STATISTICAL ANALYSIS OF THREE FOURTH-ORDER ORDINARY DIFFERENTIAL EQUATION SOLVERS

by

BO HE, M.S.E.E., B.S.E.E.

A THESIS

IN

STATISTICS

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

MASTER OF SCIENCE

Approved

Clyde F. Martin
Chairperson of the Committee

James G. Surles

Mara Neusel

Accepted

Fred Hartmeister
Dean of the Graduate School

December, 2007

ACKNOWLEDGMENTS

I would like to gratefully acknowledge my adviser, Dr. Clyde F. Martin for his kindness, patience and generous help on my work with his broad knowledge and experiences. I also express my honest appreciation to Dr. James G. Surles and Dr. Mara Neusel for serving on my committee. Dr. Surles gives enlightened suggestions to the work in this thesis. In addition, I would like to thank the faculty, staff and graduate students of the Department of Mathematics and Statistics at Texas Tech University for their encouragement, friendship and support. Finally, I would like to express my gratitude to my parents and my husband, who are far away. Their constant support and love always encourage me to move on in my life.

CONTENTS

ABSTRACT

We develop an autoregressive integrated moving average model ($ARIMA$) to study the statistical behavior of the numerical error generated from three fourth-order ODE solvers: Milne's method, Adams-Bashforth method and a new method which randomly switches between Milne and Adams-Bashforth methods. With the actual error data based on three differential equations, we desire to identify an $ARIMA$ model to each data series. Results show that some of data series can be described by $ARIMA$ models but others can not. Based on the mathematical form of the numerical error, other statistical models should be investigated in the future. Finally we assess the multivariate normality of the sample mean vectors which are generated by the switching method as an application of the multivariate central limit theorem.

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

Ordinary differential equations (ODEs) arise in many instances when using mathematical modeling techniques for describing phenomena in science, engineering, economics, etc. In most cases the model is too complex to be solved analytically, numerical methods must be used. The numerical solution is only an approximation of the exact solution, so an amount of error is introduced in the answer. In our study we define the numerical error as the difference between the numerical solution and the analytical solution of a differential equation. So far many research topics related to error analysis focus on minimizing the numerical error. But the statistical properties of the numerical error are less studied. Our goal is to analyze the numerical error statistically and try to develop a time series model to fit the numerical error.

Our work focuses on three fourth-order numerical methods: Milne's method, Adams-Bashforth method and the switching method which randomly switches between these first two methods. We apply these three fourth-order methods to solve one linear ordinary differential equation and two nonlinear differential equations. All of these differential equations have well known analytical solutions. So the difference between the numerical solution and the analytical solution is the error data that we investigate in the thesis.

An autoregressive integrated moving average model ($ARIMA$) is applied to describe the numerical error. Since the error data obtained from each differential equation has a visible trend (upward or downward), we first use a cubic/quartic regression model to remove the trend. The transformed data are the residuals from the regression model, and then we try to build an $ARIMA$ model to describe the transformed data instead of the original error data. There are three stages in $ARIMA$ model building: identification, estimation and diagnostic checking. We follow these three stages and try to specify a fitted model to each set of error data. Finally we found that the numerical error generated by the switching method from the linear differential

equation can be described by $ARIMA$ models but the error data from Milne's and Adams-Bashforth methods can not. And the error data from two nonlinear differential equations generated by all three methods fail to be fitted by $ARIMA$ models. A mathematical interpretation is given in the thesis and we suggest that the error behavior caused by the step size h should be considered in future model building.

In the switching method, we can treat the error data $\mathbf{E} = (e_4, e_5, \ldots, e_{99})'$ as a random vector because each element in this vector is a random variable. If we choose a large sample of observations of $\mathbf{E}$ with the sample size n, then the sample mean vector $\mathbf{M} = \frac{1}{n}\sum_{i=1}^{n} E_i$ should have an approximate multivariate normal distribution by the Multivariate Central Limit Theorem. We assess the multivariate normality of $\mathbf{M}$ by applying *Mardia's skewness and kurtosis* test. It turns out that all the sample mean vectors $\mathbf{M}$ based on three differential equations have a limiting multivariate normal distribution. So we conclude that the numerical error can be normalized by random switching between numerical methods.

CHAPTER II
DATA PREPARATION

The numerical error is the data set that we investigate in this thesis. In this chapter, we generate all sets of the numerical error based on three ordinary differential equations. All of these equations can be solved analytically. Also we apply three fourth-order numerical methods to obtain their numerical solutions and generate the error from each method.

2.1 One Linear Differential Equation

We start with a simple linear ordinary differential equation:

$$\frac{dy}{dt} = \lambda y, \quad 0 \le t \le 1, \quad y(0) = 1.$$

Its analytical solution is given by:

$$y(t) = e^{\lambda t}.$$

Numerical methods are referred to as discrete variable methods and generate a sequence of approximate values $\{y_n\}$ at points $\{t_n\}$, where $n = 0, 1, \ldots, N-1$. These y_n values are obtained in a step-by-step manner, that is, we use one or more previous points to compute the successive point [1]. Our work focuses on fourth-order numerical methods, so information from four previous points y_n, y_{n-1}, y_{n-2} and y_{n-3} are necessary to determine the value of y_{n+1}. y_n is an approximation to $y(t)$ at point t_n, that is, $y_n \approx y(t_n)$, where $t_{n+1} = t_n + h$ for $n = 0, 1, \ldots, N-1$. The parameter h is the step size given by $h = \dfrac{t_N - t_0}{N}$. In this example, we have $t_0 = 0$, $t_N = 1$. Usually the smaller the step size, the more accurate the approximation. In our study, we let $N = 100$ and $h = 0.01$.

First we apply two well known fourth-order methods: Milne's method and Adams-Bashforth method [2] to solve the linear differential equation. The Milne's fourth-

order algorithm is given as follows:

$$y_{n+1} = y_{n-3} + \frac{4h}{3}[2f(t_n, y_n) - f(t_{n-1}, y_{n-1}) + 2f(t_{n-2}, y_{n-2})]. \tag{2.1}$$

And the fourth-order Adams-Bashforth algorithm is:

$$y_{n+1} = y_n + \frac{h}{24}[55f(t_n, y_n) - 59f(t_{n-1}, y_{n-1}) + 37f(t_{n-2}, y_{n-2}) - 9f(t_{n-3}, y_{n-3})], \tag{2.2}$$

where $f(t, y) = \dfrac{dy}{dt}$. For this example $\dfrac{dy}{dt} = \lambda y$, so algorithm (2.1) and (2.2) can also be written as follows: Milne's fourth-order algorithm:

$$y_{n+1} = y_{n-3} + \frac{4\lambda h}{3}(2y_n - y_{n-1} + 2y_{n-2}). \tag{2.3}$$

Fourth-order Adams-Bashforth algoritm:

$$y_{n+1} = y_n + \frac{\lambda h}{24}(55y_n - 59y_{n-1} + 37y_{n-2} - 9y_{n-3}). \tag{2.4}$$

In these two fourth-order methods, we need four initial values to calculate the successive point value. In our work, we set analytical solutions when $t = 0, t = h, t = 2h, t = 3h$ as our initial values [3]. So $y_0 = e^{0\lambda h} = 1, \quad y_1 = e^{\lambda h}, \quad y_2 = e^{2\lambda h}, \quad y_3 = e^{3\lambda h}$.

We obtain the numerical error from the following formula:

$$e_n = y_n - e^{\lambda n h}, \quad \text{for} \quad n = 0, 1, 2, \ldots, 99. \tag{2.5}$$

Now we collect the error data generated by the third fourth-order numerical method. The third method is performed by randomly switching between Milne's and Adams-Bashforth methods. We are interested in whether this switching method could achieve better results. First we also let analytical values $y_0 = 1, y_1 = e^{\lambda h}, y_2 = e^{2\lambda h}, y_3 = e^{3\lambda h}$ be four initial values, then we choose Milne's algorithm (2.3) or Adams-Bashforth algorithm (2.4) randomly to determine the successive numerical value y_4.

Similarly y_5 is calculated by using algorithm (2.3) or (2.4) randomly and with information of four preceding values y_1, y_2, y_3, y_4. Other point values $y_6, y_7, \ldots, y_{99}$ can be obtained in the same way. There are a total of 2^{96} possible sets of numerical solutions $\{y_0, y_1, \ldots, y_{99}\}$ by using the switching method. We run the switching method 1000 times randomly and then produce the error data from each set of numerical solution ${y_n}^i$, where n indicates the numerical value taken at nth step, and i indicates the ith set of solution. So each set of the error data has the form ${e_n}^i = {y_n}^i - e^{\lambda nh}$. In our work, we let the average error $e_n = \dfrac{1}{1000}\sum_{i=1}^{1000} {e_n}^i$, for $n = 0, 1, \ldots, 99$ be the numerical error generated by the switching method.

Figures 2.1 and 2.2 show scatter plots for the error data from the linear equation $\dfrac{dy}{dt} = \lambda y$ with $\lambda = 1$ and $\lambda = -1$ based on Milne's Method, Adams-Bashforth method and the switching method. We can see that the switching method is more accurate than Adams-Bashforth method but less than Milne's method.

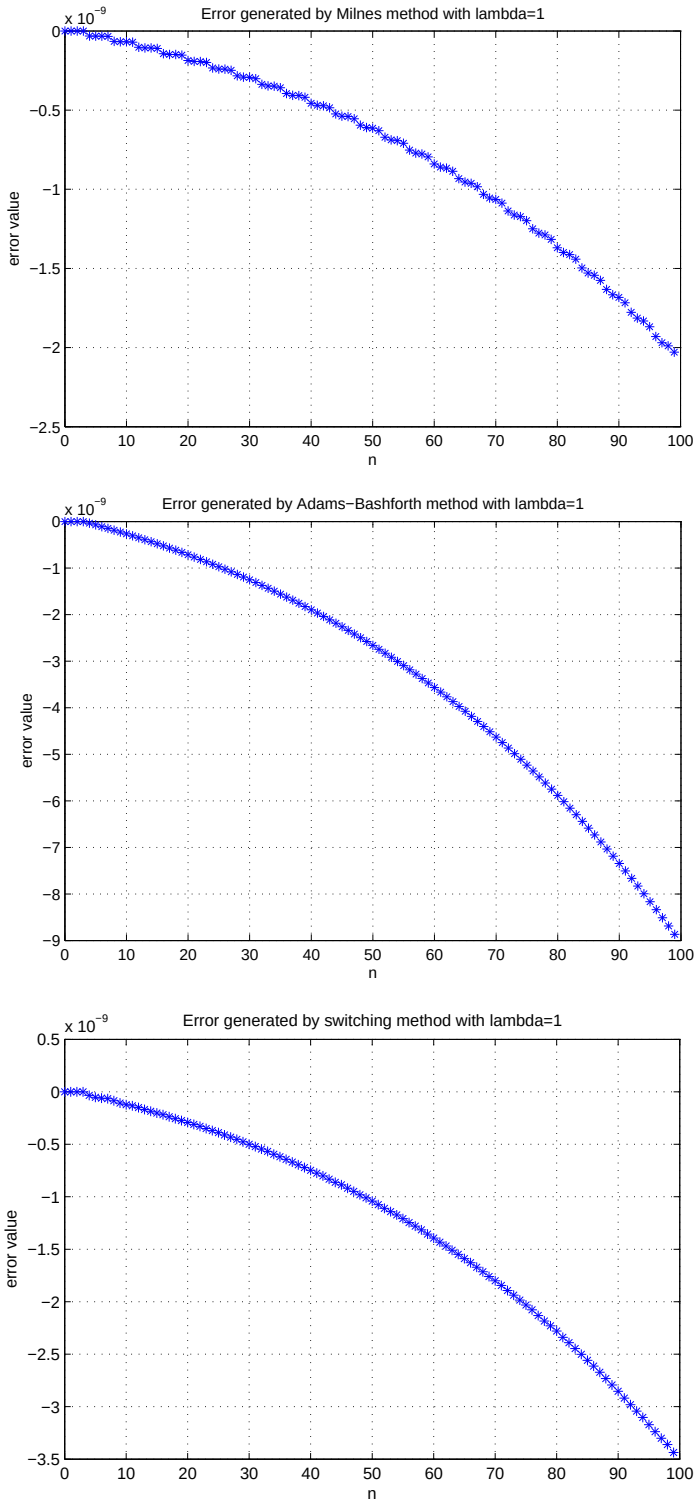

Figure 2.1: Error data from linear equation $\frac{dy}{dt} = y$ based on three fourth-order numerical methods.

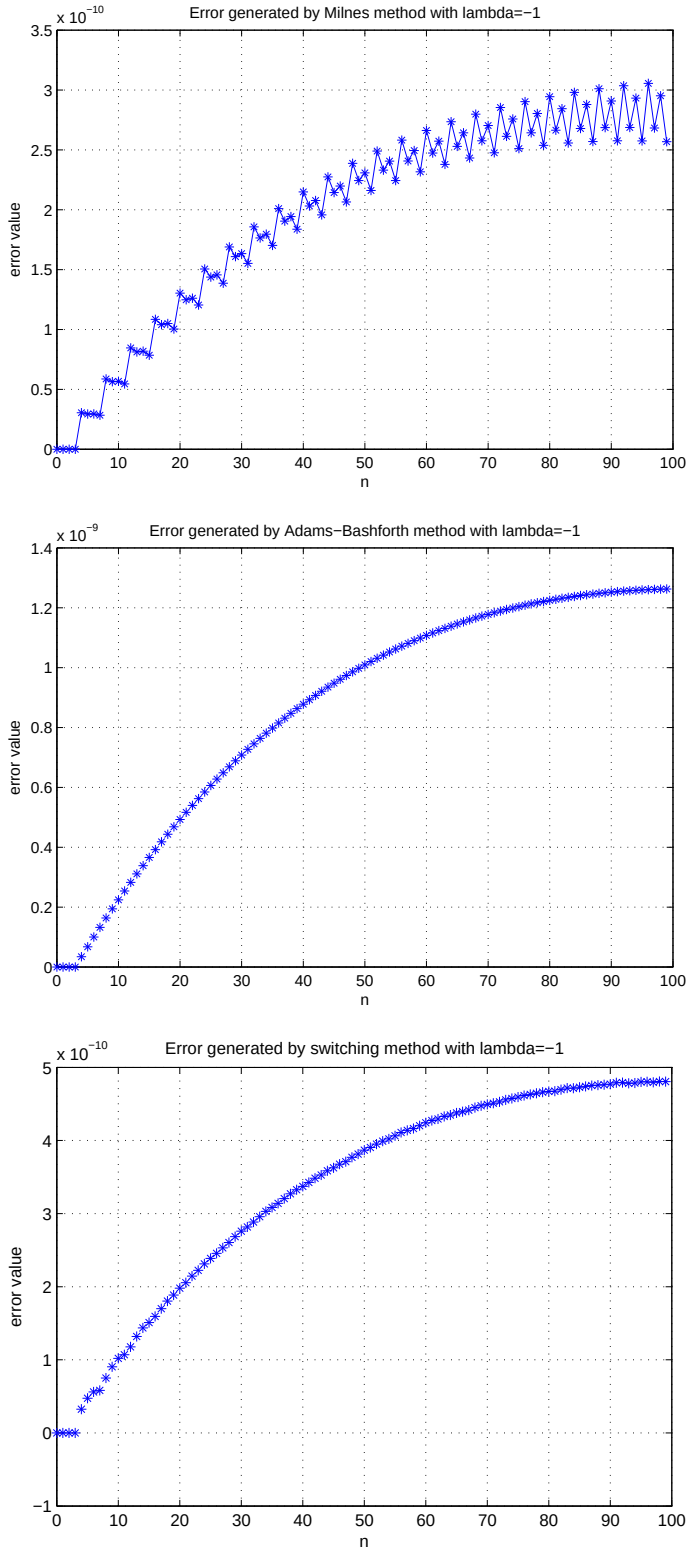

Figure 2.2: Error data from linear equation $\frac{dy}{dt} = -y$ based on three fourth-order numerical methods.

