A COMPARATIVE STUDY OF TIME-SERIES FORECASTING APPLIED TO STOCK MARKET PRICE

A Thesis

submitted in partial fulfillment

of the requirements for the Degree

of

Master of Science in Statistics

in the

University of Canterbury

by

Issarest Weeraprajak

University of Canterbury

2000

TABLE OF CONTENTS

Abstract

This thesis is a comparative study on forecasting New Zealand stock market daily closing prices by treating them as a time series. The methods used here are Box and Jenkins autoregressive integrated moving average (ARIMA) model, Bayesian dynamic linear model and Fuzzy neural networks. These methods are compared by using simple trading strategies, resulting in potentially profitable forecasting especially through the fuzzy neural networks.

In addition, the final part of this thesis are summary and comments on different methods that have been used by researchers to predict the stock prices.

Acknowledgments

I would like to thank my Lord and Savior Jesus Christ who is without doubt my ultimate source of strength and inspirations. Then I would like to thank admiral Charn and Korapin Weeraprajak, my parents, for their love and support, Paul and Nicky Clark for being the greatest guardians, Dr. Easaw Chacko, my supervisor, for his wisdom and guidance.

Finally thanks to all my friends who have really support me physically and spiritually.

Chapter 1

1 Introduction

This thesis is concerned with forecasting stock market prices, which is regarded as forecasting of a suitable time series model. We shall start with a general introduction to the basic concepts of a time series that is used in this thesis. Following the general introduction is a general outline of what the thesis aims to do. This is then followed by the layout of the thesis and finally the main results are given.

1.1 General concepts of Time Series

People have always been interested in predicting the future. Before the era of "scientific forecasting" prophets have, in many cases with remarkable accuracy predicted future events. These were of a general nature but modern scientific time series analysis tends to focus on specific events such as stock market prices.

Before the 1920's, scientific forecasting was done by extrapolating the series through a global fit in the time domain. The beginning of "modern" time series prediction might be set at 1927 when Yule (1927) invented the autoregressive technique in order to predict the annual number of sunspots. His model predicted the next value as a weighted sum of previous observations of the series. In order to obtain "interesting" behavior from such a linear system; outside intervention in the form of "external shocks" must be assumed.

Generally, time series analysis has 3 aims: *modeling*, *characterization* and *forecasting*. The aim of modeling is to find a description that accurately captures features of long-term behavior of the system. The aim of characterization is to determine fundamental properties with little or no *a priori* knowledge. The aim of forecasting is to accurately predict the short-term evolution of the system. This thesis deals with forecasting.

A time series is a special case of a stochastic process. We introduce here a brief introduction to the basic concept of a stochastic process as required for time series analysis. A *stochastic process* $\{Y_t\}$ is a collection of random variables where the parameter t belongs to a parameter space (or index) T.

$$\{Y_t\} = (\ldots Y_1,\ Y_2,\ \ldots,\ Y_n).$$

The set of possible values which each Y_t can take (commonly called the *state space),* is here assumed to be $\subset \mathbb{R}$. In our case the parameter space T is assumed to be the set of integers $\{0,\ 1,\ 2,\ \ldots\}$.

The three assumptions that are normally made for stochastic processes are that the processes are *stationary*, are *Normally distributed* and are *ergodic.* Stationarity means the invariance of the probabilistic structure of a process over time. A time series (or stochastic process) $\{Y_t,\ t \in T\}$, is said to be *strictly stationary* if the joint distribution of $\left(Y_{t_1},\ \ldots,\ Y_{t_k}\right)'$ and $\left(Y_{t_1+h},\ \ldots,\ Y_{t_k+h}\right)'$ are the same for all possible collection of such Y_i. Note that a time series is defined completely in a probabilistic sense if the joint distributions for all

finite sets of random variables $\left(Y_{t_1}, \ldots, Y_{t_k}\right)$ are defined. However, in most applications, the form of the joint distribution function is not known. A great deal can be accomplished, however, by dealing only with first two moments of the time series. In lines with this approach, a time series is *weakly stationary* if:

1. $E|Y_t|^2 < \infty$ for all $t \in T$,
2. $E(Y_t)$ is a constant for all $t \in T$,
3. $\operatorname{cov}\{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k}\} = \operatorname{cov}\{Y_{t_1+h}, Y_{t_2+h}, \ldots, Y_{t_k+h}\}$ for all $k \in I_{(1,\infty)}$ and $t_1, \ldots, t_k, h \in T$.

