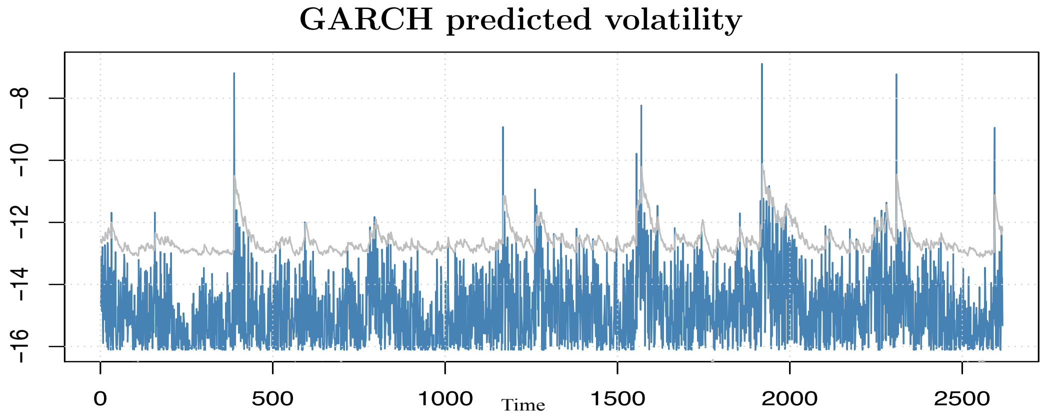


Figure 4.1: One-step-ahead predicted volatility (grey color) with Citi Bank stock exchange series (blue color).

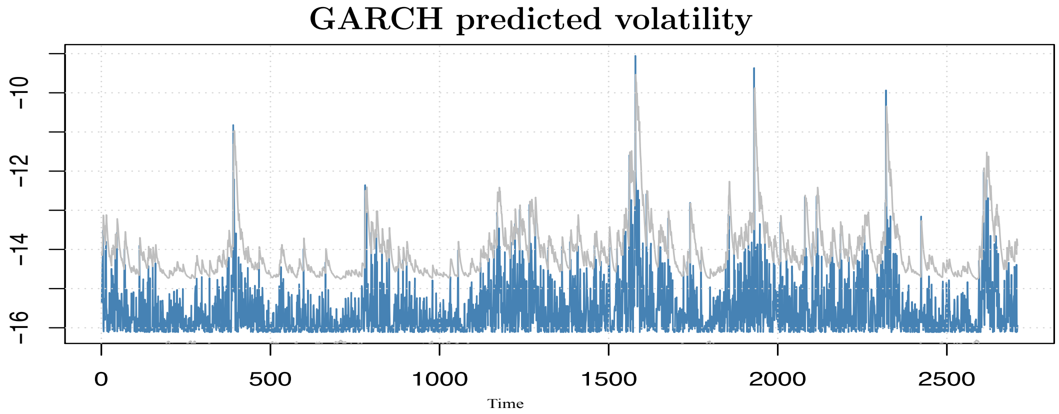


Figure 4.2: One-step-ahead predicted volatility (grey color) with Microsoft stock exchange series (blue color).

4.1.2 Results of APARCH Model

In this subsection, the results of APARCH model are discussed, which can solve a limitation of GARCH model. The leverage effects states that the large negative returns of high frequency data appear to increase volatility compared to positive returns of the same magnitude. The reason is that the GARCH model estimates the volatility as a function of squared past values, so the positive or negative returns have the same effect. We avoid this issue by using APARCH model that replaces the square function with a flexible class of non-negative functions. The estimation of time varying parameters and one-step-ahead predicted volatility for Microsoft stock market is summarized in Table 4.5. We see that the estimate of δ is 1.360 with a standard error of 0.488, so there is strong evidence that δ is not 2, which confirms the APARCH condition (see subsection 3.1.2). Also, $\hat{\gamma_1}$ is -0.99 with a standard error of 0.109 and p-values of 2E-16, so we reject the null hypothesis that $\hat{\gamma_1} = 0$, and conclude that there is a statistically significant leverage effect in the process. Fig. 4.3 shows the original high frequency returns (blue color) and corresponding predicted log-volatility (grey color), where we conclude that the APARCH volatility has higher ability to detect the extreme fluctuations compared to GARCH volatility.

Table 4.5: **APARCH statistics for the Microsoft stock exchange**

Parameter	Estimate	Error	t-statistics	p-value
a_0	0.117	0.0300	3.920	8.8E-05
a_1	0.037	0.0117	3.227	0.0012
b_1	0.087	0.026	33.08	<2E-16
γ_1	-0.999	0.109	-9.135	<2E-16
δ_1	1.360	0.428	3.175	0.0015

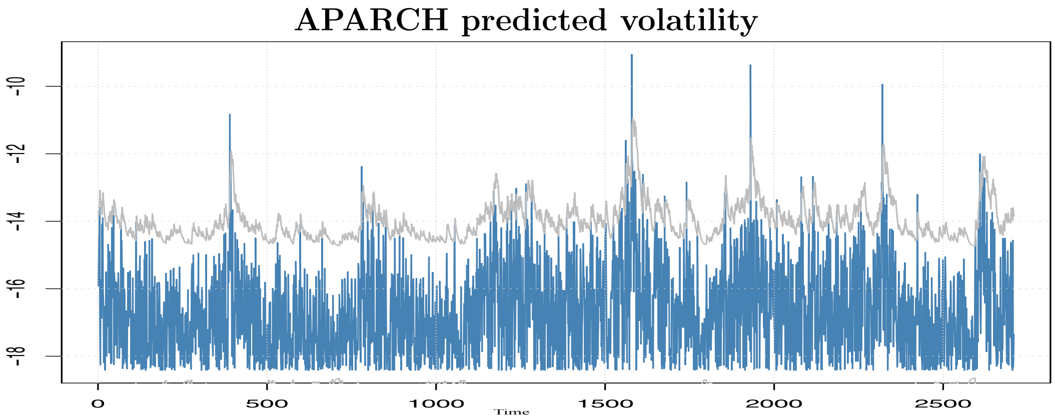
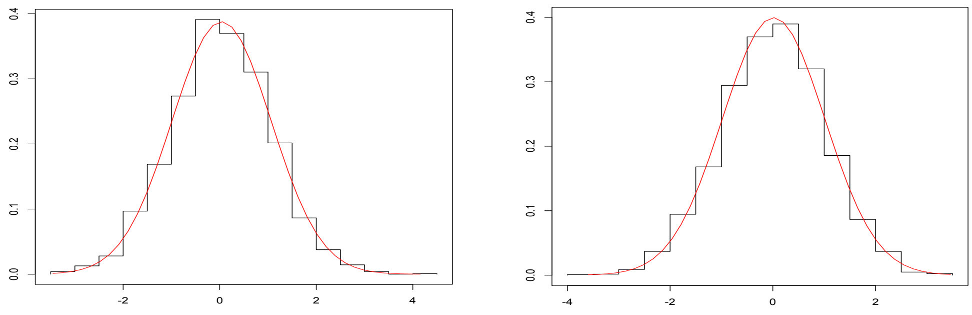


Figure 4.3: One-step-ahead predicted volatility (grey color) with Microsoft stock exchange series (blue color).

4.1.3 Results of Stochastic Volatility Model

This subsection deals with the stochastic nature of high frequency data including extreme fluctuations. The stochastic volatility model is analyzed on the Lehman Brothers collapse data, Flash Crash events data, natural earthquakes, and mining explosions data. The GARCH and APARCH models differ from the SV model in the sense that, unlike the SV model, they do not have any stochastic noise. The SV model is characterized by the fact that it invariably contains its probability density function. The histograms of natural earthquakes and mining explosions data are presented in Fig. 4.4. The maximum likelihood is computed by taking into consideration the conditional Normal distribution of the datasets.