2.2 Two Nonlinear Differential Equations

We consider the numerical error generated by two nonlinear ordinary differential equations now. The first equation is

$$\frac{dy}{dt} = 1 + y^2, \quad 0 \le t \le 1, \quad y(0) = 0, \tag{2.6}$$

and its analytical solution is:

$$y(t) = \tan t.$$

The second equation is

$$\frac{dy}{dt} = 1 - y^2, \quad 0 \le t \le 1, \quad y(0) = 0, \tag{2.7}$$

and its analytical solution is:

$$y(t) = \frac{e^{2t} - 1}{e^{2t} + 1}.$$

We apply the Milne's method, Adams-Bashforth method and the switching method to solve these two equations numerically and produce the error data respectively. When we apply Milne's algorithm (2.1) and Adams-Bashforth's algorithm (2.2) to obtain numerical solutions, we use $f(t, y) = \dfrac{dy}{dt} = 1 + y^2$ for equation (2.6) and $f(t, y) = \dfrac{dy}{dt} = 1 - y^2$ for equation (2.7). The computation procedure are very similar as we did for the linear equation. Without presenting the detailed work, we just show scatter plots for the error data from two nonlinear equations as follows.

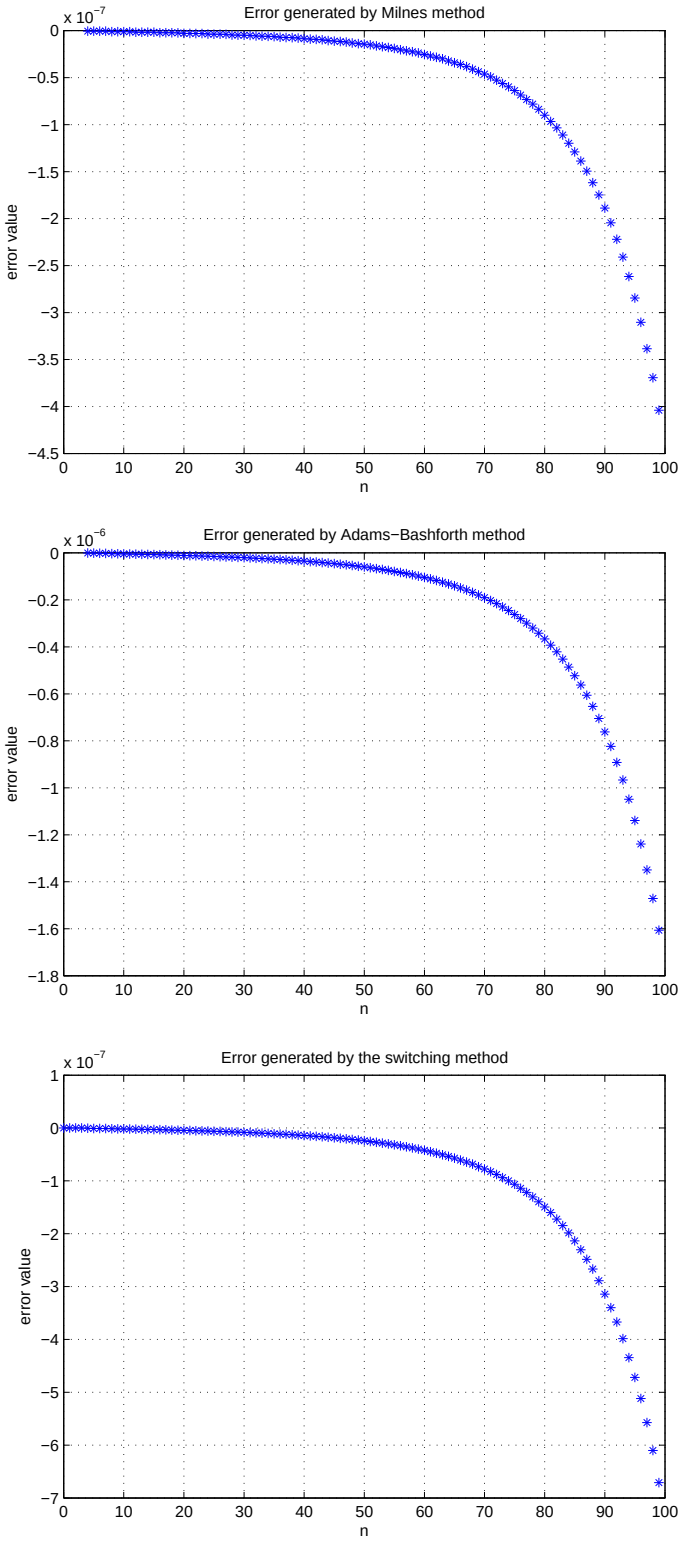

Figure 2.3: Error data from equation $\frac{dy}{dt} = 1 + y^2$ based on three fourth-order numerical methods.

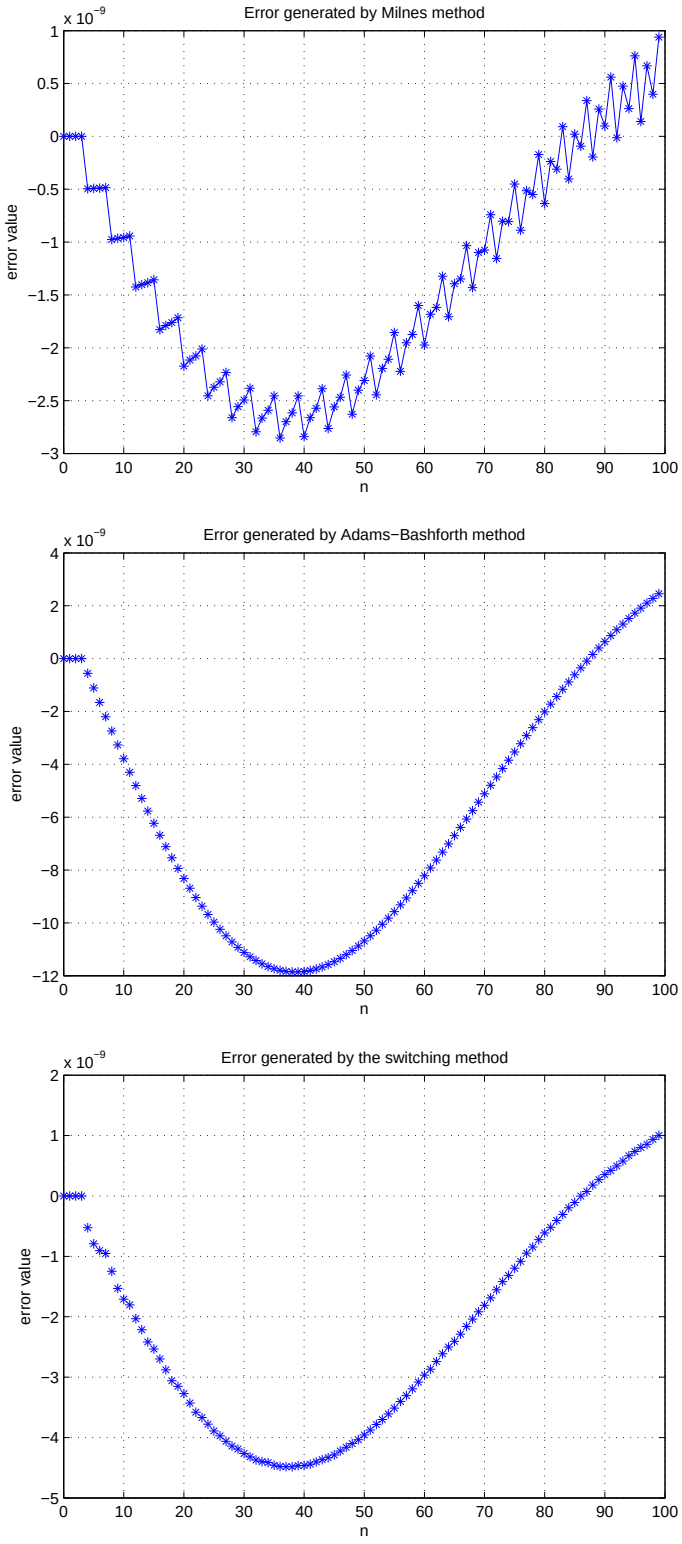

Figure 2.4: Error data from equation $\frac{dy}{dt} = 1 - y^2$ based on three fourth-order numerical methods.

CHAPTER III
TIME SERIES MODEL

A time series is a sequence of observations taken at equally spaced time intervals and adjacent observations are dependent. The main goals of time series analysis are basically two: to identify the regular patterns present in the data and to predict future values [4]. In our study, the numerical error is a typical time series because the numerical error is a sequence of dependent observations and they are taken at uniform consecutive time intervals. In this chapter, we focus on identifying time series models to all sets of numerical error discussed in Chapter II.

3.1 General Time Series Models

3.1.1 Main Characteristics of a Time Series

There are some main characteristics of a time series that we need to consider when building a time series model. A stochastic process $\{x_t\}$ is a collection of random variables, where t vary over the integers $t = 0, \pm 1, \pm 2, \ldots$. A time series is a sample realization of a stochastic process which is observed only for a finite number of intervals, indexed by $t = 1, \ldots, n$. Any stochastic process can be partially characterized by the first and second moments of the joint probability distribution [5]. The mean value function for the series at one particular time point is defined as:

$$\mu_t = E(x_t). \tag{3.1}$$

the variance at time t is:

$$\gamma(t,t) = E[(x_t - \mu_t)^2]. \tag{3.2}$$

the autocovariance function is given by:

$$\gamma(s,t) = E[(x_s - \mu_s)(x_t - \mu_t)], \tag{3.3}$$

which measures the linear dependence between two points on the same series observed at different times. Then the autocorrelation function (**ACF**) is denfined as:

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}. \tag{3.4}$$

Note that $-1 \leqslant \rho(s,t) \leqslant 1$ for all s and t.

An important class of stochastic processes is the stationary process. We define a weakly stationary time series as a process whose mean and variance are constant over time and the autocovariance function depends only on the time shift or lag [5]. That is:

$$E(x_t) = \mu, \forall t, \tag{3.5}$$

and

$$\gamma(s,t) = E[(x_s - \mu)(x_t - \mu)] = \gamma(s-t) \tag{3.6}$$

so the autocovariance function $\gamma(s,t)$ depends on s and t only through their difference. In other words, letting $s = t + h$, where h represents the lag, then

$$\gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)] \tag{3.7}$$

Also the variance at any time t is:

$$\gamma(t,t) = \gamma(0) = E[(x_t - \mu)^2] \tag{3.8}$$

Finally the ACF for a stationary process at lag h is:

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}. \tag{3.9}$$

Another useful tool to describe a stationary process is the partial autocorrelation function (**PACF**). The partial autocorrelation function ϕ_{hh} at lag h measures the correlation between x_t and x_{t-h} after the linear effect of the intermediate values

$\{x_{t-1}, \ldots, x_{t-(h-1)}\}$ has been removed. In [6], the formula of PACF is well presented.

For a specific stationary time series, we can estimate the mean μ (3.5) of the stochastic process by the sample mean

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{3.10}$$

and the theoretical autocovariance (3.7) by the sample autocovariance at lag h:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}). \tag{3.11}$$

So the sample autocorrelation at lag h is given as follows:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \tag{3.12}$$

3.1.2 Time Series Models

Autoregressive models (**AR**), Moving average models (**MA**) and autoregressive-moving average models (**ARMA**) are linear stationary models. Both nonstationary and stationary time series can be described by a more powerful model, autoregressive integrated moving average models (**ARIMA**).

Autoregressive models. In this model, the current value of the series is expressed as a finite, linear aggregate of previous values of the series and a white noise series [6]. The form of autoregressive model of order p, abbreviated as **AR**(**p**), is given as follows:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \omega_t, \tag{3.13}$$

where $\phi_1, \phi_2, \ldots, \phi_p$ are parameters and ω_t is a white noise series with mean zero and constant variance ${\sigma_\omega}^2$. In this form we assume that the mean of x_t is zero. If the mean μ of x_t is not zero, we should substitute $x_t - \mu$ for x_t in (3.13), or write

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \omega_t, \tag{3.14}$$

where $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$. A second useful form of $AR(p)$ models follows by using the *backshift operator* B, which is defined by $Bx_t = x_{t-1}$; hence $B^m x_t = x_{t-m}$. So the $AR(p)$ model given in (3.13) can also be written as:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)x_t = \omega_t \tag{3.15}$$

or more concisely as

$$\phi(B)x_t = \omega_t, \tag{3.16}$$

where the autoregressive operator $\phi(B)$ is

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{3.17}$$

Moving Average models. Unlike the $AR(p)$ model, the moving average model of order q, abbreviated as **MA(q)**, assumes the white noise ω_t are combined linearly to determine the current observed data. Thus

$$x_t = \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_q \omega_{t-q}, \tag{3.18}$$

where $\theta_1, \theta_2, \ldots, \theta_q$ are parameters.

If we define a moving average operator of order q by

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q, \tag{3.19}$$

the moving average model may be written as

$$x_t = \theta(B)\omega_t. \tag{3.20}$$

Autoregressive Moving Average models. To achieve greater flexibility in fitting of actual time series, it is sometime advantageous to include both autoregressive and moving average terms in the same model [7]. This leads to the mixed

autoregressive-moving average model, abbreviated as **ARMA(p,q)**:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \omega_t + \theta_1 \omega_{t-1} + \cdots + \theta_q \omega_{t-q} \tag{3.21}$$

or

$$\phi(B) x_t = \theta(B) \omega_t \tag{3.22}$$

where p and q are the autoregressive and the moving average orders, respectively. As before, if the mean μ of x_t is not zero, we let $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$ and rewrite the model (3.21) as

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \omega_t + \theta_1 \omega_{t-1} + \cdots + \theta_q \omega_{t-q} \tag{3.23}$$

An $ARMA(p, 0)$ model is the same as $AR(p)$ model; likewise an $ARMA(0, q)$ model is the same as $MA(q)$ model.