The Normal assumption means that for every k random variables belonging to the process $\{Y_t\}$ their joint density function follows a multivariate normal density function

$$f\left(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k}\right) = (2\pi)^{-\frac{k}{2}} |\Pi|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu)' \Pi (y-\mu)\right\}, \quad y \in \mathbb{R}^k$$

where y is a vector of states, μ is a vector of mean values and Π is the inverse covariance matrix of the k variables.

Therefore the Gaussian process is determined if we know μ and Π whose generic element $\sigma_{r,s}$ is

$$\sigma_{r,s} = \operatorname{cov}(Y_r, Y_s) = E\left[(Y_r - \mu_r)(Y_s - \mu_s)\right], \quad \forall\, r,s = 1, 2, \ldots, n.$$

If the process is weakly stationary, μ will be constant for every t and the elements of $\prod$ will depend only on the difference between the indices r and s which is invariant with respect to a translation over time. Hence for Gaussian processes weak stationarity and strong stationarity are coincident.

Ergodicity means that the sample moment of the observed time series $\{Y_t\}$, $t = 1, \ldots, n$ for a finite n converge to the corresponding population moments as $n \to \infty$. Stationarity is necessary for ergodicity but not vice versa. In this thesis how ever, if the time series is mentioned to be stationary then the Gaussianity and ergodicity properties are assumed.

The most fundamental process in time series is a white noise or purely random process $\{\varepsilon_t\}$, where all random variables have zero mean, constant variance and are uncorrelated, i.e.:

$$E(\varepsilon_t) = 0 \qquad \forall t;$$

$$E(\varepsilon_t \varepsilon_s) = \begin{cases} \sigma_\varepsilon^2 & \text{if} \quad t = s \\ 0 & \text{if} \quad t \neq s \end{cases} \qquad \forall\ t,\ s.$$

A Gaussian white noise process is one that is normally distributed. In this thesis ε_t is assumed to be a Gaussian white noise process.

The related process in continuous time is the Brownian motion or Wiener process $\{W_t\}$ where $\{W(t) : t \geq 0\}$ is a Gaussian process, and for any t, h

with $t+h>0$, the difference $W(t+h)-W(t)$ has expectation 0 and variance $|h|$.

1.2 Aim of the thesis

The aim of this thesis is to apply different mathematical models namely Box and Jenkins autoregressive integrated moving average (ARIMA) model, Bayesian dynamic linear model and Fuzzy neural networks to predict some New Zealand stock market closing prices. An additional aim is to find out if profitable trading can be made under such models. Several models used by researchers to predict stock prices are also briefly discussed.

1.3 Outline of the thesis

This thesis is divided in to 6 chapters. Here, chapter 1 is the general introduction of the thesis. Chapters 2 to 4 review Box and Jenkins models, Bayesian dynamic linear model and Fuzzy neural networks respectively. These chapters are in two parts, the first part where theoretical aspects of the model are examined and the remaining part where the results of using the model are shown. Chapter 5 is a summary and comments on different methods that have been used by researchers to predict the stock prices. Lastly, chapter 6 is the conclusion of this thesis.

1.4 Results of the thesis

The New Zealand stock market daily closing prices analyzed in this thesis are Air New Zealand Ltd. "A", Brierley Investments Ltd., Carter Holt Harvey Ltd., Lion Nathan Ltd. and Telecom Corporation of New Zealand Ltd. The data span 1872 trading days, with the first observation being 1st of May 1992 and the last observation being 7th of October 1999.

The above stock prices were analyzed under ARIMA model, Bayesian dynamics linear model and fuzzy neural networks with similar trading strategies (details in sections 2.5, 3.7 and 4.5). The results given in table 2.1, 3.1 and 4.1 show that fuzzy neural networks out-performed other models in every stock price except Brierley Investments stocks where the Box and Jenkins model performed better.

Chapter 2

2 The Box and Jenkins Approach

2.1 Introduction

The paper by Box and Jenkins (1962) is concerned with developing the relationship between linear predictors and control theory. In doing so, it sets out the relationship between exponential smoothing procedures and certain linear times series models. Exponential smoothing procedures are concerned with forecasting nonstationary series, and Box and Jenkins show how *differencing* enables the classical theory of stationary stochastic process to be applied to the problem of forecasting nonstationary series. The resulting models belong to what is called the autoregressive-integrated-moving average, or ARIMA, class of time series.