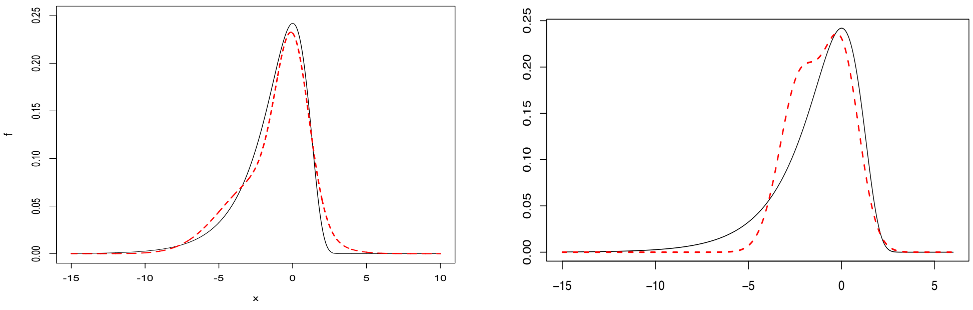
For the datasets described in this paper, the ARCH Normality assumption is kept on volatility η_t. Therefore, the $log\eta_t^2$ is distributed as the logarithm of a chi-squared random variable with one degree of freedom whose mean is -1.27 and variance is 4.93. In Fig. 4.5 (a, b), we see a clear representation of the density of $\log(\chi_1^2)$ and fitted Normal mixture of financial time series. It is obvious that the density of financial time series is skewed with a long tail on the left.



(a) Natural Earthquake

(b) Mining Explosion

Figure 4.4: The histograms of geophysical time series and the fitted Normal density (red color).



(a) The JPMorgan Chase.

(b) The Bank of America

Figure 4.5: Density plot of $\log(\chi_1^2)$ (solid curves) and the fitted Normal mixture distribution (dashed red curves) from (a) the JPMorgan Chase and (b) the Bank of America stock exchanges.

The parameters from time-varying Eqs. (3.9) and (3.10) were initialized in order to observe the performance of the SV algorithms for each dataset described above. We set the initial values to be $\alpha_0 = 0, \alpha_1 = 0.95, \sigma_\omega = 0.5, \sigma_0 = 1, \mu_1 = -3, \sigma_1 = 2$ and β (the mean of the observations). In order to maximize the Eq. (3.30), the innovation processes for Eqs. (3.9) and (3.10) were fitted to the data by taking into consideration this time-varying probability ($p_1 = 0.5$).

Tables 4.6-4.9 summarize the estimation of parameters ($\alpha_0, \alpha_1, , \sigma_\omega$, β, σ_0, μ_1 and σ_1). The estimated error in these tables makes two things evident: firstly, the estimates are close to the true parameters; secondly, the algorithm of the SV model aligns with the financial and the geophysical data. The variance σ_w^2 of the log-volatility process measures the uncertainty about the future volatility of data. If the value of σ_w^2 is zero, it is not possible to identify the SV model. The parameter α_1 is considered as a measure of the persistence of shocks to the volatility. Tables 4.6-4.9 indicate that α_1 is less than 1, which suggests that the latent volatility process and y_t are stationary. This confirms the results of section 2.2.

In these tables, we also notice that α_1 is near to unity and σ_w^2 is different from 0, which means that the evolution of volatility is not smooth over time. This also suggests that the time series used in this paper could be heteroscedastic by nature, that is, there is a non-constant conditional volatility over time. So, it is very useful to control the risk or to mitigate the effect of hazards. Figs. 4.6-4.9 show forecasting behaviors of two stock companies, one natural earthquake dataset and one mining explosion dataset. The original high frequency data (blue color) are shown with corresponding log-predicted volatility (red color) in these figures. It is clear the stochastic volatility is able to detect the financial crashes or extreme fluctuations for both financial and geophysical data. For the results of other datasets, the reader is referred to [34] and [22].

Table 4.6: Summary statistics for the JPMorgan Chase stock exchange.

Parameter	Estimate	Standard Error
α_0	-0.047	0.057
α_1	0.961	0.014
σ_ω	0.252	0.050
β	-13.15	1.394
σ_0	0.947	0.047
μ_1	-2.348	0.083
σ_1	1.050	0.055

Table 4.7: Summary statistics for Bank of America stock exchange.

Parameter	Estimate	Standard Error
α_0	-0.018	0.057
α_1	0.961	0.014
σ_ω	0.228	0.042
β	-14.06	1.451
σ_0	0.952	0.046
μ_1	-2.242	0.092
σ_1	1.082	0.056

Table 4.8: Summary statistics for Earthquake data from TUC station.

Parameter	Estimate	Standard Error
α_0	0.032	0.021
α_1	0.998	0.001
σ_ω	0.428	0.034
β	-11.97	1.720
σ_0	0.642	0.038
μ_1	-2.474	0.085
σ_1	2.323	0.051

Table 4.9: Summary statistics for Explosion data from TUC station.

Parameter	Estimate	Standard Error
α_0	0.021	0.051
α_1	0.984	0.003
σ_ω	0.693	0.017
β	-10.14	1.935
σ_0	2.2E-6	0.050
μ_1	-2.300	0.073
σ_1	2.062	0.045

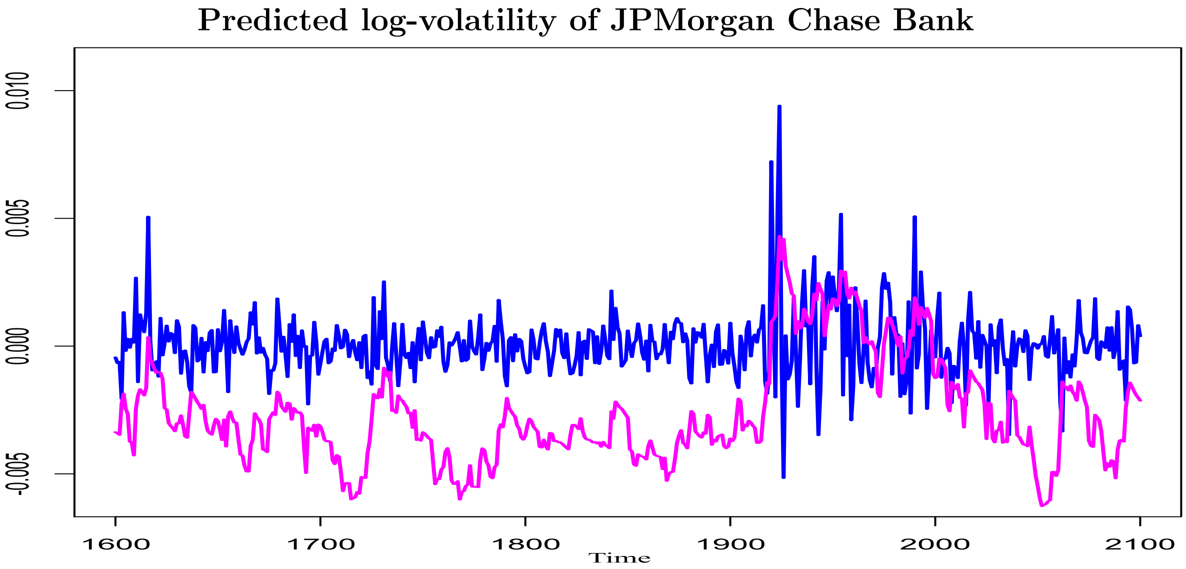


Figure 4.6: One-step-ahead predicted log-volatility with five hundred observations from JPMorgan Chase stock.