The models presented so far are based on the stationarity assumption. But many series actually exhibit nonstationary behavior and, in particular, do not vary about a fixed mean. The autoregressive integrated moving average model (**ARIMA**) can be applied to describe many such series.

Autoregressive Integrated Moving Average models. The autoregressive integrated moving average model is a broadening of the class of $ARMA$ models with differencing process. We define the *backward difference operator* $\bigtriangledown$ as $\bigtriangledown x_t = x_t - x_{t-1} = (1 - B)x_t$; hence $\bigtriangledown^d x_t = (1 - B)^d x_t$. Usually the differencing process can transform a nonstationary series to a stationary one [8]. So a series x_t is said to be **ARIMA(p,d,q)** if the differenced series $\bigtriangledown^d x_t = (1 - B)^d x_t$ is a stationary $ARMA(p, q)$. In general, we write the model as

$$\phi(B)(1 - B)^d x_t = \theta(B)\omega_t, \tag{3.24}$$

where the order d is the number of differences needed to achieve stationarity. The

order p and q are the autoregressive and the moving average orders, respectively. Also if the mean of the differenced series $\bigtriangledown^d x_t$ is not zero, then we let $\alpha = \mu(1-\phi_1-\cdots-\phi_p)$ and rewrite the model (3.24) as $\phi(B)(1-B)^d x_t = \alpha + \theta(B)\omega_t$.

As we can see, the $ARIMA(p,d,q)$ model is a general form of the $ARMA(p,q)$, $AR(p)$ and $MA(q)$ models with appropriate selections of p, d and q. Also when it comes to identifying a nonstationary time series, the $ARIMA$ model may be a suitable choice.

3.2 ARIMA Model Building

The upward or downward trend behaviors in error data plots obtained in chapter II suggest modeling the data using $ARIMA$ models. Our task is to identify an appropriate subclass of $ARIMA$ model to represent our error data. The techniques which are most commonly used for model identification were proposed by Box and Jenkins [6]. The procedure involves making successive approximations through three stages: *identification*, *estimation* and *diagnostic checking*. In this section we explain this procedure in detail.

Identification Stage. Our objective in this stage is to transform the nonstationary time series to a stationary one and to specify the autoregressive order p, the moving average order q and the order of the differencing d. We use the sample autocorrelation function ACF and the partial autocorrelation function PACF as the basis for judgment of the stationarity of the series and to provide criteria for specifying values of p, q and d.

The first step, which is taken before starting the identification cycle, is to examine the time plot of the data and to judge whether or not it is stationary. If a trend is evident in the data, then it must be removed. From figures 2.1 to 2.4, we can see all sets of our data series have visible trends, so it is necessary to remove those trends. To try and achieve an approximate constant mean and variance in our data sets posed a challenge. We try many transformations to stabilize the variance such as: the Box-Cox class of power transformation [5], the differencing transformation and fitting of

parametric curves. After examinations of all these transformations, it is concluded that fitting a cubic regression curve through the error data best removes trends. That is,

$$e_n = \beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3 + \eta_n, \tag{3.25}$$

where e_n is the numerical error, $\beta_0, \beta_1, \beta_2$ and β_3 are parameters that can be estimated and η_n is a sequence of residuals [9], which is an approximate stationary time series. In the following procedure, it is the residuals η_n that $ARIMA$ models are applied to fit. We call the residuals η_n as the transformed series in the thesis to avoid confusion with other residuals.

Usually to determine whether stationarity has been achieved, either by trend removal or by differencing, one may examine the ACF sequence of the processed series. The sequence corresponding to a stationary process should converge quite rapidly to zero as the value of the lag increases. Therefore a slow decay in ACF sequence indicates a nonstationary series [10].

After appropriately transforming the data, the next step is to specify the value of d, p and q. All these order values are selected by inspecting the sample ACF and PACF by the following rules.

The Order of Differencing d: Sometimes the differencing process is still needed even the transformation process have been performed. If there is a slow decay in the ACF of the transformed series η_n, then it probably needs differencing. After differencing η_n once, inspect the time plot and ACF of $\bigtriangledown\eta_n$. If additional difference is necessary, then try differencing again. If the ACF at lag 1 is zero or negative or the ACF are all small and patternless, then the series does not need a higher order of differencing. The value of d is the number of differences needed to achieve stationarity.

The Autoregressive Order p and Moving Average Order q: Having tentatively decided what d should be, we next study the ACF and PACF of the appropriately differenced series $\bigtriangledown^d\eta_n$ to provide clues about the choice of order p and q. Table 3.1 can be used as a guide to choose the value of p and q [6].

Table 3.1: Behavior of the ACF and PACF for ARMA models

	AR(p)	**MA(q)**	**ARMA(p, q)**
ACF	Tails off (damped exponentials and/or sine waves)	Cuts off after lag q	Tails off (damped exponentials and/or sine waves)
PACF	Cuts off after lag p	Tails off (damped exponentials and/or sine waves)	Tails off (damped exponentials and/or sine waves)

To determine whether the ACF and PACF show evidence of cut-off behavior, we need some scales to assess the significance of the ACF and PACF values at each lag. Usually a simple measure of the scale of ACF and PACF is provided by the limits of $\pm 1.96/\sqrt{n}$, where n is the length of the data series. These bounds are presented by the dashed horizontal lines on the ACF and PACF graphs.

In this stage two or more related models of the same series may be obtained because of the difficulties of selecting an appropriate model. Then we investigate them further at the estimation and diagnostic checking stages.

Estimation Stage. Our goal in this stage is to estimate the parameters $\phi_1, \ldots, \phi_p$; $\theta_1, \ldots, \theta_q$ and ${\sigma_\omega}^2$ of the selected $ARIMA$ model $\phi(B)(1-B)^d \eta_n = \theta(B)\omega_n$. We employ the least squares method to estimate parameters. In [6], the detail estimation procedure is well presented. Also it can be accomplished by *R-programming*.

Diagnostic Checking Stage. Once we have identified and estimated the candidate $ARIMA$ models, we want to assess whether the models fit the data well enough. This investigation includes the analysis of the residuals and model comparisons.

Analysis of the residuals. If the fitted model is adequate, then the residuals should be approximately white noise ω_n, where ω_n are *iid* normal random variables with mean zero and finite variance ${\sigma_\omega}^2$.

The time plot of the residuals can be inspected to check the mean and constant variance. To test the independence assumption, we could inspect the sample ACF of the residuals. The theoretical ACF of white noise process take value zero for all

nonzero lags. So if the model is appropriate, most values of the sample ACF should be close to zero. We are also interested in whether a set of the ACF of the residuals are jointly zero or not, that is, in testing $H_0 : \rho_{1res} = \rho_{2res} = \cdots = \rho_{Mres} = 0$. The most usual test statistic is the Ljung-Box statistic Q_{LB} [5]. For a good fit model we expect that the Ljung-Box statistic Q_{LB} takes small values or the large corresponding p values. Finally we can use Q-Q plots to assess the normality of the residuals.

Model Comparisons. Once several models have been identified and estimated, it is possible that more than one of them is not rejected in the diagnostic of the residuals. Now we need to decide which model fits better. Akaike Information Criterion, AIC is a goodness of fit measure used to assess which of two ARIMA models are better when both have acceptable residuals [11]. The lower the AIC, the better the model.

3.3 Analysis of ARIMA Models

We will follow three stages of $ARIMA$ model building to specify an ARIMA model for each set of the error series obtained from Chapter II.

3.3.1 Analyzing Error Data from the Equation $\mathbf{\frac{dy}{dt} = y}$

The Switching Method: Figure 3.1 shows the time plot, ACF and PACF of the original error data e_n, the transformed data η_n and the differenced series $\bigtriangledown\eta_n$. Since there is an obvious downward trend in the original error data e_n, we fit a cubic regression model $e_n = \beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3 + \eta_n$ to remove the trend. The quantities $\eta_1, \eta_2, , \ldots, \eta_n$ are the residuals of the cubic regression model and also the transformed data. We try to identify an $ARIMA$ model to present η_n. The PACF of η_n almost cuts off at lag 4 and the ACF tails off slowly, which indicate that an $ARIMA(4, 0, 0)$ model could be appropriate for the η_n series. Also the ACF of η_n decays slowly, so the differencing process may be needed. The right column of figure 3.1 displays the time plot of the differenced series $\bigtriangledown\eta_n$ and its estimated ACF and PACF. The graph of $\bigtriangledown\eta_n$ shows a series that moves around a constant mean with approximately constant variance. The ACF decreases quickly and the PACF cuts off at lag 3. So an

$ARIMA(3,1,0)$ model should be tried as well.

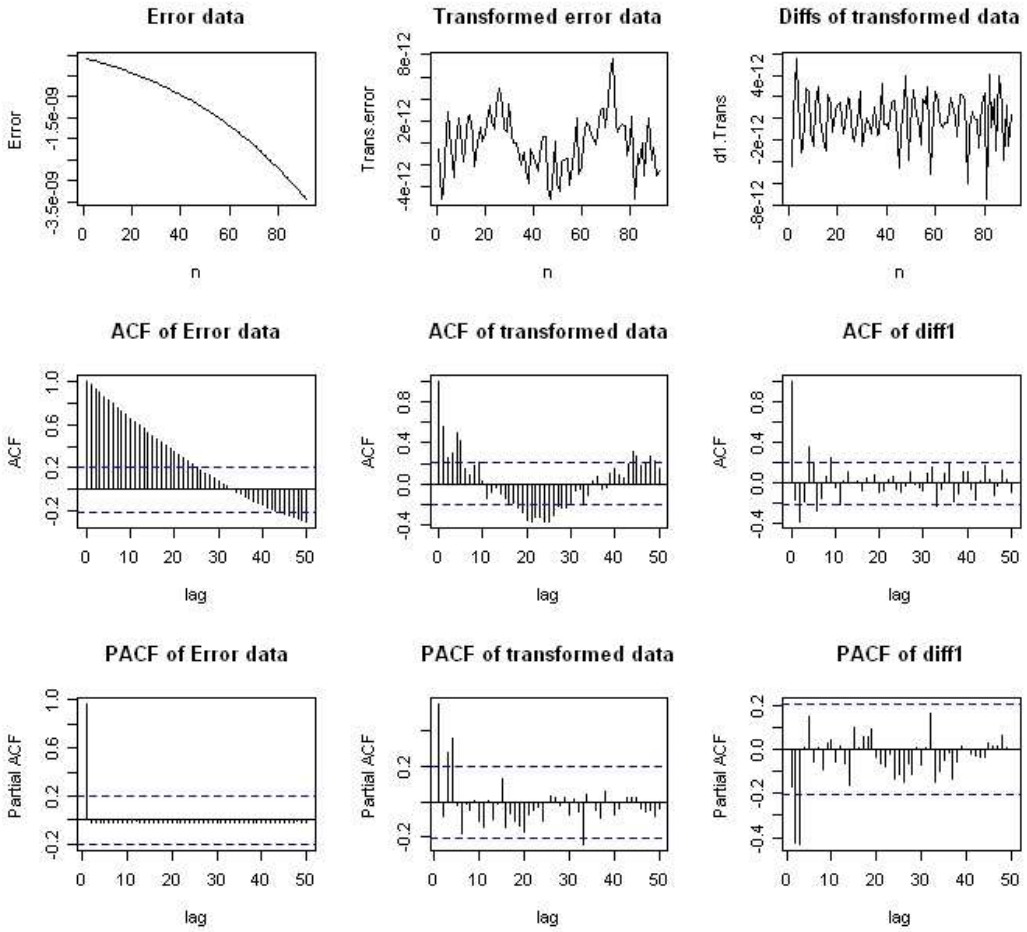

Figure 3.1: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ based on the switching method

After estimating parameters of these two models, we assess their adequacy by analyzing their residuals. Figure 3.2 shows the standardized residuals, their ACF graph and p values for Ljung-Box statistic. Figure 3.3 presents Q-Q plots of the residuals from two models. As we can see the residuals of these two $ARIMA$ models could be treated as white noise series. However the **AIC** value for $ARIMA(4,0,0)$ is $AIC = -4700.89$ and the **AIC** value for $ARIMA(3,1,0)$ is $AIC = -4649.81$. So we select the $ARIMA(4,0,0)$ model.

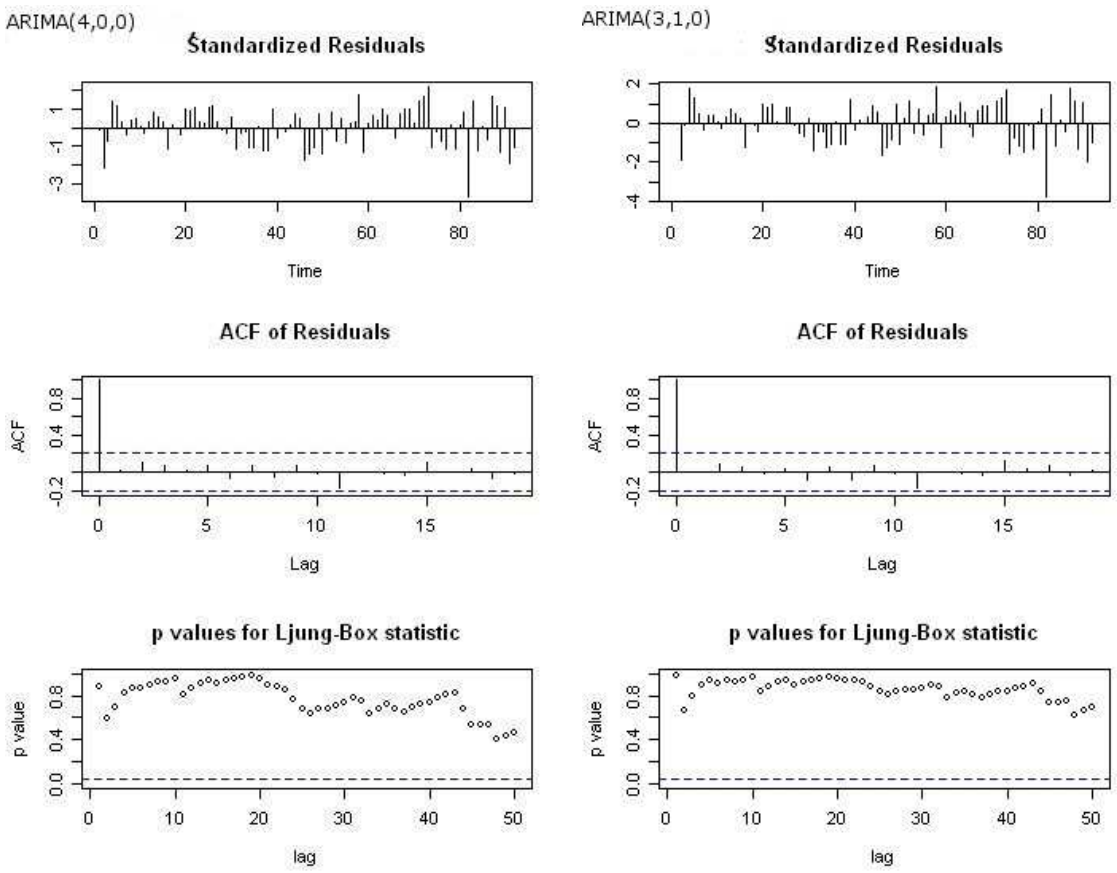

Figure 3.2: The residuals analysis of ARIMA(4,0,0) and ARIMA(3,1,0) models

Finally the statistical model for the error data is

$$e_n = \beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3 + \phi_1 \eta_{n-1} + \phi_2 \eta_{n-2} + \phi_3 \eta_{n-3} + \phi_4 \eta_{n-4} + \omega_n, \quad (3.26)$$

where

$$\eta_{n-1} = e_{n-1} - (\beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3), \quad (3.27)$$

$$\eta_{n-2} = e_{n-2} - (\beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3), \quad (3.28)$$

$$\eta_{n-3} = e_{n-3} - (\beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3), \quad (3.29)$$

$$\eta_{n-4} = e_{n-4} - (\beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3), \quad (3.30)$$