Box and Jenkins published a number of other papers in 1960s and at the end of the decade provided a unified approach to ARIMA modeling in their book; see Box and Jenkins (1970). Since then the Box and Jenkins method has been widely used in many time series applications and because of its popularity, we investigate here how the method performs in forecasting stock market prices. In this chapter the Box and Jenkins models will be briefly reviewed. Section 2 reviews the models for stationary time series such as the autoregressive (AR) model and the moving average (MA) model. Section 3 reviews how non-stationarity is modeled. Section 4 reviews model identification techniques that

are used to identify stationarity and the order of Box and Jenkins model. Section 5 looks at stock trading simulation under Box and Jenkins method.

2.2 Model for Stationary Time series

To model a time series using the Box and Jenkins methods we first reduce the time series to a stationary one (see section 2.3). This is usually done by differencing. Then the stationary time series is analyzed to determine the order of the *moving average* (MA) or *autoregressive* (AR) part.

2.2.1 Model Operators

Define the operator, B the backward shift operator; by $BY_t = Y_{t-1}$.

Define the operator, ∇ the difference operator; by $\nabla Y_t = Y_t - Y_{t-1} = (1-B)Y_t$

2.2.2 Autoregressive (AR) models

An autoregressive (AR) process of order p is

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \ldots + \phi_p Y_{t-p} + \varepsilon_t \quad \text{or} \quad \Phi(B) Y_t = \varepsilon_t$$

where ε_t is Gaussian white noise with zero mean and constant variance and $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$.

For an AR process to be stationary, the roots of $\Phi(B)=0$ should lie outside the unit circle.

2.2.3 Moving average (MA) models

The moving average (MA) process of order q is

$$Y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \theta_3\varepsilon_{t-3} - \ldots - \theta_q\varepsilon_{t-q} \quad \text{or} \quad \Theta(B)\varepsilon_t = Y_t$$

where $\Theta(B)=1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q$.

Although no restrictions on the root of $\Theta(B)=0$ are required for an MA process to be stationary. It is generally desirable to impose the restrictions that the root of $\Theta(B)=0$ should lie outside the unit circle to ensure that the process satisfies a condition called *invertibility*. This condition assures that in an MA process, the influence of the remote past has very little influence on the present. This is always true for an AR process.

2.2.4 Mixed autoregressive - moving average (ARMA) models

ARMA process of order p and q is

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q} \quad \text{or}$$

$$\Phi(B)Y_t = \Theta(B)\varepsilon_t$$

where Φ, Θ and ε_t are as above.

For an ARMA process to be stationary, we require that the roots of $\Phi(B)=0$ should lie outside the unit circle.

2.3 Non-stationary (ARIMA) models

Most economic time series are non-stationary. A first step in practical modeling is to transform the observed series to a stationary series. Transformations to stabilize variance, familiar from standard regression analysis, are supplemented in time series analysis with time dependent transformations call *differencing.* A time series exhibiting a constant drift in linear trend may be transformed to a stationary series by taking first difference, $W_t = Y_t - Y_{t-1} = \nabla Y_t$.

Higher order differences remove polynomial trends; for example, the second order (or twice repeated) difference $W_t = \nabla^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$ removes a second order drift in trend. When a series is differenced in this way the original undifferenced series is called *integrated* and is referred to as an **ARIMA** model i.e. autoregressive integrated moving average.

2.4 Model identification

The process of model identification determines

(i) The order of differencing needed to reduce an ARIMA model to stationary.

(ii) The order of the resulting MA and/or AR parameters θ and ϕ.

2.4.1 Determining Non-stationarity by unit root test

The unit root tests are methods for detecting the presence of a unit root in parametric time series models. Such tests can be helpful to determine stationarity of time series under Box and Jenkins models since an AR or ARMA process is stationary if and only if all the roots of $\Phi(B)=0$ are outside the unit circle. There are several unit root tests. This thesis uses the test by Said and Dickey (1984) with the null hypothesis H_0 : One of the roots in an AR or an ARMA process is unity. Given time series Y_t, the $\hat{\tau}$ statistics suggest by Said and Dickey (1984) is

$$\hat{\tau}=\left[\hat{V}\left\{\hat{\theta}_1\right\}\right]^{-\frac{1}{2}}\left(\hat{\theta}_1-1\right),$$

$$\hat{\tau}\xrightarrow{\mathcal{L}}\frac{\frac{1}{2}\left(W^2(1)-1\right)}{\int_{t=0}^{1}W^2(t)\,dt},$$

where $\hat{\theta}_1$ is the regression coefficient by regressing Y_t on Y_{t-1}, $V\left\{\hat{\theta}_1\right\}$ is ordinary least square estimator of the variance of $\hat{\theta}_1$ and $W(t)$ is a standard Brownian motion. In this thesis the test is conducted at the .05 level of significance and the MA's order are determined in the next section.