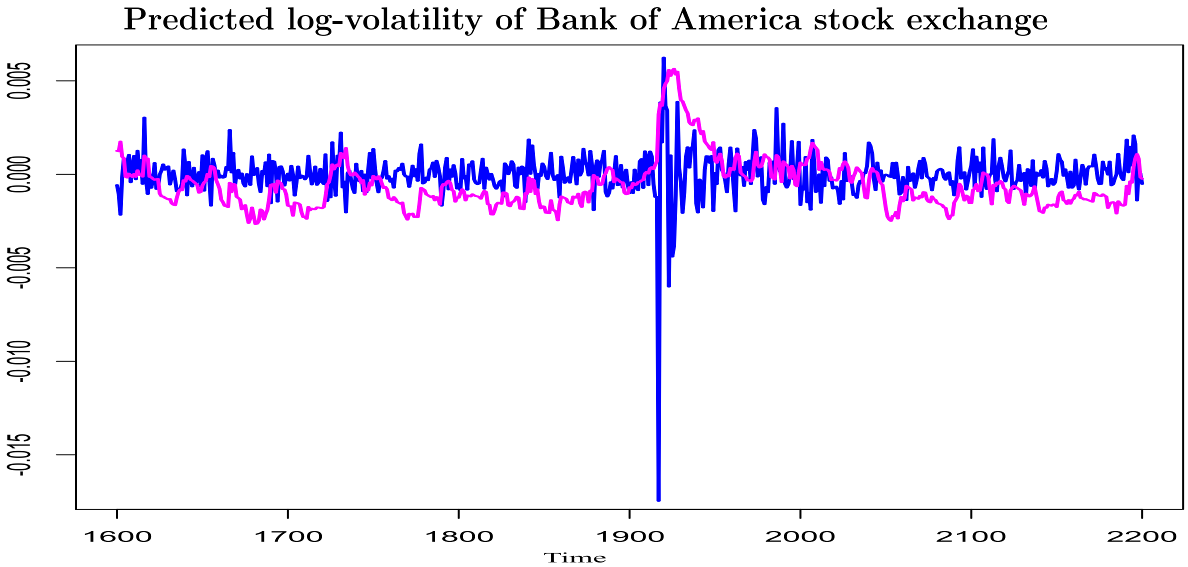


Figure 4.7: One-step-ahead predicted log-volatility with six hundred observations from Bank of America stock.

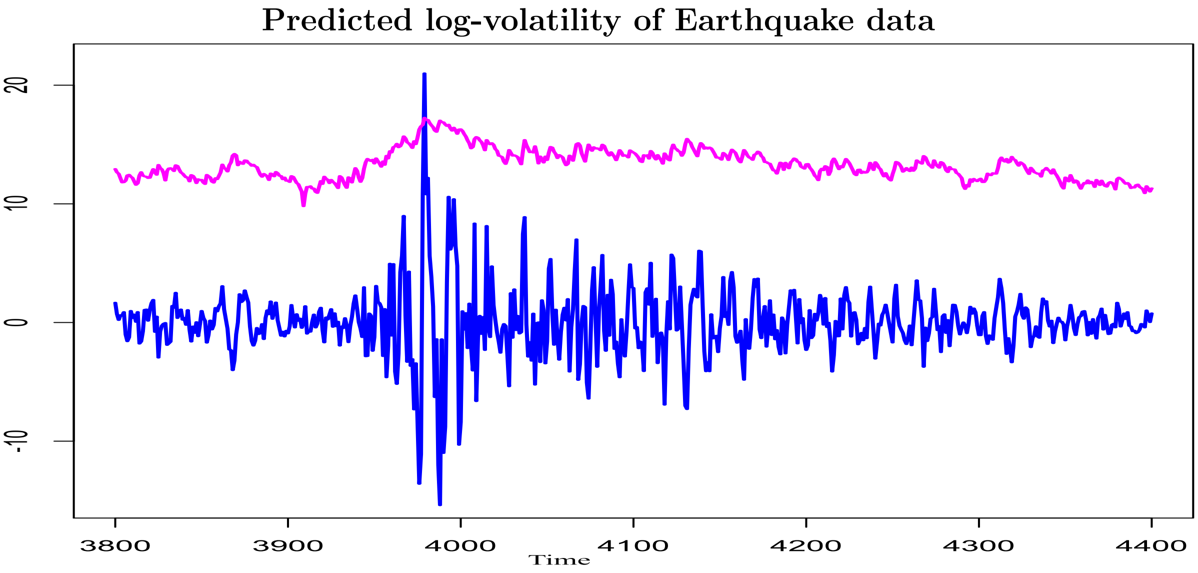


Figure 4.8: One-step-ahead predicted log-volatility with sixteen hundred observations from natural earthquake event.

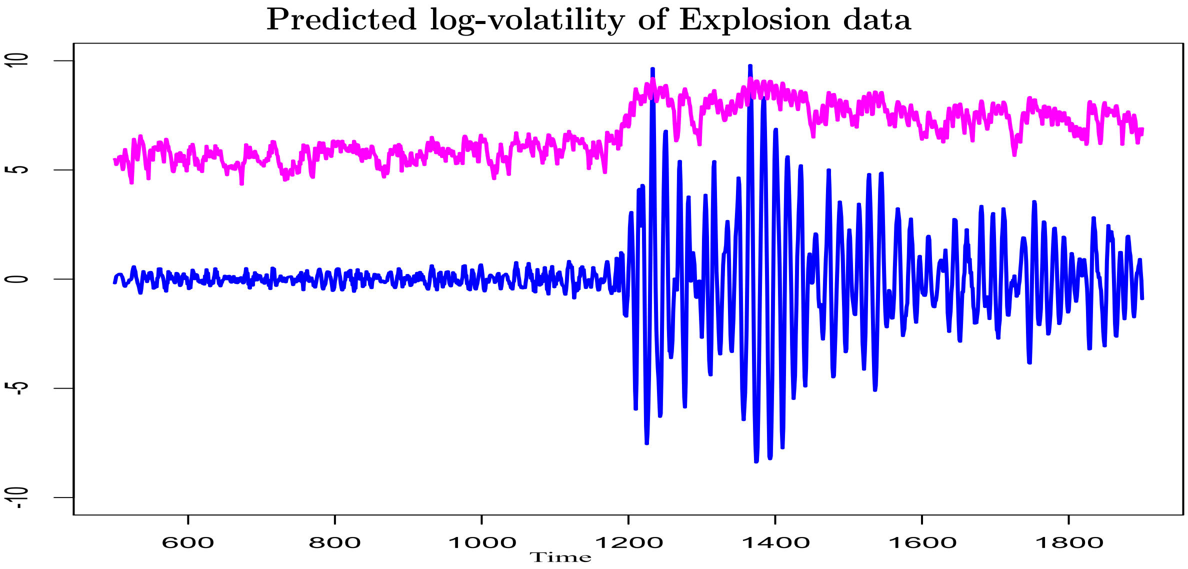


Figure 4.9: One-step-ahead predicted log-volatility with fourteen hundred observations from mining explosion event.