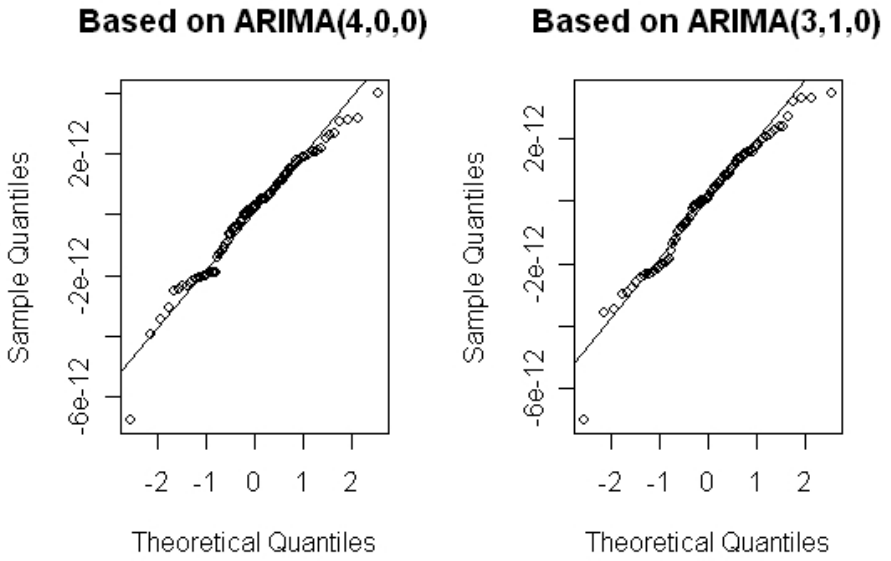

Figure 3.3: The Q-Q plot of the residuls from ARIMA(4,0,0) and ARIMA(3,1,0) models

By substituting equations (3.27), (3.28), (3.29) and (3.30) to the equation (3.26), we obtain another expression for the error data e_n in which all the parameters can be estimated.

$$e_n = \phi_1 e_{n-1} + \phi_2 e_{n-2} + \phi_3 e_{n-3} + \phi_4 e_{n-4} + C(1 - \phi_1 - \phi_2 - \phi_3 - \phi_4) + \omega_n, \quad (3.31)$$

where $C = \beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3$.

Equation (3.31) tells us that the error value at each step depends on four previous error values. This result seems like a logical model because the error series is generated by fourth-order methods.

The Milne's Method: The time plot, ACF and PACF of the original error data e_n, the transformed series η_n and the differenced series $\bigtriangledown^4 \eta_n$ are presented in figure 3.4. We also apply a cubic regression model to remove the trend. And the ACF of η_n strongly suggests that the differencing is necessary. Actually we perform the differencing process four times to make the ACF converge to zero faster. The right

column of figure 3.4 displays the time plot of the differenced series $\bigtriangledown^4\eta_n$ and its estimated ACF and PACF. Now the ACF of $\bigtriangledown^4\eta_n$ looks like a damped sine wave and the PACF cuts off at lag 5. However its time plot doesn't show stationary. The time plot decreases rapidly at the right end part. We try $ARIMA(5,4,0)$ model to fit the data, but *R-programming* reports an error: nonstationary data.

The Adams-Bashforth Method: We also apply a cubic regression model and four times differencing to achieve stationary. The right column of figure 3.5 tells us that the ACF of $\bigtriangledown^4\eta_n$ tails off in a damped sine curve and the PACF cuts off rapidly at lag 1. So the $ARIMA(1,4,0)$ model may be appropriate for the transformed series. However $ARIMA(1,4,0)$ model fails the residuals diagnostic checking. We don't find a suitable model for this set of error data.

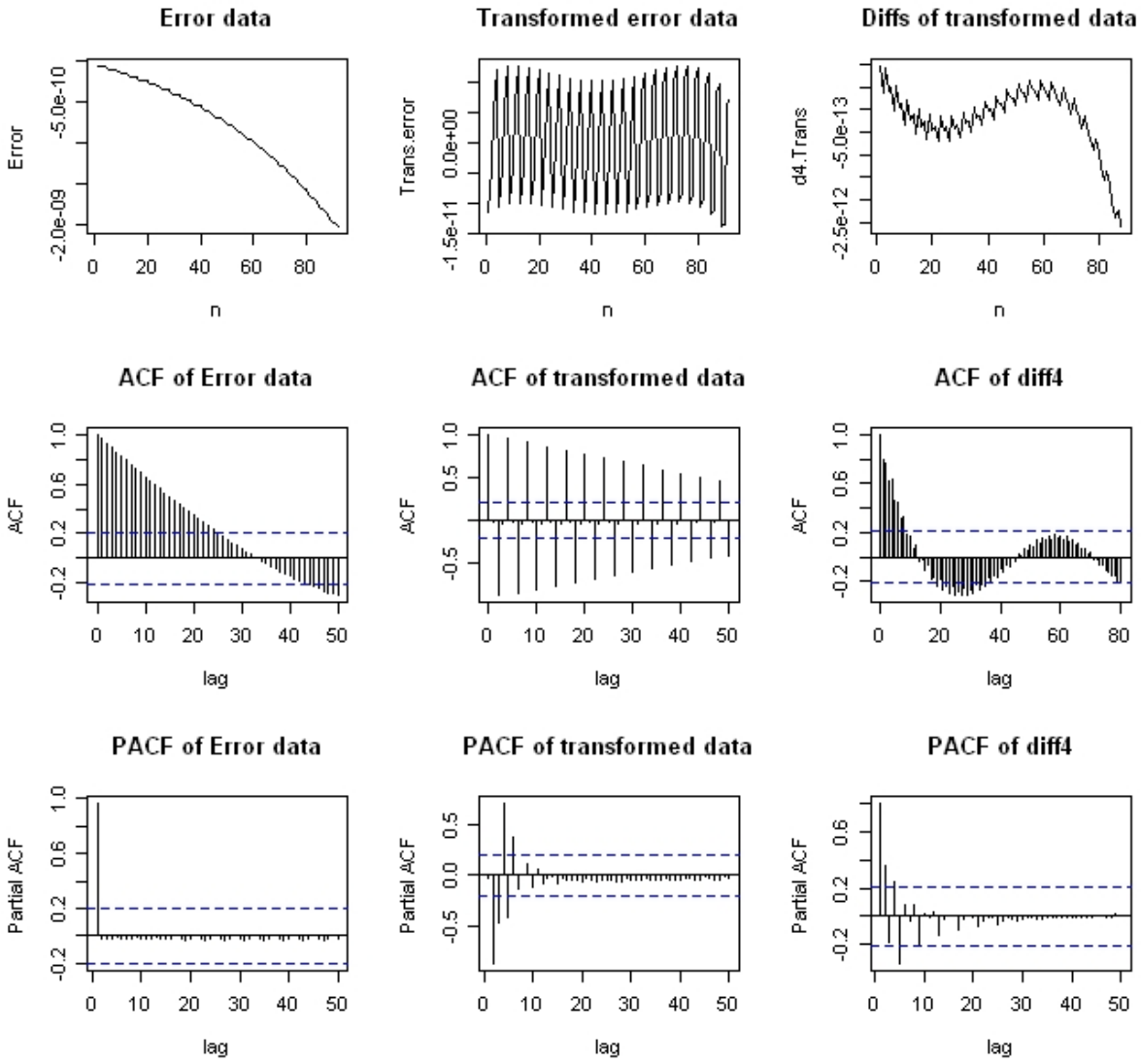

Figure 3.4: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ based on Milne's method

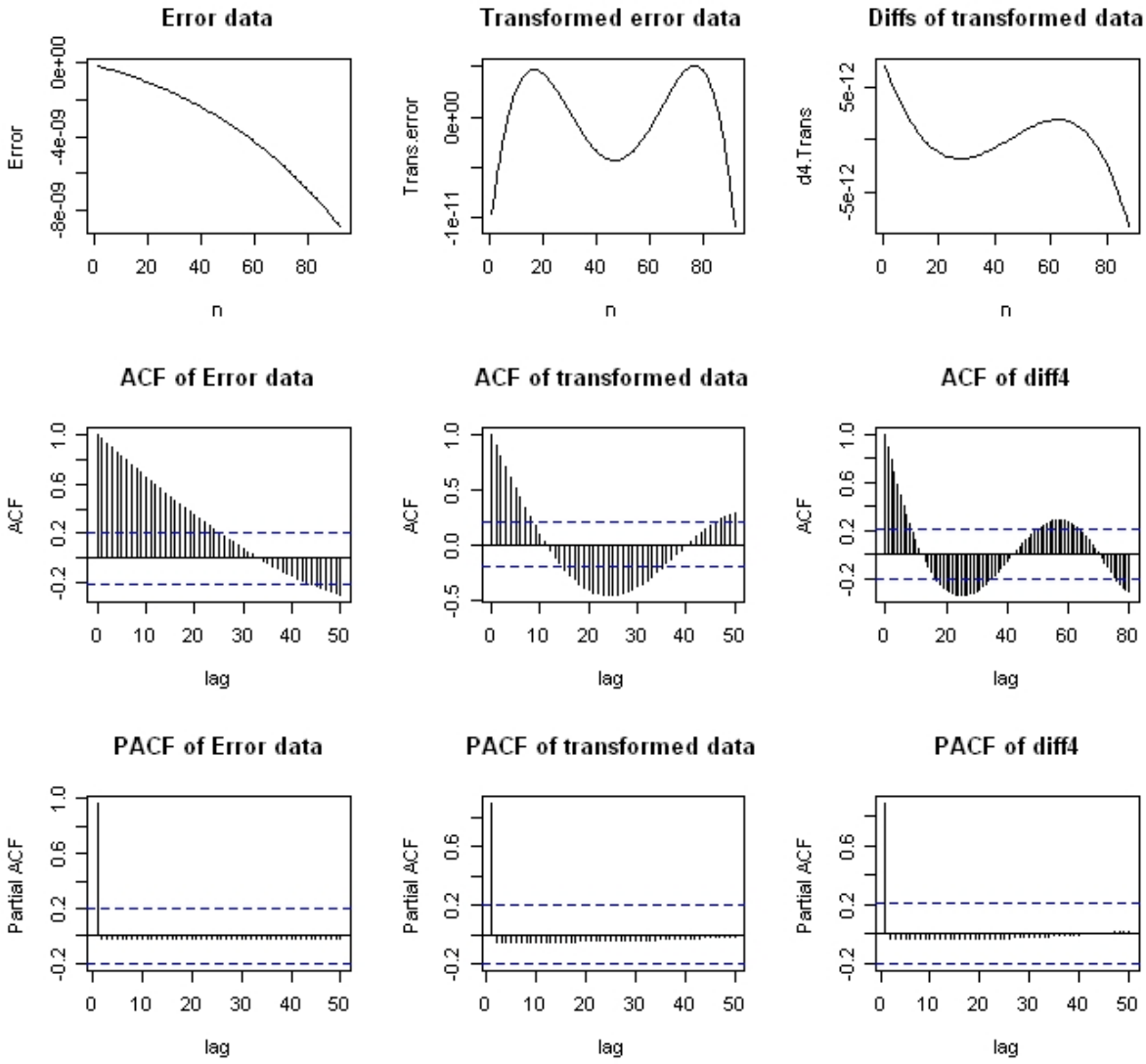

Figure 3.5: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ based on Adams-Bashforth method

3.3.2 Analyzing Error Data from the Equation $\frac{\mathbf{dy}}{\mathbf{dt}} = -\mathbf{y}$

The Switching Method: A cubic curve and one time differencing process are applied to the error data e_n. Figure 3.6 shows the time plot, ACF and PACF of e_n, the transformed error data η_n and the differenced series $\bigtriangledown\eta_n$. The third plot in the first row is the time plot of the differenced data. As we can see, it has a decreasing variance. So it is still not stationary enough. According to the ACF and PACF of η_n and $\bigtriangledown\eta_n$ we use $ARIMA(7,0,0)$ and $ARIMA(3,1,0)$ model to fit the transformed data η_n.

The diagnostic checking results of these two models are given in figure 3.7 and figure 3.8. The residuals of these two $ARIMA$ models could be treated as white noise series. However the **AIC** value for $ARIMA(7,0,0)$ is $AIC = -4838.79$ and the **AIC** value for $ARIMA(3,1,0)$ is $AIC = -4779.58$. So we select the $ARIMA(7,0,0)$ model.

Finally the statistical model for this set of error data is

$$e_n = \beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3 + \phi_1\eta_{n-1} + \phi_2\eta_{n-2} + \cdots + \phi_7\eta_{n-7} + \omega_n, \tag{3.32}$$

where

$$\eta_{n-1} = \quad e_{n-1} - (\beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3), \tag{3.33}$$

$$\vdots$$

$$\eta_{n-7} = \quad e_{n-7} - (\beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3) \tag{3.34}$$

So equation (3.32) can also be written as:

$$e_n = \phi_1 e_{n-1} + \phi_2 e_{n-2} + \cdots + \phi_7 e_{n-7} + C(1 - \phi_1 - \phi_2 - \cdots - \phi_7) + \omega_n, \tag{3.35}$$

where $C = \beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3$.