2.4.2 Determining the order of an ARMA process by SCAN method

There are many ways to determine the order of an ARMA process. In this thesis, the smallest canonical correlation (SCAN) method proposed by Tsay and Tiao (1985) is used. This method was chosen because it is computationally cheap. The SCAN method can be described as follow.

Consider a stationary ARMA(p,q) model as describe in section 2.2.4. Let k be the AR order and i be the MA order, $k,i \in I_{0,\infty}$. Defined a $(k+1)\times(k+1)$ matrix $A(k,i)$ given by

$$A(k,i) = \Sigma^{-1}(k+1,0)\Sigma(k+1,i+1)^{T}\,\Sigma^{-1}(k+1,0)\Sigma(k+1,i+1).$$

where

$$\Sigma(k,i) = \begin{pmatrix} \sigma(i) & \sigma(i-1) & \cdots & \sigma(i-k+1) \\ \sigma(i+1) & \sigma(i) & \cdots & \sigma(i-k+2) \\ \vdots & \vdots & \cdots & \vdots \\ \sigma(i+k-1) & \sigma(i+k-2) & \cdots & \sigma(i) \end{pmatrix},$$

and

$$\sigma(j) = \operatorname{cov}(Y_t, Y_{t+j}), \quad j = 0,\ \pm 1,\ \pm 2,\ \ldots.$$

Denote the smallest eigenvalue of $A(k,i)$ by $\mu(k,i)$. Tsay and Tiao have shown that for $k = p, p+1, \ldots$ and $i = q, q+1, \ldots$ we have $\mu(k,i) = 0$, otherwise $\mu(k,i)$ is nonzero. Therefore the orders of an ARMA process can be determined by finding the maximum values of k and i (denote by k^{max} and i^{max} respectively) that makes $\mu(k,i)$ nonzero. Hence the AR and MA orders of an ARMA process under the SCAN method are $k^{max}+1$ and $i^{max}+1$.

2.5 Trading Stock price under Box and Jenkins method

The trading strategy simulation is summarised as follows.

1. Start with an initial amount (cash balance) of \$1000 and n the number of observations equal to 200.
2. Analyse the first n observations of each stock using Box and Jenkins approach.
3. Predict the price for the $(n+1)^{th}$ observation.
4. If the predicted price in step 3 is less than the price for n^{th} observation and the cash balance is more than 0, buy the stock at the $(n+1)^{th}$ observation price using all cash available.
5. If the predicted price in step 3 is more than price for n^{th} observation and the number of stocks held is more than 0, sell all of the stock at the $(n+1)^{th}$ observation price.
6. If the situation does not satisfy conditions 4 or 5 do not trade.
7. Increase n by 1 and go to step 2.

This trading strategy is chosen because the data are stock closing prices, which are available at the end of the trading day. Therefore by predicting tomorrow's prices we can only buy, sell or not trade at closing tomorrow. The strategy is only a simple one because it is not the intention of this thesis to research into trading strategies, which can be very complex. This strategy is compared with the adjusted stock price which can be describe as buying a thousand dollars worth of stocks at the first trading day and never trade again so that the worth will go up and down according to the stock prices. This is consistent with the commonly held view of the Martingale hypothesis of stock prices (see section 5.2). In this view, the expected price tomorrow is the same as today's value so that the above strategy will lead to not trading at all but holding on to the initial stock (or initial capital).

The tables 2.1 and 2.2 give the summary statistics of the results and figures 2.1 to 2.5 show the actual results against time for the trading strategy above. Note that the results are for either cash balance or the amount of stocks held in dollars noted here as "worth" and in figures 2.1 to 2.5 worth is compared with adjusted stock price.

The annual returns under Box and Jenkins method are 9.43, 1.73, 9.32, 2.03 and 16.46 percents for Air New Zealand, Brierley, Carter Holt Harvey, Lion Nathan and Telecom respectively while the adjusted stock prices gives 6.13, -10.56, -0.86, -0.42, 20.73 percents respectively. Thus, the results show that the Box and Jenkins method performed slightly better than the adjusted stock price in the long run except for the Telecom stock where the adjusted price performance is better.

Stock	Mean	Standard Deviation	Maximum	Minimum	Final Cash Balance
AIRVA	1489.25	297.72	2201.66	985.51	1912.70
BRY	1156.54	196.24	1541.49	560.33	1131.04
CAH	1414.02	257.51	1964.03	741.11	1899.76
LNN	1254.40	148.24	1516.47	941.34	1155.46
TEL	2376.75	797.43	3858.27	1000.00	2996.06

Table 2.1 Summary statistics of the trading simulation results (worth).