4.2 A new approach for fitting SV model

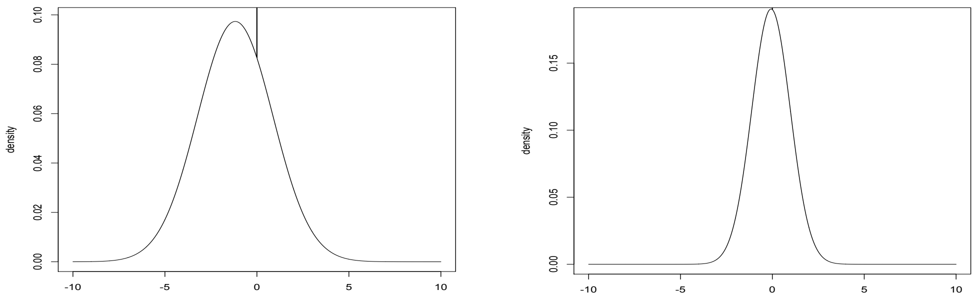
This subsection deals with a new approach for fitting SV model, when the high frequency data contain noisy observations. The stochastic volatility (SV) model described in section 4 is used to estimate the volatility of a log-squared observations via MLE. However, we observed that the financial data does not always fit into SV methodology and so convergence of MLE is not always guaranteed. It is because the high frequency data contain some unusual observations and zero returns as well. Thus, the stochastic volatility model is not always applicable to forecast data volatility. To avoid this problem, the model described in section 3.3 is employed to simulate the data based on their dynamic behavior.

We first used the superposed Ornstein-Uhlenbeck model with the Lévy driven process. The BDLP for the superposed $\Gamma(u, v)$ Ornstein-Uhlenbeck model has been shown to be a compound Poisson process ([7]). The data is simulated using the solution of the stochastic differential equation via the BDLP. A Matlab program was developed to simulate the process using different time steps. Table 4.10 summarizes the estimation of parameters $\lambda_1, \lambda_2, w_1, w_2$ for NASDAQ, S&P 500, BVSP and SETI stock exchanges when the superposed $\Gamma(u, v)$ Ornstein-Uhlenbeck model was applied to the data. We obtained λ_2 by adjusting λ_1 in order to fit the superposed $\Gamma(u, v)$ Ornstein-Uhlenbeck model.

Table 4.10: Simulation results for the developed and emergent market indices.

Stock Exchange	λ_1	λ_2	w_1	w_2	RMSE
NASDAQ	0.003	8	0.47	0.53	1.34
S&P 500	3.2E-04	32	0.30	0.70	1.99
BVSP	0.0017	12	0.20	0.80	1.66
SETI	0.0074	38	0.30	0.70	0.32

The simulated data mimics the original financial time series data. This is observed from the estimates and RMSE in Table 4.10. Also, the simulated data exhibits long memory behavior which facilitate prediction using the SV model. In order to fit the SV model, the distribution of errors of the data are analyzed (Fig. 4.10). The following figures show that the error terms follow almost normal distributions. So the ARCH Normality condition is assumed on the basis of volatility η_t described in Eq. (3.1).



(a) NASDAQ

(b) BVSP

Figure 4.10: Density plot of the fitted Normal mixture distribution from (a) NASDAQ and (b) BVSP stock exchanges.

The above figures show that the density of financial time series is skewed with a little tail on the left. Table 4.11 and Figs. 4.11 - 4.12 summarize the estimation of parameters (α_0, α_1, σ_w, λ, σ_0, ϕ_1 and σ_1) and predicted volatility with ±2 errors for NASDAQ and BVSP stock market data. For the results of other stock market datasets, we refer the reader to [22]. The advantage of this methodology is that the estimates obtained are stable around the true value and also the low errors indicate that the estimation procedure is accurate, therefore producing a higher forecasting accuracy. Thus, our estimation algorithm is feasible with large data sets and have good convergence properties.

Table 4.11: Summary statistics for the developed and emergent market indices.

Parameter	S&P 500	Standard Error	BVSP	Standard Error
α_0	9.53E-02	0.07950	8.906E-02	0.0227
α_1	9.79E-01	0.00870	9.807E-01	0.0029
σ_ω	0.01740	3.46E-02	7.845E-02	0.0041
β	5.54E+00	3.19750	1.405E+01	0.7599
σ_0	1.863E-07	0.02150	5.119E-07	0.0036
μ_1	8.490E-02	0.03890	-7.087E-02	0.1199
σ_1	2.490E-01	0.03332	1.048E+00	0.0845

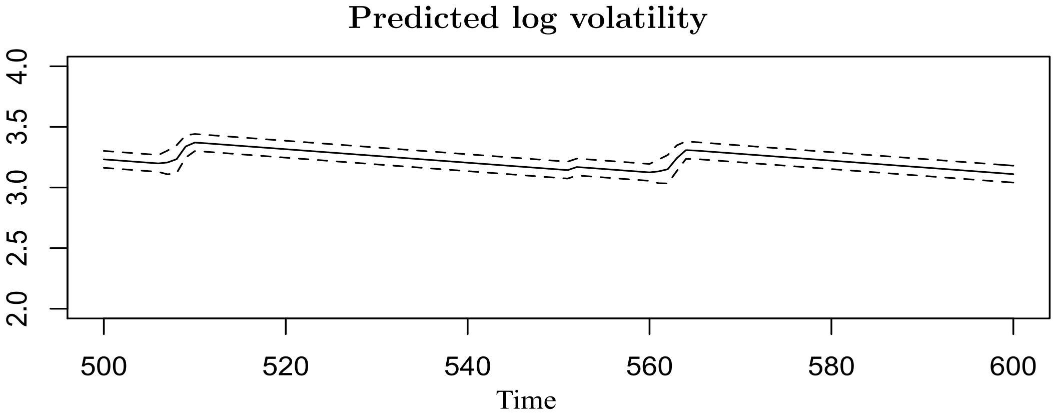


Figure 4.11: One-step-ahead predicted log volatility (solid lines), with ±2 standard prediction errors (dashed lines) for one hundred observations from NASDAQ stock exchange.

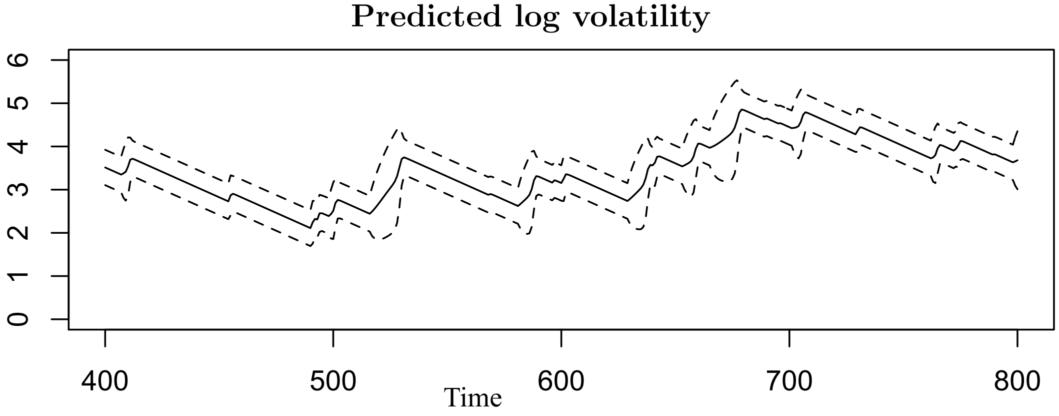


Figure 4.12: One-step-ahead predicted log volatility (solid lines), with ±2 standard prediction errors (dashed lines) for four hundred observations from BVSP stock exchange.

Chapter 5

Other Researches

This chapter deals with our other researches on different types of data. The machine learning (ML) techniques (supervised and upsupervised), Dynamic Fourier models, wavelet models are studied on public health (e.g., breast cancer data, heart disease data, diabetes data, autism spectrum disorder data), financial and geophysical time series. In this chapter, we briefly discuss on the machine learning techniques when they are applied to the diabetes and heart disease datasets. The readers are referred to [13], [30], [45] and [46] for the details of data backgrounds, machine learning algorithms, Dynamic Fourier techniques, wavelet models and the results of other datasets.