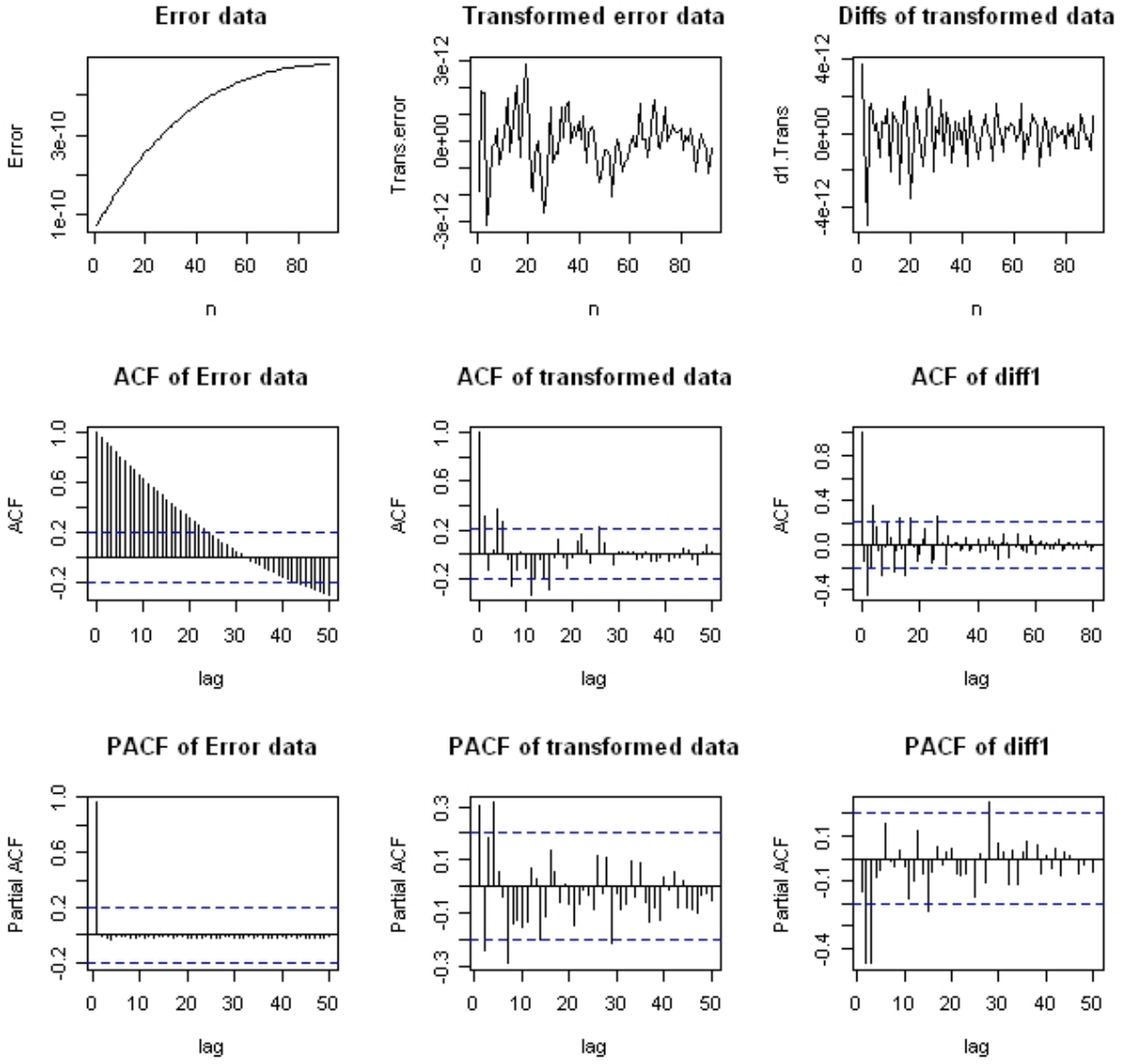

Figure 3.6: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ based on the switching method

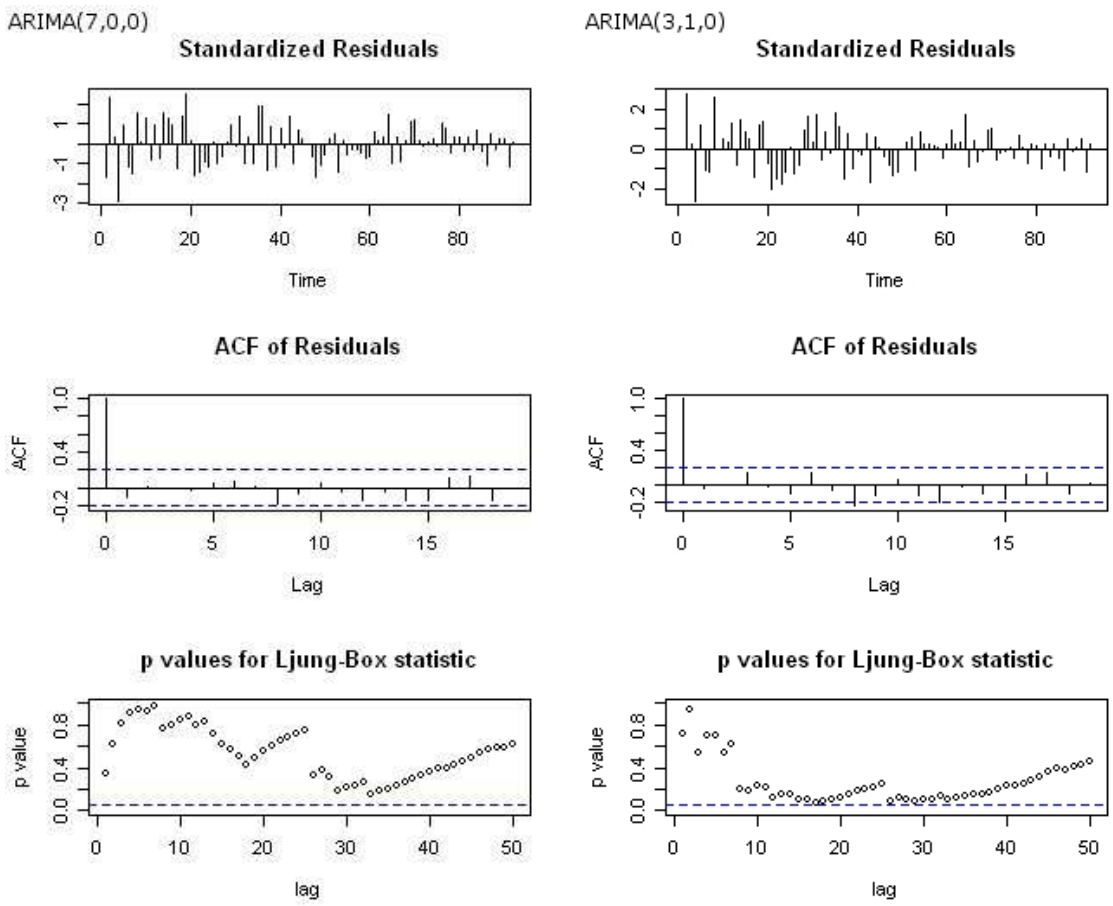

Figure 3.7: The residuals analysis of ARIMA(7,0,0) and ARIMA(3,1,0) models

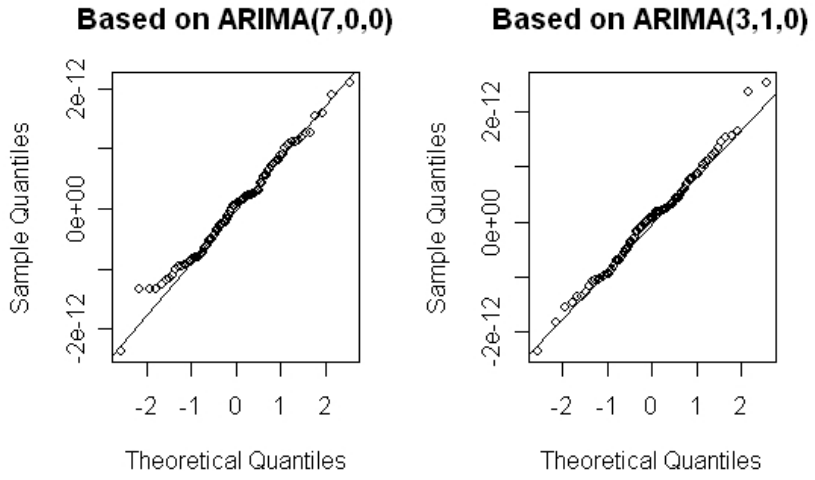

Figure 3.8: The Q-Q plot of the residuls from ARIMA(7,0,0) and ARIMA(3,1,0) models

The Milne's Method: The right column of figure 3.9 tells us that the ACF of $\bigtriangledown\eta_n$ tails off in a damped exponential curve and the PACF cuts off at lag 5. So the $ARIMA(5,1,0)$ model may be appropriate for the transformed data. However $ARIMA(5,1,0)$ model is rejected by the residuals diagnostic checking. We don't find a suitable model for this set of error data.

The Adams-Bashforth Method: After a cubic regression model and eight times differencing, we still can not achieve a stationary series. Figure 3.10 shows the ACF of $\bigtriangledown^8\eta_n$ are still outside the two bands around zero. No ARIMA model is found for this set of data.

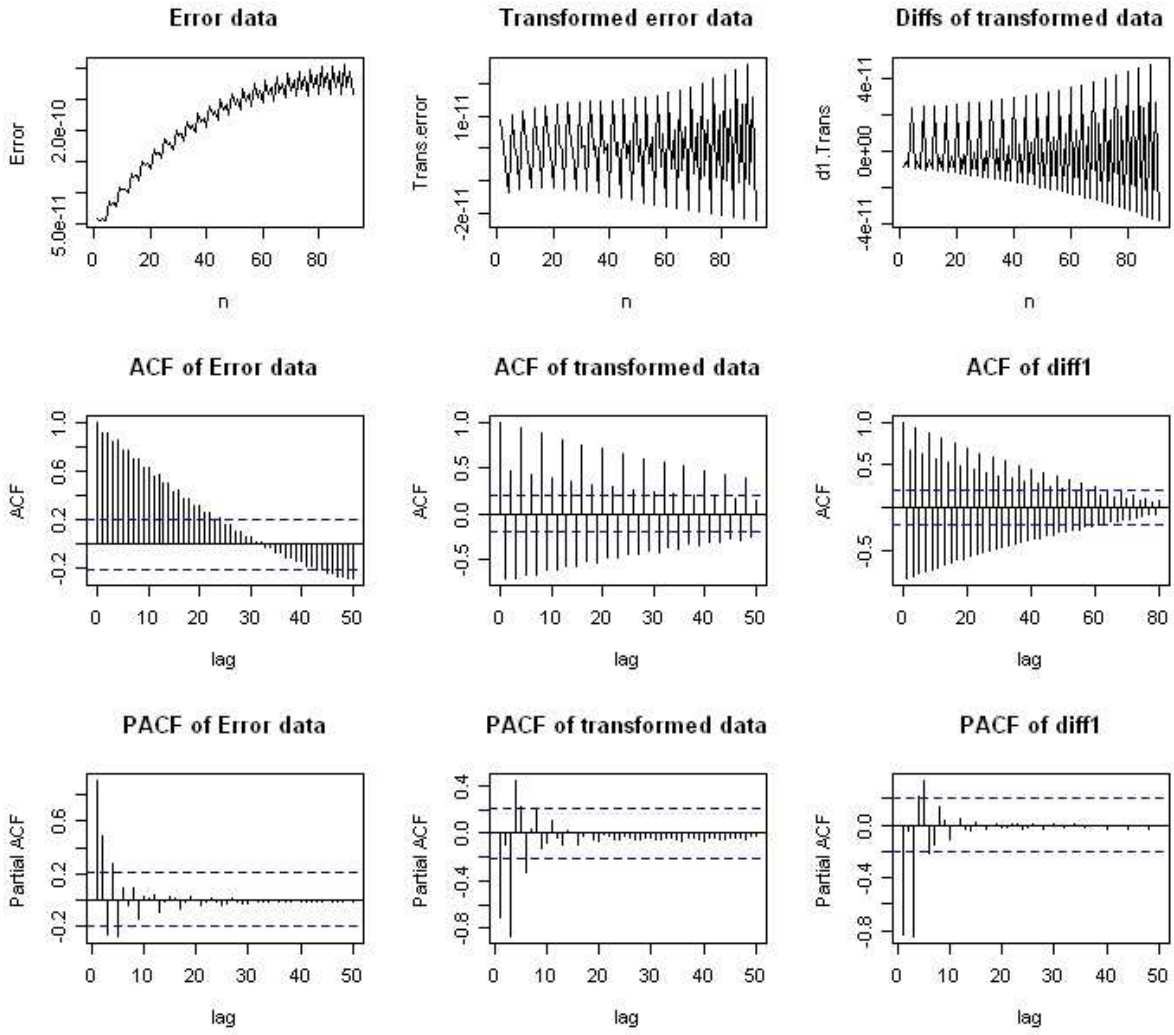

Figure 3.9: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ based on Milne's method

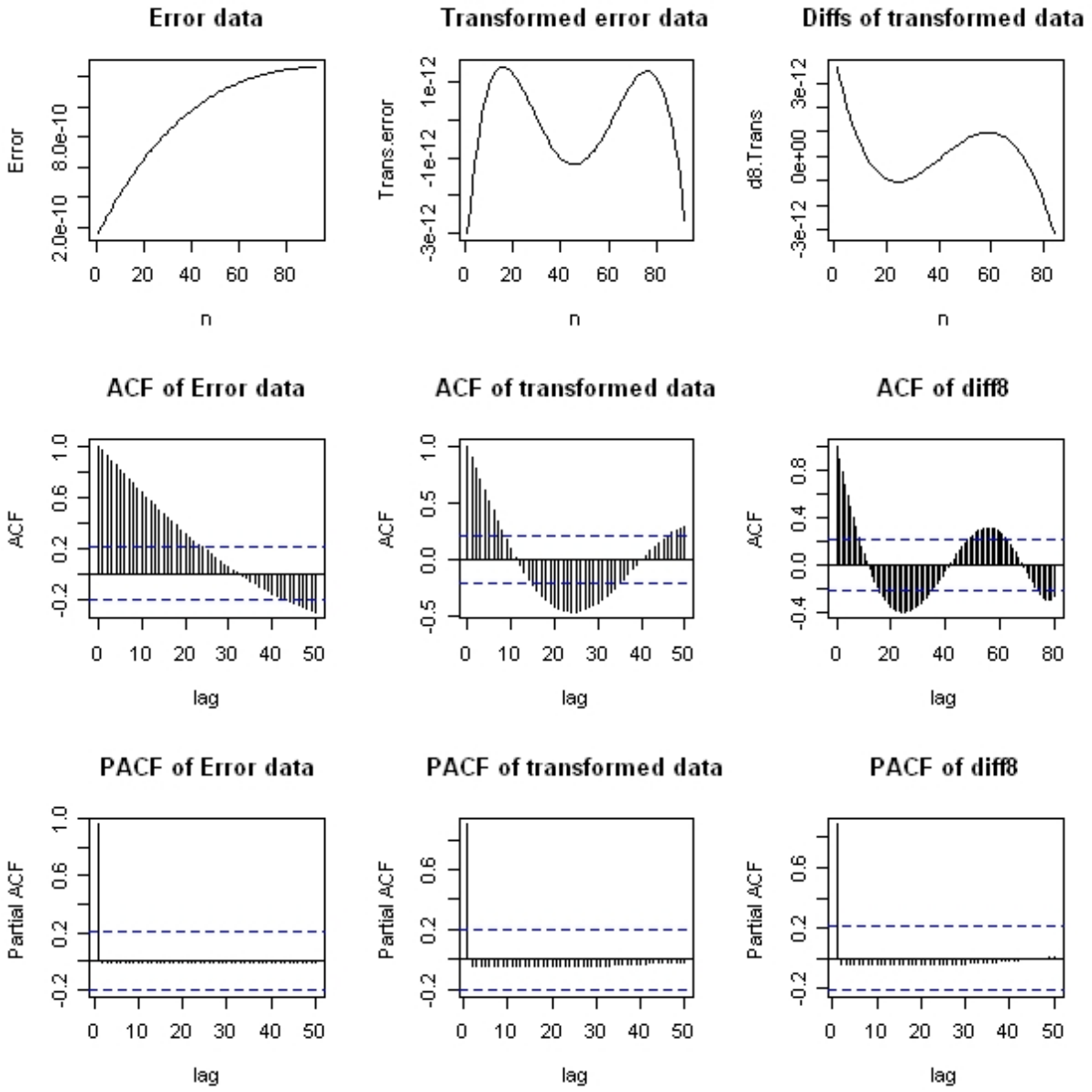

Figure 3.10: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ based on Adams-Bashforth method

3.3.3 Analyzing Error Data from Two Nonlinear Equations

For the error data from two nonlinear equation $\frac{dy}{dt} = 1 + y^2$ and $\frac{dy}{dt} = 1 - y^2$, we apply a quartic regression model to remove trends and then perform the differencing process eight times to achieve stationarity.