Stock	Mean	Standard Deviation	Max.	Min.	Final Cash Balance
AIRVA	1606.09	371.00	2281.77	784.53	1535.91
BRY	1122.45	303.32	1541.67	312.50	447.92
CAH	1199.52	263.89	1664.00	460.00	940.00
LNN	867.68	128.82	1210.05	598.17	970.32
TEL	3065.42	1059.54	5000.00	994.90	3882.65

Table 2.2 Summary statistics of adjusted stock price.

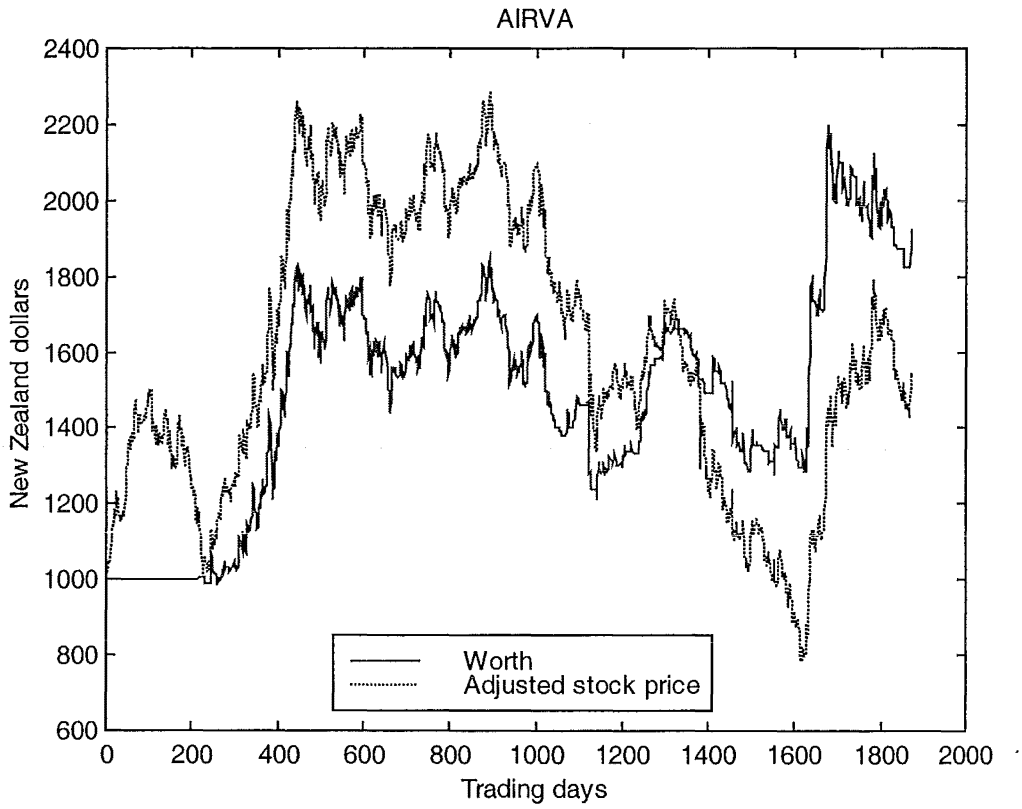

Figure 2.1 Simulation result for Air New Zealand Ltd. "A" (AIRVA)