5.1 Machine Learning Algorithms

Machine learning (ML) techniques help to detect and classify chronological diseases like cancer, heart disease, tumor, diabetes and so on. These techniques are useful to many statistical, probabilistic, and optimization processes that allow computers to consider past observations and to detect the pattern of the disease. Researchers have recently attempted to use ML techniques (especially the supervised and unsupervised ones) in this case. The reason is that the techniques help us to identify the sources and to order the variables in terms of importance in causing chronological disease.

In this study, we first did the exploratory analysis to see the association, correlation and frequency distribution of the features, which are very effective in capturing the characteristic and patterns of data correctly. Then the data is randomly split into training set (67% for building a predictive model) and test set (33% for evaluating the model). The prediction mean squared error (PMSE), miss-classification rate (MCR), and prediction accuracy using the Receiver Operating Characteristic (ROC) curve are computed in order to validate the fitted model with data. The analyses were performed by Python programs.

5.1.1 Analysis of Fitted models

The machine learning models are trained with data to accurately predict whether an individual has been diagnosed with disease (breast cancer, heart disease, diabetes or Autism spectrum disorder). The models used in this research are as: Logistic regression, Ridge regression, Principal component regression, Random Forest and Support vector machine. However, the problem of machine learning techniques is to fit them into the data due to over-fitting, under-fitting, and bias-variance issues. In this case, the regularization technique and cross-validation are used to estimate the tuning parameters of each model corresponding to the low mean squared error. The adequacy and predictive ability of the data are determined by computing sensitivity, specificity, accuracy and confidence interval. Now the results of five ML algorithms are briefly presented when they are applied to diabetes and heart disease datasets.

Results of Logistic regression

The logistic regression technique have been used for train data to build a predictive model. In this case, the lasso regularization (L_1 norm) is employed to obtain the tuning parameter λ via cross-validation. The L_1 penalty is used for both variable selection and shrinkage, since it has the effect of forcing some of the coefficient estimates to be zero. Table 5.1 represents important predictors using the best predictive model with L_1 penalty. It is clear that *number of times pregnant (preg)*, *plasma glucose concentration (plas)*, *body mass index (mass)*, *diabetes pedigree function (pedi)*, and *age* are important predictors for identifying a diabetes patient and *exercise induced angina (Exang)*, *number of major vessel (Ca)*, *the slope of the peak exercise ST segment (Slope)*, *thalassamia (Thal)* , *chest pain (Cp)*, and

ST depression induced by exercise relative to rest (oldpeak) are important predictors for identifying a heart disease patient. The test data is predicted using this predictive model and evaluated the model using different metrics.

Variables	Coefficients
preg	0.0323
plas	0.0250
pres	0
skin	0
insu	0
mass	0.0407
pedi	0.2950
age	0.0121

Table 5.1: Coefficients of important predictors using LGR (L_1) model for Diabetes data.

Results of Ridge regression

The Logistic regression model with L_2 penalty term (Ridge regression) is also studied on the datasets. An advantage of L_2 penalty is that it overcomes the multicollinearity issue of the datasets. It has more power to reduce the over-fitting issue of data compared to L_1 regularization. The parameter λ is tuned (optimized) until we find a model that fits well to the train data. In this case, the tuning parameter is selected by 10-fold cross validation procedure. From Table 5.3, it is clear that the prediction mean squared error and miss-classification rate of this model are very low for both diabetes and heart disease data.

Variables	Coefficients
Age	0
Sex	0
Cp	0.2847
Trestbps	0
Chol	0
Fbs	0
Restecg	0
Thalach	-0.0073
Exang	0.4792
Oldpeak	0.1605
Slope	0.2172
Ca	0.4164
Thal	0.3021

Table 5.2: Coefficients of important predictors using LGR(L_1) model for heart Disease data

Results of Principal Component Regression

A dimension reduction tool, namely principal component regression (PCR) is studied to reduce the set of predictors of datasets. An advantage of principal component analysis is that it transforms the high dimensional correlated data into uncorrelated data with same amount of variation. In this case, the entire dataset is transformed into three principal components to build the predictive model. The first principal component contains most of the variability in the data. PCR also overcomes the multicollinearity issue of the datasets used in this study. We obtained a very good prediction accuracy on test data, which are 88.17% for heart disease data. The predictive performance of PCR compared to the other methodologies are shown in Table 5.3.

Results of Random Forest

The Random forest model is fitted into train data with optimized number of trees. For the heart disease dataset, the tuning parameters are 500 trees and 3 sampled variables at each split. We obtained a very good prediction accuracy on test data, which are 74.80% for diabetes disease. An important feature of random forest is that we made an order of the predictors in terms of importance using Mean Decrease Accuracy and Mean Decrease Gini indices [45]. From Fig. 5.1, it is clear that the plasma-glucose-concentration (*plas*) is the most important variable and triceps-skin-fold-thickness (*skin*) is the least important variable in causing diabetes.

Results of Support Vector Machine

Last, a support vector machine (SVM kernel) technique is studied to classify the disease datasets. The reason of using SVM kernel function is that it maps the non-linear separable dataset into a higher dimensional space where a hyperplane is able to separate the classes (target variable) linearly. To fit the model, the dataset is first standardized and trained with 10-fold cross-validation procedure. The model is evaluated with different cost levels

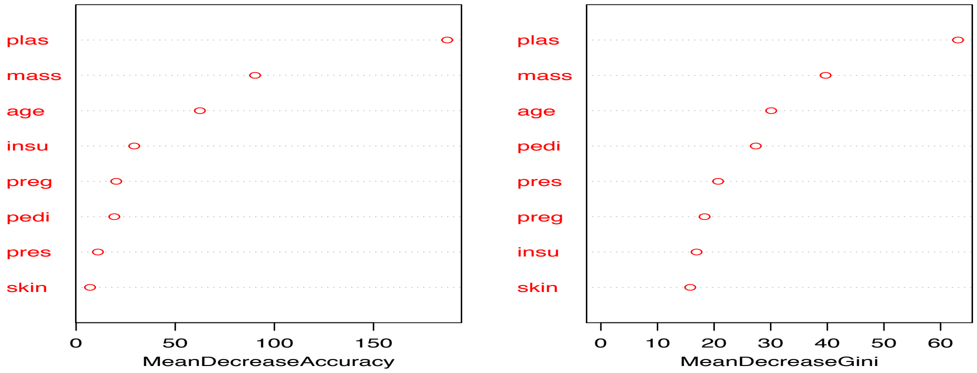


Figure 5.1: Variable importance plot using Random Forest model for Diabetes data.

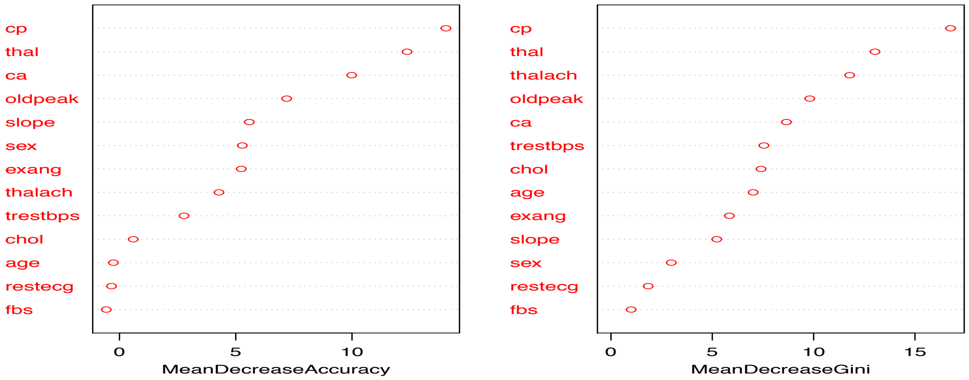


Figure 5.2: Variable importance plot using Random Forest model for Heart-Disease data

(C) in order to obtain the best predictive model with optimal cost. The highest accuracy is achieved when γ is 0.01 for diabetes dataset and 0.1 for heart disease dataset. Table 5.3 summarizes the prediction mean square error (PMSE) for the predicted probabilities and miss-classification rate (MCR) for both diabetes and heart disease datasets.