Figure 3.11, 3.12 and 3.13 show the time plot, ACF and PACF of the original error data e_n based on equation $\frac{dy}{dt} = 1 + y^2$, the transformed error data η_n and the differenced series $\bigtriangledown^8 \eta_n$ from three methods. The ACFs of three sets of processed data all decay very slowly and PACFs display an extreme cut off at lag 1. We investigate $ARIMA(1, 8, 0)$ model and $ARIMA(1, 0, 0)$ model, but neither of them can fit the transformed data sets.

Same thing happens to the error data based on the differential equation $\frac{dy}{dt} = 1 - y^2$. Figure 3.14, 3.15 and 3.16 show the time plot, ACF and PACF of the original error data e_n based on equation $\frac{dy}{dt} = 1 - y^2$, the transformed error data η_n and the differenced series $\bigtriangledown^8 \eta_n$ from three methods. We test the adequacy of $ARIMA(1, 8, 0)$ model and $ARIMA(1, 0, 0)$ model for error data sets generated by Adams-Bashforth method and the switching method. It turns out they are not suitable. Also the $ARIMA(3, 8, 0)$ model does not fit the data obtained by Milne's method.

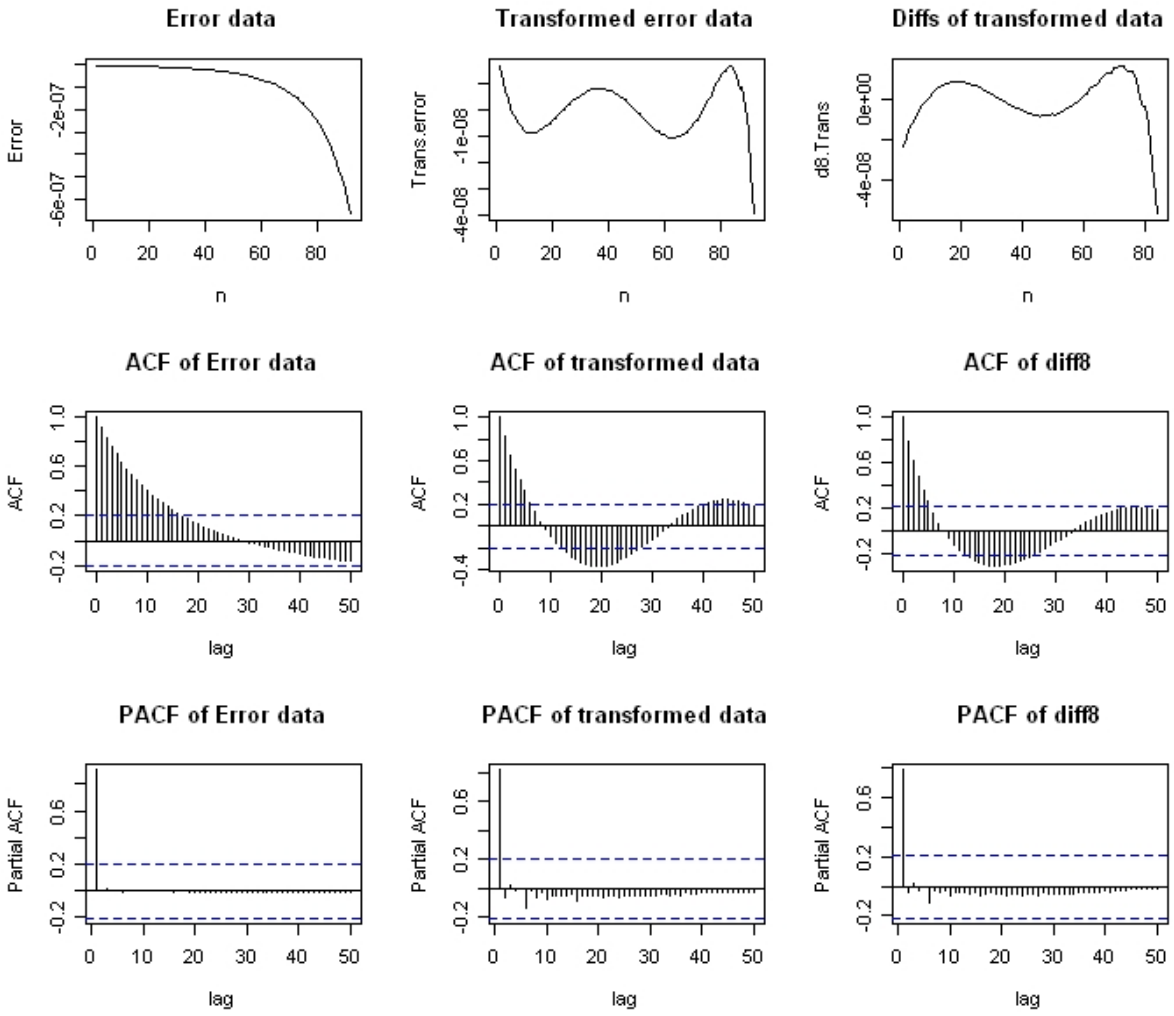

Figure 3.11: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ from equation $\dfrac{dy}{dt} = 1 + y^2$ based on the switching method

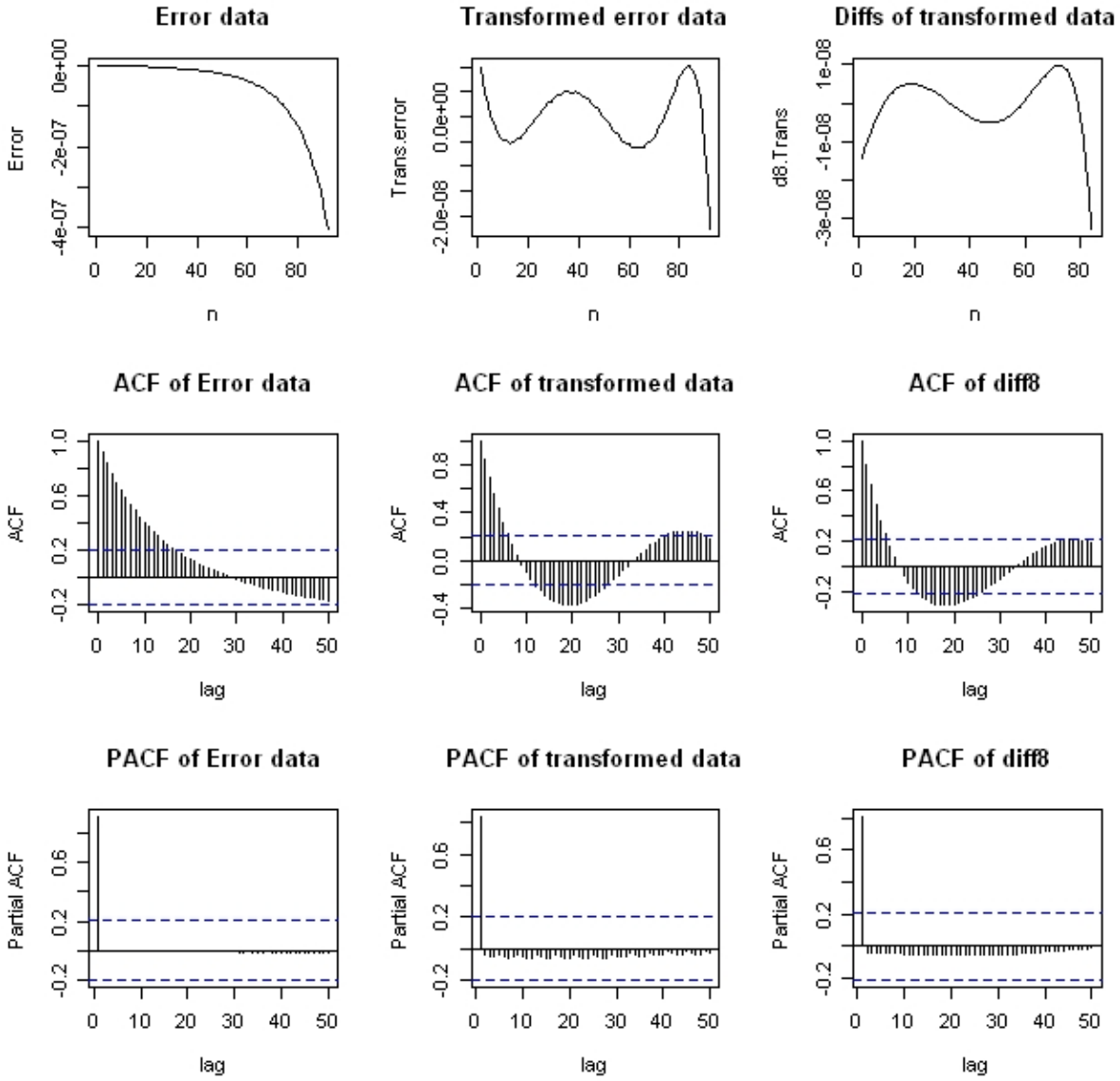

Figure 3.12: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ from equation $\frac{dy}{dt} = 1 + y^2$ based on Milne's method

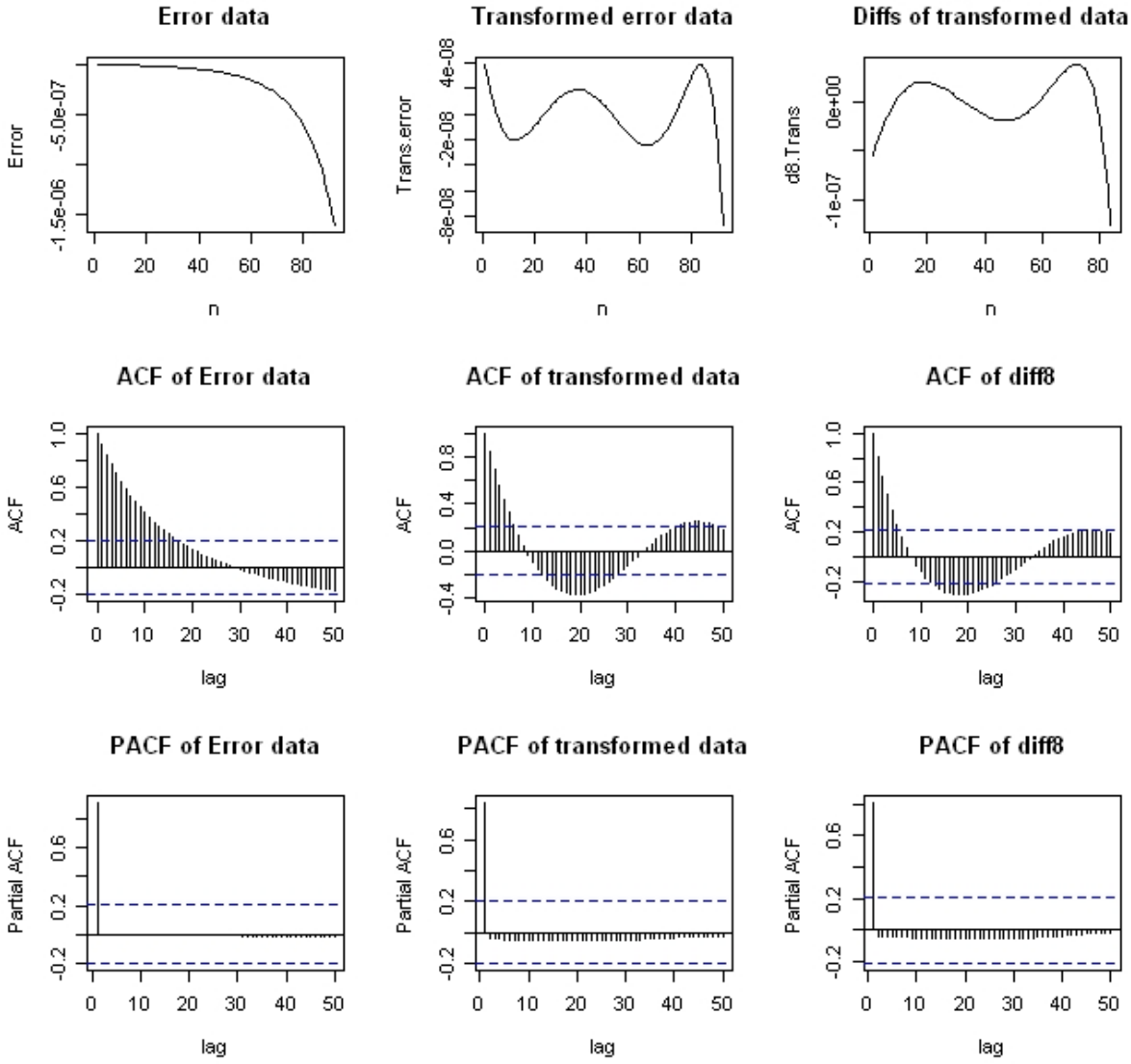

Figure 3.13: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ from equation $\frac{dy}{dt} = 1 + y^2$ based on Adams-Bashforth method