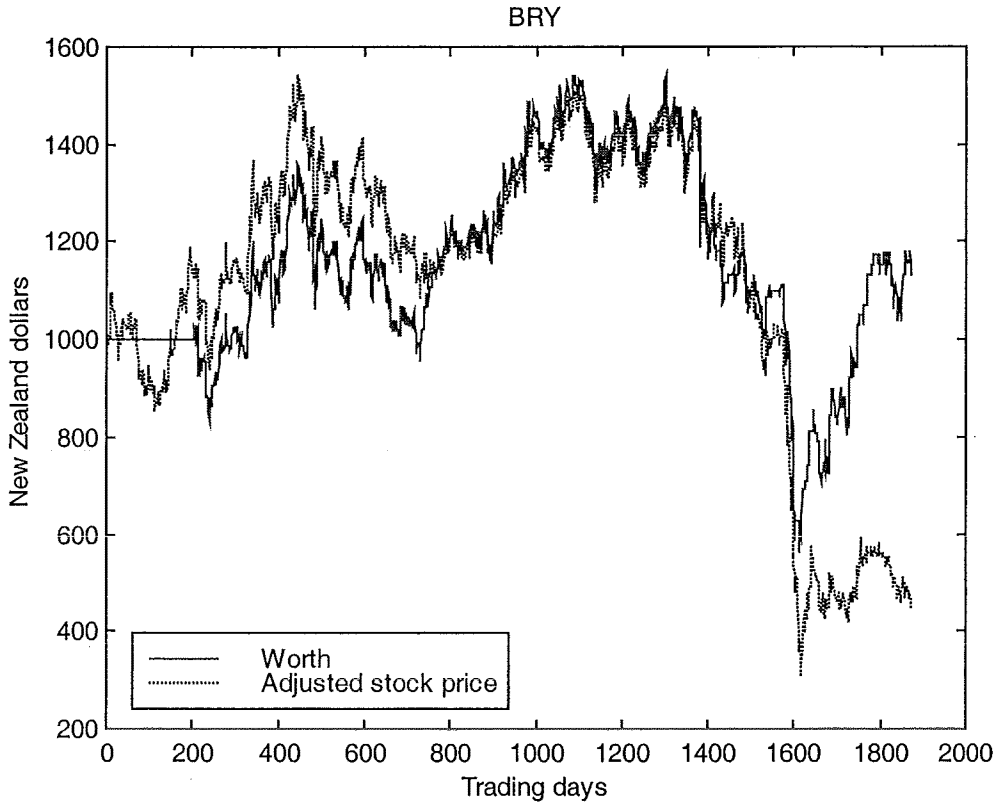

Figure 2.2 Simulation result for Brierley Investments Ltd. (BRY)

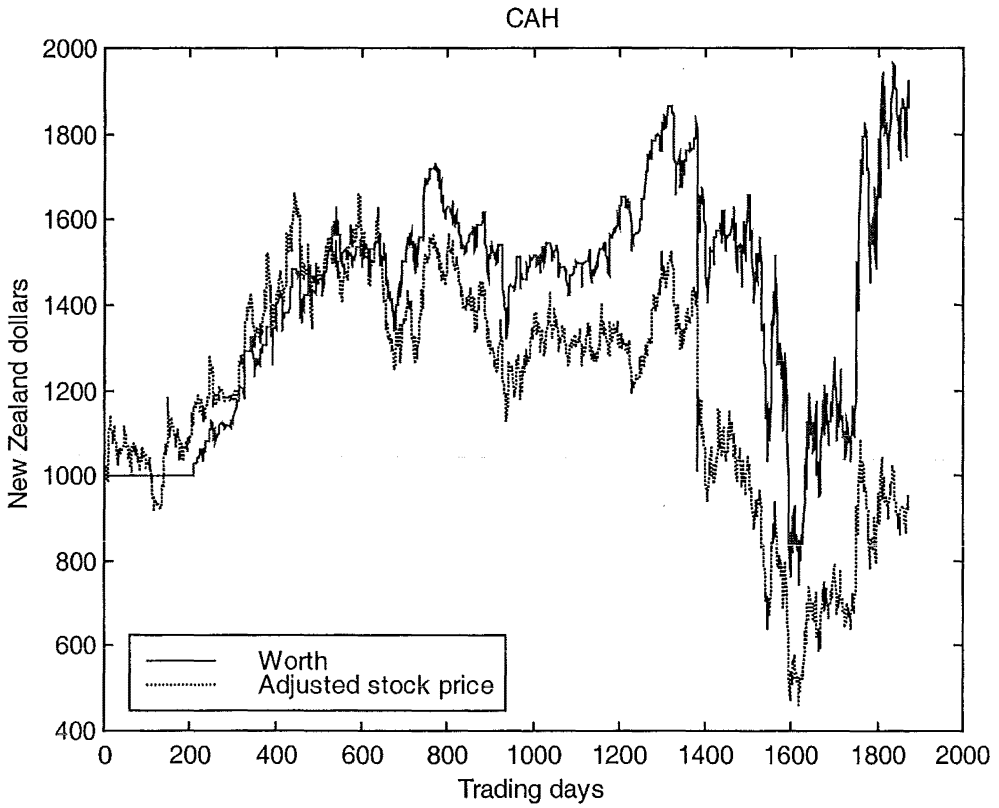

Figure 2.3 Simulation result for Carter Holt Harvey Ltd. (CAH)