Table 5.3: **Model Evaluation**

Models	Diabetes Dataset		Heart Disease Dataset	
	PMSE	MCR	PMSE	MCR
LGR-L_1	0.173	0.256	0.151	0.172
RGR	0.174	0.264	0.149	0.193
PCR	0.194	0.308	0.107	0.118
RF	0.169	0.252	0.145	0.172
SVM	0.169	0.240	0.144	0.162

5.1.2 Model Evaluation

This subsection presents the accuracy of our predictive models obtained from above analysis. Tables 5.4 & 5.5 show the sensitivity, specificity, accuracy and confidence interval with 95% significance level for both diabetes and heart disease data. The sensitivity of models measures the proportion of people with the disease (diabetes and heart disease) who will have a positive result. So the highly sensitive test is one that correctly identifies patients with a disease. For example, the Logistic regression (L_1) test is 75.81% sensitive for diabetes data, that is, the model classifies 75 individuals out of every 100 patient correctly who have the diabetes (see Table 5.4). On the other hand, the specificity measures the proportion of people without the disease (diabetes or heart disease) who will have a negative result. For instance, the logistic regression (L_1) for diabetes test is 70.31% specific, meaning that the model identifies 70% of individuals correctly who do not have the diabetes (see Table 5.4). We plotted the ROC curves between True Positive Rate (Y-axis) and False Positive Rate

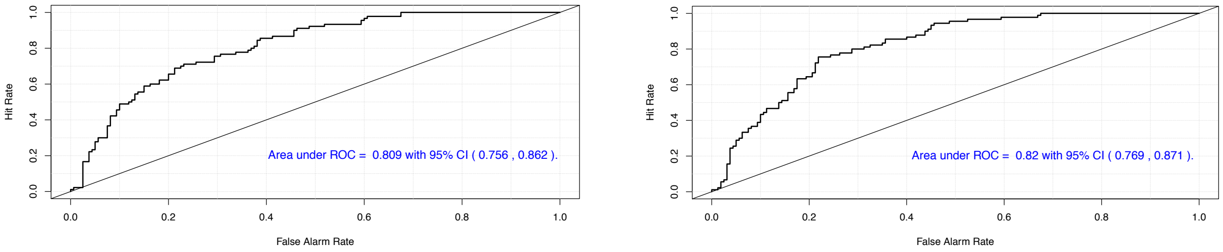
(X-axis) of our predictive models in Figs. 5.3 and 5.4. In these figures, the diagonal line represents the threshold (0.5) of ROC curves. The areas under the curve are almost 0.8 for all models fitted with diabetes data. For heart disease data, the area under the ROC curve is 0.942 for principal component regression. Thus it is concluded that ML techniques have good predictive ability on diabetes and heart disease data.

Table 5.4: Model Evaluation using ROC Curve for Diabetes data.

Models	Sensitivity (%)	Specificity (%)	Accuracy (%)	Conf. Interval (%)
LGR-L_1	75.81	70.31	74.44	(68.52 - 79.69)
RGR	85.19	60.87	74.00	(68.10 - 79.32)
PCR	87.50	40.00	70.40	(64.32 - 75.99)
RF	86.25	54.44	74.80	(68.94 - 80.06)
SVM	79.41	68.75	76.00	(70.21 - 81.16)

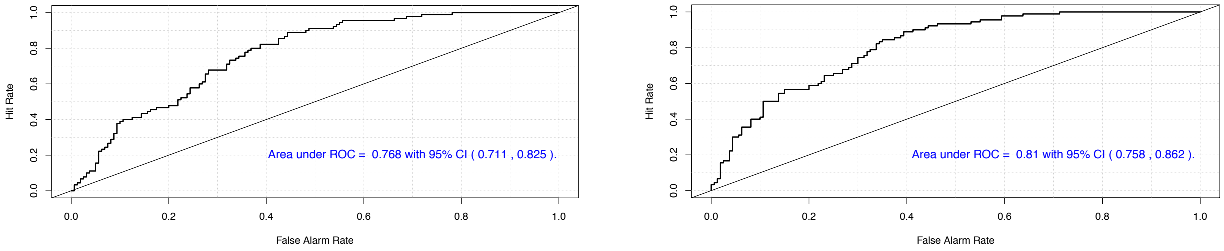
Table 5.5: Model Evaluation using ROC Curve for Heart Disease data.

Models	Sensitivity (%)	Specificity (%)	Accuracy (%)	Conf. Interval (%)
LGR-L_1	77.36	90.00	82.20	(73.57 - 89.83)
RGR	91.11	72.92	81.72	(72.35 - 88.92)
PCR	95.56	81.25	88.17	(79.82 - 93.95)
RF	78.43	88.10	82.80	(73.57 - 89.83)
SVM	78.00	86.05	81.72	(72.35 - 88.98)