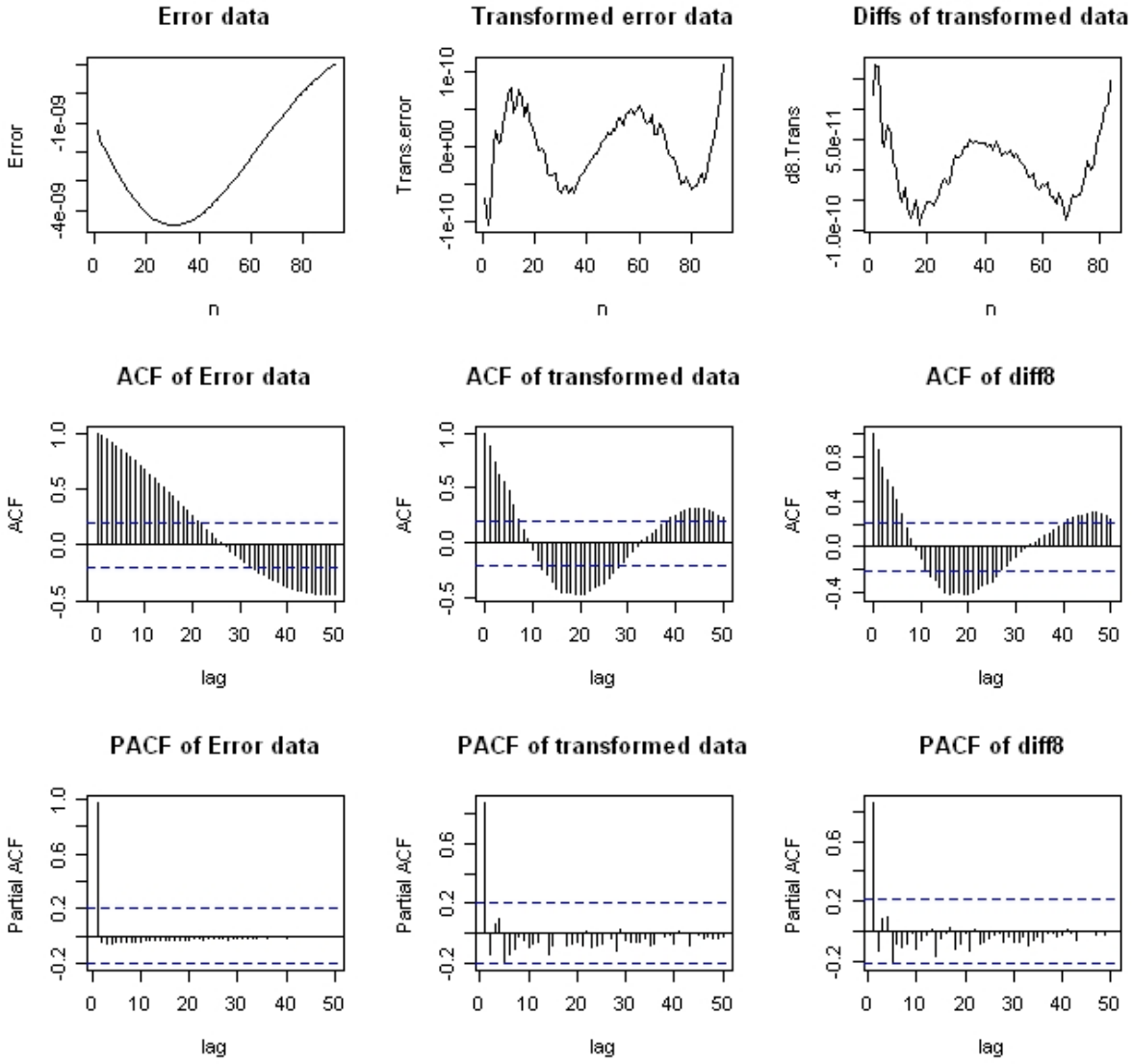

Figure 3.14: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ from equation $\frac{dy}{dt} = 1 - y^2$ based on the switching method

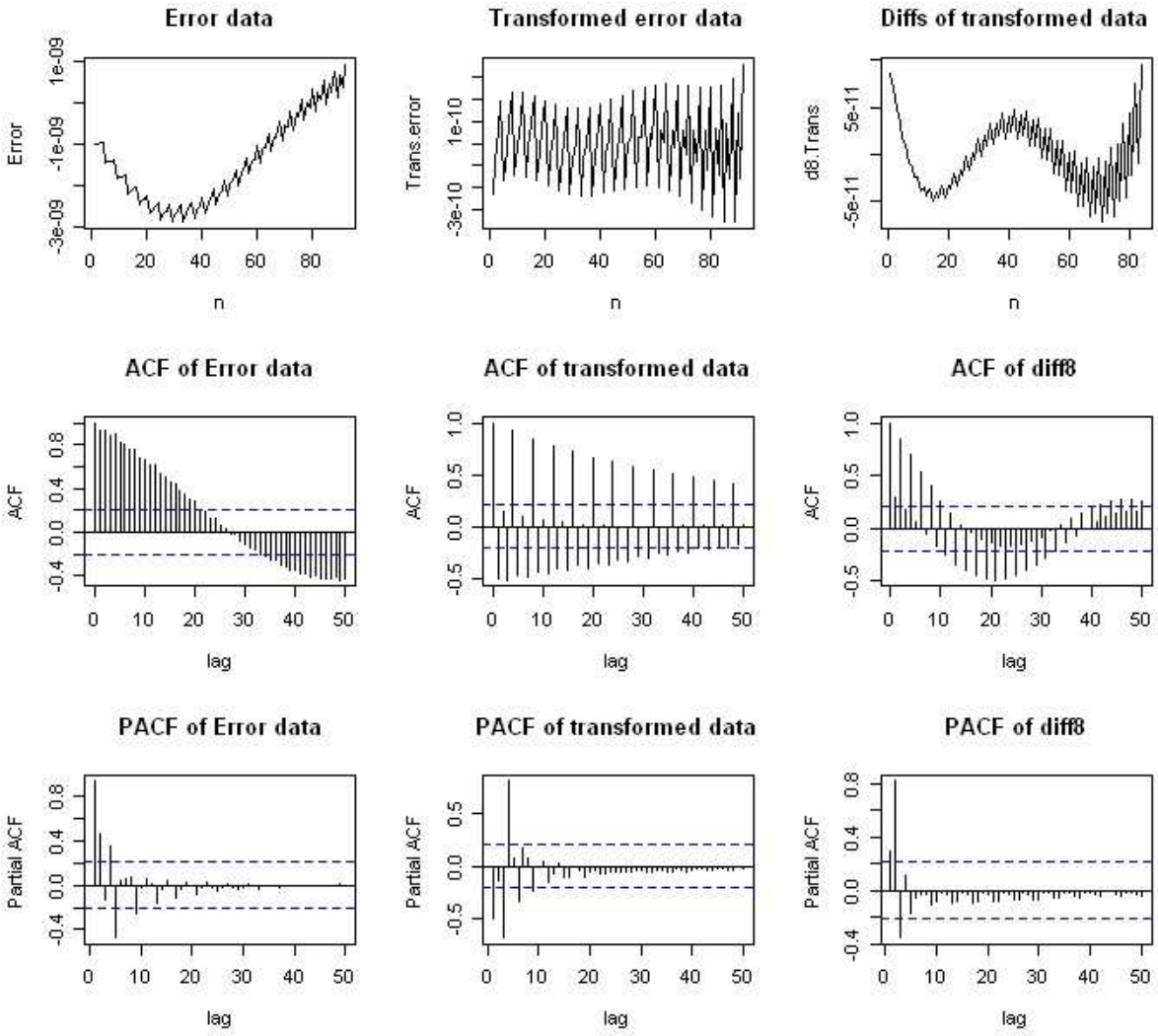

Figure 3.15: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ from equation $\dfrac{dy}{dt} = 1 - y^2$ based on Milne's method

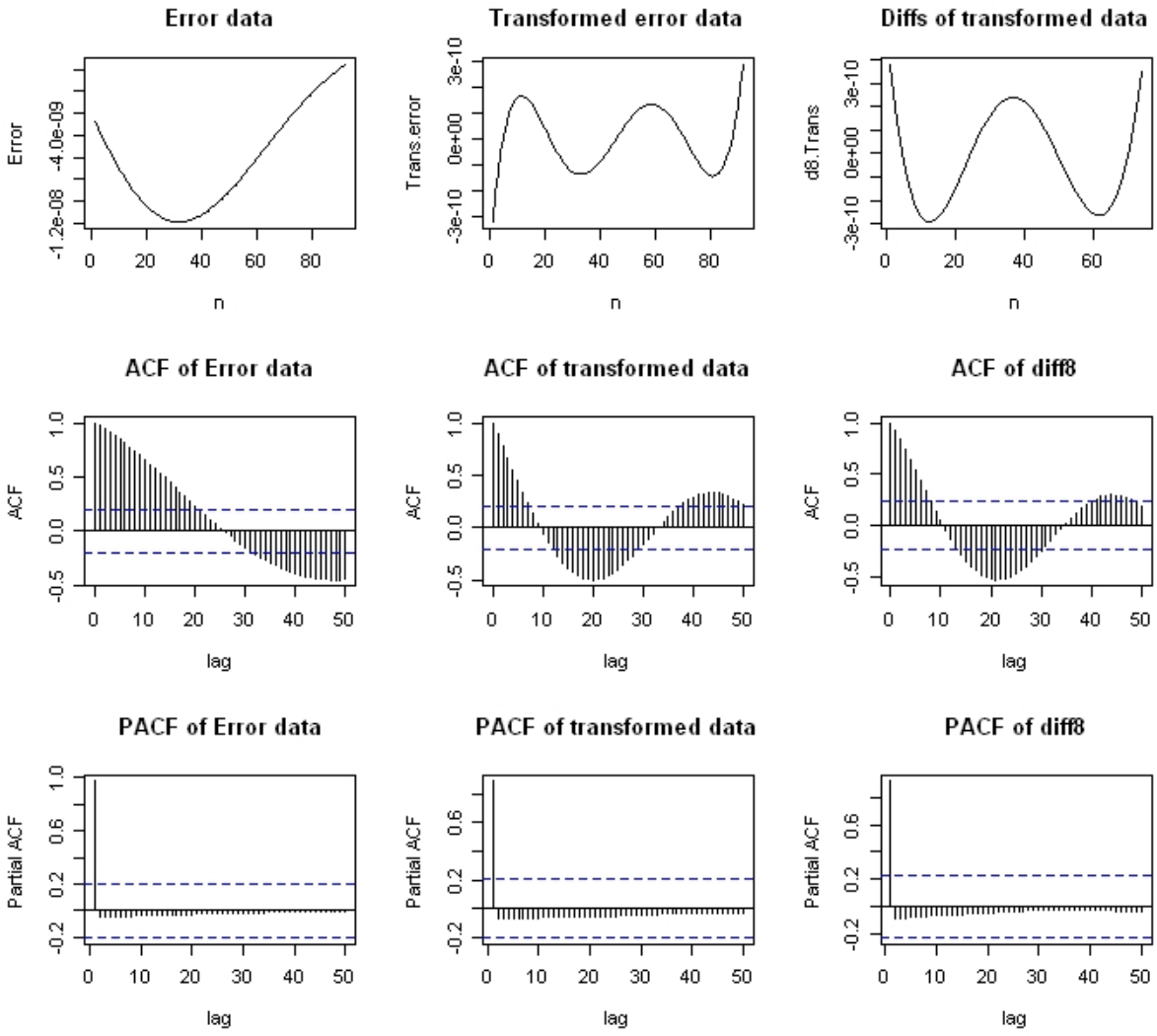

Figure 3.16: The time plot, ACF and PACF of e_n, η_n and $\bigtriangledown\eta_n$ from equation $\frac{dy}{dt} = 1 - y^2$ based on Adams-Bashforth method

3.3.4 Mathematical Interpretation

So far we identify an $ARIMA(4,0,0)$ model to fit the transformed data η_n generated by the fourth-order switching method from the differential equation $\frac{dy}{dt} = y$ and an $ARIMA(7,0,0)$ model for the transformed data η_n generated by the switching method from the differential equation $\frac{dy}{dt} = -y$. However adequate models for other sets of error data haven't been identified yet. In our work we apply parametric curves to remove trends in the error data. The cubic or quartic regression model can remove the trend well but they can not stabilize the variance. So other kinds of transformation should be tried in the future.

We examine the mathematical form of the error data and try to find some factors that affect the error's behavior but we fail to account for in our statistical model. We use the error data e_n generated by Milne's method from the differential equation $\frac{dy}{dt} = y$ as an example to explain this problem. From the result (2.5) we know the error data is obtained by

$$e_{n+1} = y_{n+1} - e^{(n+1)h}, \tag{3.36}$$

actually the random error due to the finite precision of the computing machinery may also be introduced in the answer, so we write the numerical error in a more accurate form:

$$e_{n+1} = y_{n+1} - e^{(n+1)h} + \tau_{n+1}, \tag{3.37}$$

where we assume τ_{n+1} is a white noise series. By Milne's fourth-order algorithm (2.3) $y_{n+1} = y_{n-3} + \frac{4h}{3}(2y_n - y_{n-1} + 2y_{n-2})$, equation (3.37) can also be written as

$$e_{n+1} = y_{n-3} + \frac{4h}{3}(2y_n - y_{n-1} + 2y_{n-2}) - e^{(n+1)h} + \tau_{n+1}. \tag{3.38}$$

where

$$y_{n-3} = e_{n-3} + e^{(n-3)h} + \tau_{n-3} \tag{3.39}$$

$$y_{n-2} = e_{n-2} + e^{(n-2)h} + \tau_{n-2} \tag{3.40}$$

$$y_{n-1} = e_{n-1} + e^{(n-1)h} + \tau_{n-1} \tag{3.41}$$

$$y_n = e_n + e^{nh} + \tau_n \tag{3.42}$$

By substituting equation (3.39), (3.40), (3.41) and (3.42) to equation (3.38), we obtain another expression for the error data e_n:

$$e_{n+1} = \underbrace{\frac{8h}{3}e_n - \frac{4h}{3}e_{n-1} + \frac{8h}{3}e_{n-2} + e_{n-3}}_{\mathbf{A}} + \underbrace{\tau_{n+1} + \frac{8h}{3}\tau_n - \frac{4h}{3}\tau_{n-1} + \frac{8h}{3}\tau_{n-2} + \tau_{n-3}}_{\mathbf{B}}$$
$$+ \underbrace{(e^{-3h} + \frac{8h}{3}e^{-2h} - \frac{4h}{3}e^{-h} + \frac{8h}{3} - e^h)e^{nh}}_{\mathbf{C}}$$

where term $\mathbf{A} = \frac{8h}{3}e_n - \frac{4h}{3}e_{n-1} + \frac{8h}{3}e_{n-2} + e_{n-3}$ means that each step value depends on four previous point values. And term **B** is a linear combination of white noise series. So we can use an $AR(4)$ model to describe these two terms.

In term C we can use Taylor Series for the exponential function:

$$e^h = 1 + h + \frac{h^2}{2} + \frac{h^3}{6} + \dots \tag{3.43}$$

to simplify the expression. It turns out that $\mathbf{C} = O(h^5)e^{nh}$ and this is the error behavior we do not describe in our statistical model.

The model building results tell us that we should consider the error behavior caused by the step size h in the future. The $ARIMA$ model can not fit the data well alone by itself.