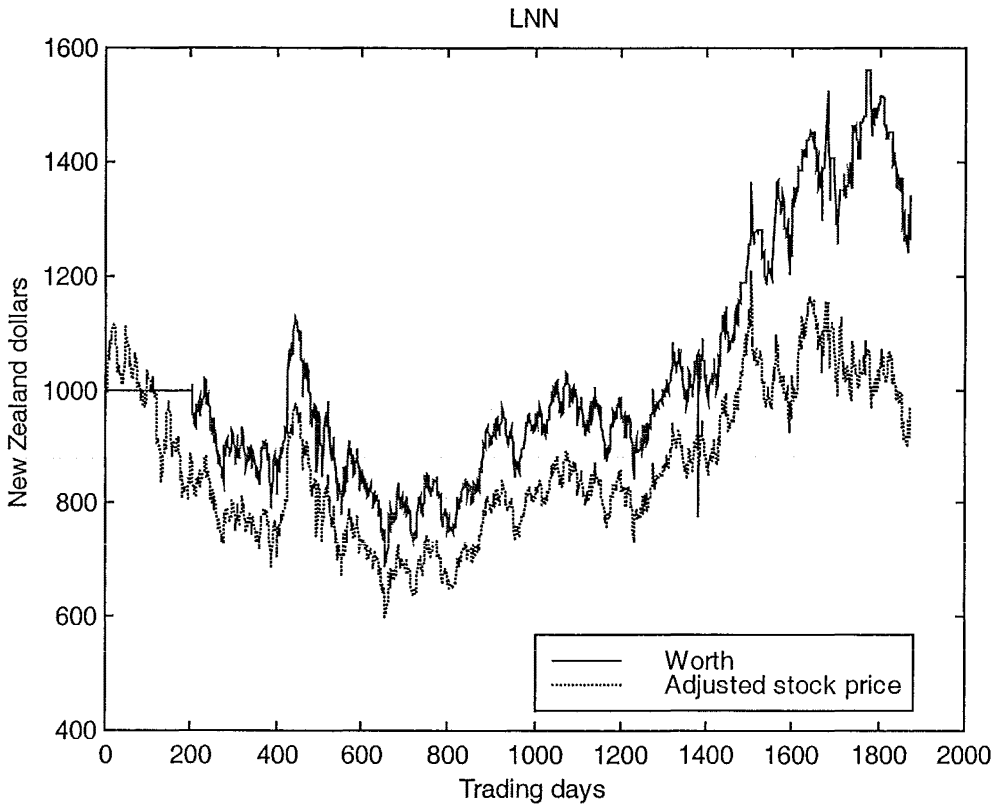

Figure 2.4 Simulation result for Lion Nathan Ltd. (LNN)

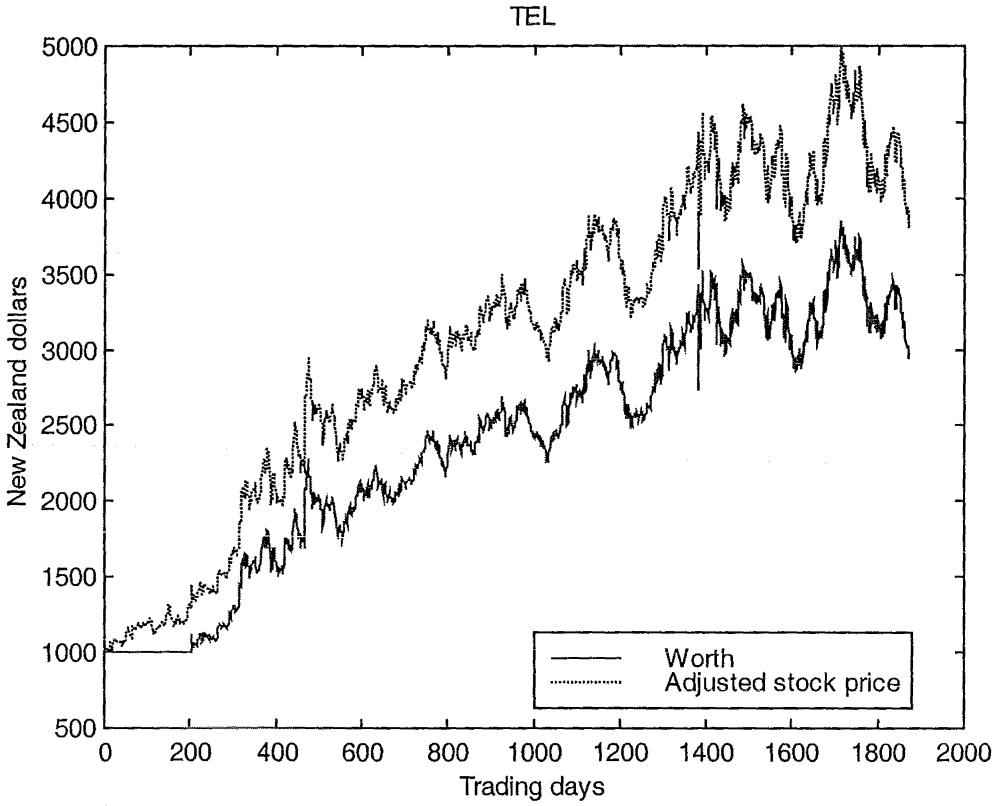

Figure 2.5 Simulation result for Telecom Corporation of New Zealand Ltd (TEL)