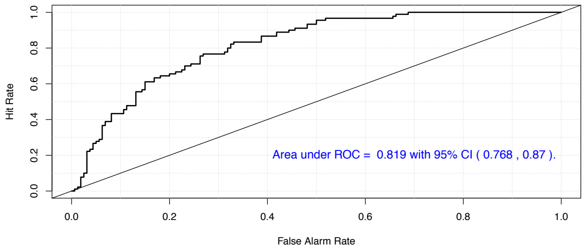
(a) Fitted Logistic Regression

(b) Fitted Ridge Regression



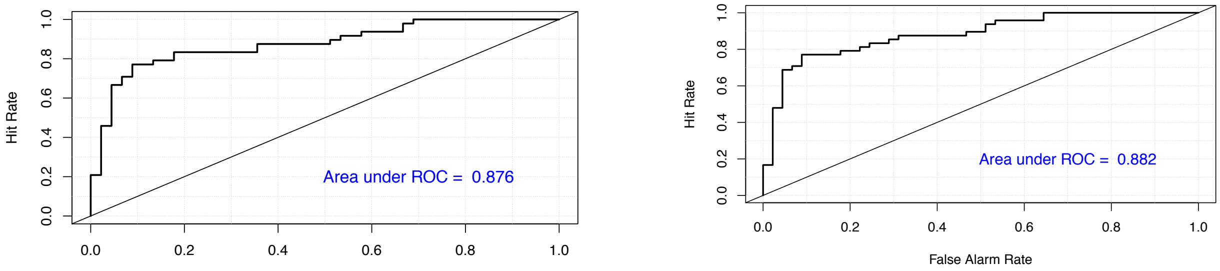
(c) Fitted Principal Comp. Regression

(d) Fitted Random Forest

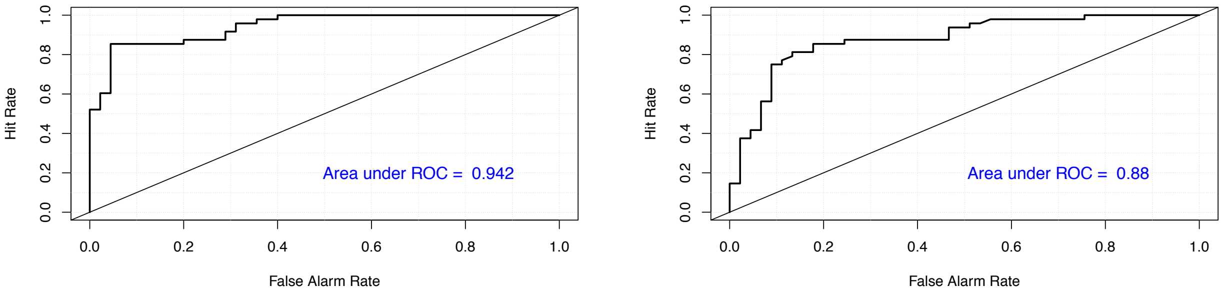


(e) Fitted SVM model

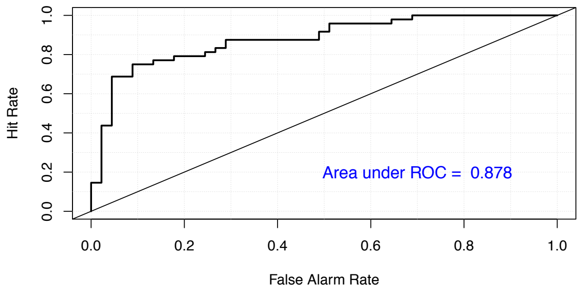
Figure 5.3: Model Evaluation using ROC Curve for Diabetes data.



(a) Fitted Logistic Regression (b) Fitted Ridge Regression



(c) Fitted Principal Comp. Regression (d) Fitted Random Forest



(e) Fitted SVM model

Figure 5.4: Model Evaluation using ROC Curve for Heart Disease data.

Chapter 6

Concluding remarks

The dissertation deals with the predicting stochastic volatility for extreme fluctuations of high frequency data arising in finance and geophysics. As the high frequency data appears very quickly, the fluctuations of these data contain specific reasons for the dynamics of data. In chapter 1, it is shown that the deterministic model does not encapsulate the full evolution of volatility and some volatility models can not capture the financial crashes or extreme fluctuations of seismic events. The deterministic approach often does not allow the level of conservatism present to be quantified, often leading to over-conservatism or inadvertent non-conservatism. In this case, a type of stochastic volatility models that incorporates time-varying parameters are implemented with stationary conditions at different levels. The stationary process is relatively easy to model the high frequency time series and facilitates prediction with higher accuracy and fidelity.

The SV model is used with a kalman filtering technique to estimate the parameters and standard errors. The kalman filtering is advantageous as it filters out unnecessary information (noise) in the high frequency data. The MLE is computed to find the optimal estimation of time-varying parameters in order to forecast the data volatility. The aims of this research focused on analyzing three volatility models: the GARCH, the APARCH, and the SV models in order to predict data volatility.

The parameter estimates of the GARCH (1, 1) and APARCH models indicate that there exists a stationary solution in the conditional volatility of high frequency financial returns (see subsections 4.1.1 and 4.1.2). But for the SV model, we were able to estimate the volatility parameters of time series including extreme fluctuations. It is because the one-step-ahead predictions along with the estimated standard error of stochastic volatility model do not show any limitations unlike the GARCH and APARCH models. We notice that the APARCH model is able to solve the leverage effect of GARCH model. But the GARCH and the APRACH both are more sensitive to noise or unexpected shocks compared to stochastic volatility model. The reason is that the SV model takes into account a stochastic component of the data volatility and estimates the time-varying parameters using filtering techniques. The estimates obtained from SV model are stable around the true value. In Figs. 4.1-4.2, it is clear that the stochastic volatility is able to detect the financial crashes or extreme fluctuations for both financial and geophysical data.

From data background, it is observed that there are some noisy observations like the values of zeroes in high frequency data which do not fit with the stochastic volatility model. A new approach has been developed in order to fit the SV model with noisy data. In this case, a stochastic differential equation has been used to first simulate the high frequency data. We used the superposed Ornstein-Uhlenbeck model with the Lévy driven process and simulated data via the solution of the stochastic differential equation. The high frequency data aligns with this technique because of its stochastic behavior. The low estimated errors (see Table 4.11) implies that the estimation procedure is accurate and capable of generating a higher forecasting accuracy.

6.1 Future works

The summary of this work opens avenues for predicting the other high frequency data like volcanic eruptions data, atmospheric data, other seismic events and financial crashes data using SV models. For the noisy part of data, other types of simulation technique may be applied, for example, systems of stochastic differential equations, machine learning and deep learning algorithms. As we know, the high frequency data follow almost log-normal distribution, for any finite-variance Lévy process, randomizing time is equivalent to randomizing variance. Thus, the time-varying Lévy process generates stochastic volatility (SV) by randomizing time, which may improve the forecasting performance. The methodologies can be used for market analysis, portfolio design or applied in other disciplines such as geophysics, social sciences, medicine and other public health datasets. In particular, the machine learning algorithms can be applied to analyze the novel COVID-19 disease.

References

[1] Fong, S.J., and Nannan, Z. (2011), Towards an Adaptive Forecasting of Earthquake Time Series from Decomposable and Salient Characteristics, *The Third International Conferences on Pervasive Patterns and Applications*, ISBN : 978-1-61208-158-8, 53-60.

[2] Engle, R.F. (1982), Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, **50(4)**, 987-1007.

[3] Bollerslev, T. (1986), Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics*, **31**, 307-327.

[4] Balland P. (2002), Deterministic implied volatility models, *Quantitative Finance*, **2**, 31-44.

[5] Franzini L., Harvey, A.C. (1983), Testing for deterministic trend and seasonal components in time series models, *Biometrika* , **70(3)**, 673-682.

[6] Guarnaccia C., Quartieri, J., Tepedino, C. (2017), Deterministic decomposition and seasonal ARIMA time series models applied to airport noise forecasting, *AIP Conference Proceedings* , **1836**, 020079.

[7] Mariani, M.C., Bhuiyan, M.A.M., Tweneboah, O.K. (2018), Estimation of stochastic volatility by using Ornstein-Uhlenbeck type models, *Physica A- Statistical Mechanics and its Applications*, **491**, 167-176.

[8] Mariani, M.C. Bhuiyan, M.A.M., Tweneboah, O.K., Gonzalez-Huizar, H., Florescu, I. (2018) Volatility models applied to geophysics and high frequency financial market data, *Physica A- Statistical Mechanics and its Applications*, **503**, 304-321.

[9] Mariani, M.C., Bhuiyan, M. A. M, and Tweneboah, O.K., Beccar Varela M.P., Florescu, I. (2018), Analysis of stock market data by using Dynamic Fourier and Wavelets techniques, *Physica A- Statistical Mechanics and its Applications*, **537**, 122785.

[10] Hamiel, Y., Amit, R., Begin, Z.B., Marco, S., Katz, O., Salamon, A., Zilberman, E., and Porat, N. (2009), The seismicity along the Dead Sea fault during the last 60,000 years, *Bulletin of Seismological Society of America*, **99(3)**, 2020-2026.

[11] Brockman, P., and Chowdhury, M. (1997), Deterministic versus stochastic volatility: implications for option pricing models, *Applied Financial Economics*, **7**, 499-505.

[12] Danielsson, J. (2011), Financial Risk Forecasting: The Theory and Practice of Forecasting Market Risk with Implementation in R and Matlab, *The Wiley Finance Series*, DOI:10.1002/9781119205869, 5-29.

[13] Mariani, M.C., Gonzalez-Huizar, H., Bhuiyan, M.A.M. and Tweneboah, O.K. (2017), Using Dynamic Fourier Analysis to Discriminate Between Seismic Signals from Natural Earthquakes and Mining Explosions, *AIMS Geosciences*, **3(3)**, 438-449.