CHAPTER IV
MULTIVARIATE NORMALITY TEST

In Chapter II, we define a new fourth-order ODE solver, the switching method. In this method, we choose Milne's or Adams-Bashforth method randomly to obtain the numerical value y_n at each step. Therefore any $\{\mathbf{y_n}, n = 4, 5, \ldots, 99\}$ can be treated as a random variable and each of them has at most 2^{n-3} possible distinct values. Let $\mathbf{Y} = (y_4, y_5, \ldots, y_{99})'$, then $\mathbf{Y}$ is a random vector with 96 components. So is the error vector $\mathbf{E} = (e_4, e_5, \ldots, e_{99})'$ because $e_n = y_n - y(t_n)$, where $y(t_n)$ is the analytical solution at each step. We observe values of $\mathbf{E}$ and obtain a sample $\{E_i, i = 1, 2, \ldots, 10^4\}$. Under random sampling, $\{E_i, i = 1, 2, \ldots, 10^4\}$ are considered to be 10^4 observations of the population $\mathbf{E}$. Assume the population $\mathbf{E}$ has a mean vector μ and a finite covariance matrix $\mathbf{\Sigma}$. Then by central limit theorem, the sample mean vector $\mathbf{M} = \frac{1}{10^4}\sum_{i=1}^{10^4} E_i$ should have an approximate multivariate normal distribution. In this chapter, we assess the multivariate normality of the sample mean vector $\mathbf{M}$.

4.1 Central Limit Theorem

With multivariate statistics, the sampling distributions of the statistics are often difficult to derive. However the *Central Limit Theorem*(CLT) can provide an approximation to the distribution of the sample mean vector. The multivariate version of central limit theorem is given as follows [12] [13].

Theorem 4.1 (The Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be independent observations from any population with mean μ and finite covariance $\mathbf{\Sigma}$. Then*

$$\sqrt{n}(\overline{\mathbf{X}} - \mu) \text{ has an approximate } N_p(\boldsymbol{0}, \mathbf{\Sigma}) \text{distribution}$$

for large sample sizes. Here n should also be large relative to p.

The central limit theorem tells us that the distribution of the sample mean for a

large sample size is nearly normal [14]. In the switching method, we collect a sample $\{E_i, i = 1, 2, \ldots, 10^4\}$, where E_i are 10^4 independent observations of the population $\mathbf{E}$. So the sample mean vector $\mathbf{M} = \overline{\mathbf{E}}$ should have an approximate multivariate normal distribution. Next we assess the multivariate normality of $\mathbf{M}$.

4.2 Measures of Multivariate Skewness and Kurtosis

There are many good tests that we can apply to assess univariate normality. One of the most powerful tests for normality is the Skewness and Kurtosis Test [15]. They are third and fourth moment tests respectively. Skewness is a measure of symmetry of the data around the sample mean. The skewness for a normal distribution is zero, and any symmetric data should also have a skewness near zero. Kurtosis is a measure of whether the data are peaked or flat near the mean compared to a normal distribution. A standard normal distribution has a kurtosis near three. As we can tell that measures of skewness and kurtosis have their diagnostic power to indicate any departure from normality.

Just like the case for univariate normality, we also can test the multivariate normality by testing the sample skewness and kurtosis. Mardia(1970) proposed multivariate extensions of measures of skewness and kurtosis and a test of multivariate normality based on the asymptotic distribution of these two measures. The population measures of multivariate skewness and kurtosis defined in Mardia(1970) and their sample counter-parts [16] are given as follows.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ be a random vector with mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_p)'$ and covariance matrix $\mathbf{\Sigma} = (\delta_{rs})$. The measure of multivariate skewness is defined as

$$\beta_{1,p} = E\{(\mathbf{X} - \mu)'\mathbf{\Sigma^{-1}}(\mathbf{Y} - \mu)\}^3, \tag{4.1}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are independent and identical random vectors. The measure of

multivariate kurtosis is defined as

$$\beta_{2,p} = E\{(\mathbf{X} - \mu)'\mathbf{\Sigma^{-1}}(\mathbf{X} - \mu)\}^2 \tag{4.2}$$

For a normal population, we have $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$.

Suppose that $\mathbf{X_i} = (X_{1i}, X_{2i}, \ldots, X_{pi})'$, where $i = 1, 2, \ldots, n$, are n independent observations on $\mathbf{X}$. Let $\overline{\mathbf{X}} = (\overline{X_1}, \overline{X_2}, \ldots, \overline{X_p})'$ and $\mathbf{S} = (S_{rs})$ denote the sample mean vector and sample covariate matrix respectively. For this sample, the measures of skewness and kurtosis corresponding to $\beta_{1,p}$ and $\beta_{2,p}$ are given by

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{(\mathbf{X_i} - \overline{\mathbf{X}})'\mathbf{S^{-1}}(\mathbf{X_j} - \overline{\mathbf{X}})\}^3 \tag{4.3}$$

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\{(\mathbf{X_i} - \overline{\mathbf{X}})'\mathbf{S^{-1}}(\mathbf{X_i} - \overline{\mathbf{X}})\}^2 \tag{4.4}$$

Mardia(1970) has also shown that for large samples from a multivariate normal distribution, the scaled multivariate skewness is asymptotically distributed as a chi-square random variable with f degrees of freedom.

$$A = \frac{n}{6}b_{1,p} \sim {\chi_f}^2, \text{where} f = \frac{p(p+1)(p+2)}{6} \tag{4.5}$$

and the multivariate kurtosis following appropriate scaling follows a standard normal distribution.

$$B = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \sim N(0,1) \tag{4.6}$$

Therefore we can test the multivariate normality by testing $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$ with the help of the test statistics A from result (4.5) and B from result (4.6) respectively [17] [18]. If either result is significant, then multivariate normality is rejected.

4.3 Assessing Multivariate Normality of Sample Mean Vector

Since the multivariate central limit theorem is based on asymptotic limits, the distribution approximations are only valid when the sample size is large enough, especially be large relative to the dimension of vectors. So in our study of this section we change the length of step size h and let $h = .05$ here. Then there are 20 step points in the interval $[0, 1]$. Other computation procedure to produce error vectors by using the switching method is the same with what we did in chapter II. Now our error vector becomes $\mathbf{E} = (e_4, e_5, \ldots, e_{19})'$ with 16 variables. So does the sample mean vector $\mathbf{M} = \overline{\mathbf{E}}$.

For each differential equation in chapter II, we use the switching method to obtain error vectors $\mathbf{E}$ and define the sample mean vector $\mathbf{M}$ is the average of 10^4 independent observations of $\mathbf{E}$. Our goal in the section is to test the multivariate normality of $\mathbf{M}$. We infer the distribution of the population mean vector $\mathbf{M}$ from its samples. Let's consider a sample of $\mathbf{M}$ containing 10^3 independent observations $\{M_1, M_2, \ldots, M_{1000}\}$, and each observation M_i is a random vector with 16 variables. We also can use a random matrix $\mathbf{X}$ to indicate the sample of $\mathbf{M}$ vector.

$$\mathbf{X} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p1} & M_{p2} & \cdots & M_{pn} \end{bmatrix} = [M_1, M_2, \ldots, M_n] \tag{4.7}$$

the column vectors $M_1, M_2, \ldots, M_n$ represent independent observations on $\mathbf{M}$ and the (i, j) entry in this random matrix indicate the ith variable of the jth observation of $\mathbf{M}$. Also n is the sample size and p is the dimension. In our study $n = 1000$ and $p = 16$.

One example of random matrix $\mathbf{X}$ of sample mean vector $\mathbf{M}$ based on differential

equation $\frac{dy}{dt} = y$ is given as follows:

$$\mathbf{X} = \begin{bmatrix} -1.13924E-07 & -1.13871E-07 & \cdots & -1.13854E-07 \\ -1.90038E-07 & -1.89325E-07 & \cdots & -1.91651E-07 \\ \vdots & \vdots & \ddots & \vdots \\ -1.79489E-06 & -1.79132E-06 & \cdots & -1.79926E-06 \end{bmatrix}$$

Similarly we can produce random matrix $\mathbf{X}$ based on other differential equations.

Once a sample matrix of $\mathbf{M}$ is collected, a random sample from the population $\mathbf{M}$ is formed. Then we can apply result (4.3) and (4.4) to calculate the sample measures of skewness and kurtosis. The sample mean vector $\overline{\mathbf{M}}$ can be calculated by

$$\overline{\mathbf{M}} = \frac{1}{n}\mathbf{XI} \tag{4.8}$$

and the sample covariance matrix $\mathbf{S}$ can be found by

$$\mathbf{S} = \frac{1}{n}\mathbf{X}(\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{II}')\mathbf{X}' \tag{4.9}$$

where $\mathbf{X}$ is the sample matrix on $\mathbf{M}$ in the form (4.7). Also the covariance matrix $\mathbf{S}$ must be nonsingular [19]. Actually this is the other reason that we change the step size for the normality test. If we choose error vector with dimension 100, it makes $\mathbf{S}$ close to singular.

Finally we test the multivariate normality of $\mathbf{M}$ by testing the asymptotic distribution of measures of skewness and kurtosis. The asymptotic distributions are given in results (4.5) and (4.6). If either result (4.5) or (4.6) is significant, the multivariate normality of $\mathbf{M}$ is rejected.

We collect a sample matrix of mean vector $\mathbf{M}$ from each differential equation and test the multivariate normality. Table 4.1, 4.2, 4.3 and 4.4 report sample measures of skewness $b_{1,p}$ and kurtosis $b_{2,p}$, Mardia's test statistics A and B described along with the corresponding p values. As we can see, with 5% significant level all these

mean vectors may be regarded as samples from a 16-variate normal population. We also choose other random samples of $\mathbf{M}$ to test the multivariate normality for each equation. Totally we choose 10 samples for each equation to test the normality. The result shows that all those 10 samples based on equation $\frac{dy}{dt} = 1 + y^2$ and equation $\frac{dy}{dt} = -y$ may be regarded as samples from a normal distribution. However for equation $\frac{dy}{dt} = 1 - y^2$ and equation $\frac{dy}{dt} = y$, the multivariate normality of $\mathbf{M}$ is rejected by two chosen samples. Anyway we still can say that the sample mean vector $\mathbf{M}$ has a limiting multivariate normal distribution.

Table 4.1: Test multivariate normality of $\mathbf{M}$ based on equation $\frac{dy}{dt} = y$

MULTIVARIATE NORMALITY TEST:			
	$b_{1,p}$	A	$p-value$
Based on skewness:	5.1934205	865.57008	0.1113345
	$b_{2,p}$	B	$p-value$
Based on kurtosis:	287.43534	−0.372004	0.7098897

Table 4.2: Test multivariate normality of $\mathbf{M}$ based on equation $\frac{dy}{dt} = -y$

MULTIVARIATE NORMALITY TEST:			
	$b_{1,p}$	A	$p-value$
Based on skewness:	5.0466743	841.11238	0.263752
	$b_{2,p}$	B	$p-value$
Based on kurtosis:	287.33667	−0.437007	0.6621059

Table 4.3: Test multivariate normality of $\mathbf{M}$ based on equation $\frac{dy}{dt} = 1 + y^2$

MULTIVARIATE NORMALITY TEST:			
	$b_{1,p}$	A	$p-$value
Based on skewness:	4.6998156	783.3026	0.7892705
	$b_{2,p}$	B	$p-$value
Based on kurtosis:	286.05324	-1.282542	0.1996526

Table 4.4: Test multivariate normality of $\mathbf{M}$ based on equation $\frac{dy}{dt} = 1 - y^2$

MULTIVARIATE NORMALITY TEST:			
	$b_{1,p}$	A	$p-$value
Based on skewness:	4.4182573	736.37622	0.9784307
	$b_{2,p}$	B	$p-$value
Based on kurtosis:	287.49439	-0.333099	0.7390596

CHAPTER V
CONCLUSION

In this thesis, our goal is to develop statistical models that provide suitable descriptions for the numerical error. An autoregressive integrated moving average model ($ARIMA$) is applied to fit the numerical error generated from three fourth-order numerical methods: Milne's method, Adams-Bashforth method and the switching method. Using the actual error data from three differential equations we want to specify a subclass of $ARIMA$ models to each data series. Results show that the data series generated by the switching method from linear differential equations can be described by $ARIMA$ models but others can not. Based on the mathematical form of the numerical error, we suggest that other statistical models should be applied in the future; especially the effect caused by the step size should be considered. Finally we assess the multivariate normality of sample mean vectors generated by the switching method as an application of the multivariate central limit theorem.

BIBLIOGRAPHY

[1] Uri M. Ascher and Linda R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Philadelphia,PA: Society for Industrial and Applied Mathematics, c1998, first edition, 1998.

[2] John R. Dormand. *Numerical Methods for Differential Equations: a Computational Approach*. Boca Raton: CRC Press, c1996, first edition, 1996.

[3] Huo L. Holder, D. and C. Martin. The control of error in numerical methods. *to appear*.

[4] James D. Hamilton. *Time Series Analysis*. Princeton,NJ: Princeton University Press, c1994, first edition, 1994.

[5] Robert H. Shumway. *Time Series Analysis and Its Applications*. New York: Springer, c2000, first edition, 2000.

[6] Jenkins G.M. Box, G.E.P. and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. New Jersey: Prentice-Hall, c1994, third edition, 1994.

[7] Ulf. Grenander. *Statistical Analysis of Stationary Time Series*. New York: Wiley, c1957, first edition, 1957.

[8] T.W. Anderson. *The Statistical Analysis of Time Series*. New York: Wiley, c1971, first edition, 1971.

[9] Raymond H. Myers. *Classical and Modern Regression with Applications*. Boston: PWS-KENT, c1990, second edition, 1990.

[10] Wayne A. Fuller. *Introduction to Statistical Time Series*. New York: Wiley, c1996, New York, 2nd edition, 1996.

[11] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[12] Richard Arnold Johnson. *Applied Multivariate Statistical Analysis*. Upper Sanddle River,NJ: Prentice Hall, c2002, 5th edition, 2002.

[13] Donald B. White. An application of a multivariate central limit theorem to sampling without replacement. *Journal of Multivariate Analysis*, 24(1):123–128, 1988.

[14] Gerorge Casella and Roger L. Berger. *Statistical Inference*. Australia;Pacific Grove,CA: Thomson Learning, c2002, 2nd edition, 2002.

[15] Henry C. Thode. *Testing for Normality*. Switzerland: Marcel Dekker AG, c2002, first edition, 2002.

[16] K.V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

[17] K.V. Mardia. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Biometrika*, 36(2):115–128, 1974.

[18] N.J.H.Small. Marginal skewness and kurtosis in testing multivariate normality. *Applied Statistics*, 29(1):85–87, 1980.

[19] Ravindra Khattree and Dayanand N. Naik. *Applied Multivariate Statistics with SAS Software*. Cary,NC:SAS Institute, c1999, 2nd edition, 2003.

PERMISSION TO COPY

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Texas Tech University or Texas Tech University Health Sciences Center, I agree that the Library and my major department shall make it freely available for research purposes. Permission to copy this thesis for scholarly purposes may be granted by the Director of the Library or my major professor. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my further written permission and that any user may be liable for copyright infringement.

Agree (Permission is granted.)

Bo He
Student Signature Date 11/24/2007

Disagree (Permission is not granted.)

_____
_____
Student Signature Date