[14] Rubio, F. J., and Johansen, A. M. (2013), A simple approach to maximum intractable likelihood estimation, *Electronic Journal of Statistics*, **7**, 1632-1654.

[15] Mariani, M.C., Asante, P., Bhuiyan, M.A.M., Beccar-Varela, M.P., Sebastian, J. and Tweneboah, O.K. (2020), Long-Range Correlations and Characterization of Financial and Volcanic Time Series, *Mathematics (MDPI)*, **8(3)**, 441.

[16] Evenson, R., Kislev, Y. (2020), A Stochastic Model of Applied Research, *Journal of Political Economy*, **84(2)**, 265-282.

[17] Rey, j.S. (2015), Mathematical Models in Geography - International Encyclopedia of the Social Behavioral Sciences (Second Edition), *Elsevier*, 785-790.

[18] Ruppert, D. (2010), Statistics and Data Analysis for Financial Engineering, *Springer*, 205-206.

[19] Ruppert, D. (2010), Statistics and Data Analysis for Financial Engineering, *Springer*, 207-215.

[20] Granger, C.W. and R. Joyeux (1980), An introduction to long-memory time series models and fractional differencing, *Journal of Time Series Analysis*, **1**, 15-29.

[21] Hosking, J.R.M (1981), Fractional differencing, *Biometrika*, **68**, 165-176.

[22] Mariani, M.C., Bhuiyan, M.A.M. and Tweneboah, O.K., and Gonzalez-Huizar, H. (2019), Long Memory Effects and Forecasting of Earthquake and Volcano Seismic Data, *Physica A- Statistical Mechanics and its Applications*, submitted.

[23] Commandeur, J.F., and Koopman, S.J. (2007), An Introduction to State Space Time Series Analysis, *Oxford University press*, 107-121.

[24] Mariani, M.C., Tweneboah, O.K., Gonzalez-Huizar, H., and Serpa, L.F. (2012), SKalman Filtering for Spacecraft Attitude Estimation, *Aerospace Research Centeral*, **5(5)**, AIAA 82-0070R.

[25] Cipra, T. and Romera, T. (1991), Robust Kalman Filter and Its Application in Time Series Analysis, *Kybernetika*, **27(6)**, 481-494.

[26] $https://www.kalmanfilter.net/kalman1d.html$

[27] Eliason, S. R. (1993), Maximum Likelihood Estimation-Logic and Practice , *Quantitative applications in the social sciences*, **96**, 1-10.

[28] N.K. Gupta and R. K. Mehra (1974), Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations, *IEEE Transactions on Automatic Control*, **19(6)**, 774-783.

[29] Jones, R.H. (1980), Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics*, **22(3)**, 389-395.

[30] Mariani,M.C., Bhuiyan, M. A. Masum, Tweneboah, O., Beccar M., Florescu, I. (2020), Classifying the Lehman Brothers Collapse and Flash Crash events and their effects on other financial data, Quantitative Finance, Submitted.

[31] R. Z. Wiggins, T. Piontek, A. Metrick, The Lehman Brothers Bankruptcy A:Overview, Yale program on financial stability case study, 2014-3A-v1, October 1, 2014.

[32] NANEX May 6th 2010 Flash Crash Analysis, Final Conclusion (2010), http://www.nanex.net/FlashCrashFinal/FlashCrashAnalysis_Theory.html, October 14, 2010.

[33] CFTC and SEC (2010b), Preliminary Findings Regarding the Market Events of May 6, 2010, Staff Report.

[34] Mariani, M.C., Bhuiyan, M.A.M., Tweneboah, O.K., and Gonzalez-Huizar, H. (2017), Forecasting the Volatility of Geophysical Time Series with Stochastic Volatility Models, *Intl. Science Index- Mathematical and Computational sciences*, **11(10)**, 371-377.

[35] Said, S.E. and Dickey, D. (1984), Testing for Unit Roots in Autoregressive Moving-Average Models with Unknown Order, *Biometrika*, **71**, 599-607.

[36] Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., and Shin, Y. (1992), Testing the null hypothesis of stationarity against the alternative of a unit root, *Journal of Econometrics*, **54**, 159-178.

[37] Rajarshi, M.B. (2012), Statistical Inference for Discrete Time Stochastic Processes, *Springer - ISBN: 978-81-322-0763-4*, 39-55.

[38] Brys, G., Hubert, M., and Struyf, A. (2004), A Robustification of the Jarque-bera test of normality, *Physica-Verlag/Springer*, 753-760.

[39] Shapiro, S.S., and Wilk, M.B. (1965), An Analysis of Variance Test for Normality (Complete Samples), *Biometrika*, **52(3/4)**, 591-611.

[40] Ljung, G.M., and Box, G.E.P. (1978), On a Measure of Lack of Fit in Time Series Models, *Biometrika*, **65(2)**, 297-303.

[41] Wang, W., Gelder, P.H.A.J.M.V., Virijling, J.K., and Ma, J. (2005), Testing and modelling autoregressive conditional heteroskedasticity of streamflow processes, *European Geosciences Union*, **12(1)**, 55-66.

[42] Janssen, A., and Drees, H. (2016), A stochastic volatility model with flexible extremal dependence structure, *Bernoulli*, **22(3)**, 1448-1490.

[43] Taylor, S. J. (1982). Financial returns modeled by the product of two stochastic processes, A study of daily sugar prices, 1961-79. *Time Series Analysis: Theory and Practice, ZDB-ID 7214716*, **1**, 203-226.

[44] http://web.mit.edu/kirtley/kirtley/binlustuff/literature/control/Kalman%20filter.pdf

[45] Mariani,M.C., Tweneboah, O., Bhuiyan, M.A. Masum (2019), Supervised Machine Learning Models Applied to Disease Diagnosis and Prognosis, *AIMS Public Health*, **6(4)**, 405-423.

[46] Mariani, M., Bhuiyan, M.A.M, Tweneboah, O. (2019), Statistical data mining algorithms for the prognosis of diabetes and autism, *9th Annual STEM conference*, Hawaii.

Curriculum Vitae

Md Al Masum Bhuiyan was born on January 19. Currently, he is pursuing a Ph.D. degree in Computational Science at The University of Texas at El Paso (UTEP). He is also enrolled in the Big Data Analytics Graduate Certificate program at UTEP. He received a Master degree in Mathematical Sciences at UTEP in 2015.

He earned another Master degree in Applied Mathematics and, prior to that, a Bachelor degree in Mathematics, both from University of Dhaka, one of the top ranked universities for scientific research in Bangladesh. This fuelled his interest in pursuing the doctoral degree in Computational Science Program (CPS), where he excels at theoretical, statistical, and numerical problem-solving abilities, as well as programming skills.

His research focuses on high frequency and high dimensional data analysis and other data analytics applications using statistical methods, stochastic differential equation, stochastic volatility and machine learning algorithms. While pursuing the Ph.D. in CPS, he worked as a Teaching Assistant, Research Assistant, MATH instructor, and Math Adjunct faculty of El Paso Community College as internship programs.

He has several publications, as well as several paper presentations at national and international conferences. He received several scholarships from UTEP, several travel funds from different Universities and a scholar award in the category of outstanding performance of research presentation. Recently the UTEP, the El Paso Herald, the News-wise magazine published articles in their websites on his excellent research achievements.

He served as a reviewer of several scientific journals, judge of several science-fairs and symposiums and as a treasurer of Computational Science Student Association in 2018-2019. He is a member of SIAM, AMS and ASA associations.

Residential address: 300, W Nevada Ave, Apt-01
El Paso, Texas 79902