Chapter 3

3 Bayesian Dynamic linear model

3.1 Introduction

In the 1800s the Danish astronomer T. N. Thiele developed a recursive procedure (a procedure where the data are processed sequentially rather than in batch form) resembling what is now referred to as the *Kalman filter*, for the problem of determining the distance from Copenhagen to Lund (Thiele, 1880). Further discussion on the work of Thiele is given in Hald (1981) and Lauritzen (1981). A *dynamic linear model* introduced by Kalman (1960) and Kalman and Bucy (1961) is based upon such a concept. The signal x_t is assumed to evolve according to the linear equation

$$x_t = F_{t-1}x_{t-1} + \varepsilon_t \qquad for\ t = 1, 2, \ldots ,$$

where F_t denotes the *state transition* matrix and ε_t denotes a white noise.

The Bayesian interpretations of dynamic linear models were given by a number of authors, including Ho and Lee (1964) and Jazwinski (1970). But the first comprehensive Bayesian treatment of these models in the statistical literature was by Harrison and Stevens (1976).

This chapter reviews some parts of Bayesian dynamic models that applies to this thesis on forecasting stock market. Section 2 explains the basic dynamic linear model form and the notation used where any extension models can be built upon this basic model. Section 3 shows how the model evolves using Baye's theorem and the concepts of prior and posterior information. Section 4 reviews component forms such as trend and seasonal effect, which can be included in the basic model. Section 5 shows how different component forms can be combined into one model through the block structured model. Section 6 reviews variance learning where the model variance is unknown, as is the case in forecasting stock price. Section 7 looks at stock trading simulation under Bayesian dynamic linear model. These reviews are based mainly on West and Harrison (1989) and Pole, West and Harrison (1994).

3.2 Model Form and Notation

Let Y_t be an observation on the time series over times $t = 1, 2, \ldots$, then the dynamic linear model can be written as

$$Y_t = F_t'\theta_t + v_t, \qquad v_t \sim N[0, V_t], \tag{3.1}$$

$$\theta_t = G_t\theta_{t-1} + w_t, \qquad w_t \sim N[0, W_t]. \tag{3.2}$$

where F_t is the vector of known values of independent variables (the *regression vector*), θ_t is the state or system vector thus $F_t'\theta_t$ is the mean response or level, v_t is a Gaussian white noise process with variance V_t, G_t is a matrix of known coefficients that defines the evolution of the state vector θ_t, and w_t is a Gaussian white noise process with variance W_t.

Equation (3.1) is the *observation equation* for the model, defining the sampling distribution for Y_t conditional on the quantity θ_t. The conditional independence structure applies. So, given θ_t, Y_t is independent of all the other observations and parameter values; and in general, given the present, the future is independent of the past. The observation equation relates Y_t to θ_t via a dynamic linear regression with a Guassian white noise error structure, through time varying variance vector V_t.

Equation (3.2) is the *evolution, state* or *system equation*, defining the time evolution of the state vector. The conditional independence property shows a one-step Markov evolution so that, given θ_{t-1} and the known values of G_t and W_t, θ_t is independent of I_{t-1}. That is, given θ_{t-1}, the distribution of θ_t is fully determined independently of values of Y_{t-1} and all the state vectors and observations prior to time $t-1$. The deterministic component of the evolution is the transition from state θ_{t-1} to $G_t\theta_{t-1}$, a simple linear transformation of θ_{t-1}. The evolution is completed with the addition of the random vector w_t.

The two Guassian white noise series $\{\nu_t\}$ and $\{w_t\}$ are assumed to be temporally independent and mutually independent; that is, the covariances $Cov[\nu_t,\nu_s]$, $Cov[w_t,w_s]$ for all $t \neq s$, and $Cov[\nu_t,w_s]$ for all t,s are zero. These assumptions are made for convenience rather than necessity since models with correlated stochastic terms can always be transformed into the set-up used here. We use this independence form because the analysis is thereby most easily derived and presented.

As is common in this type of work, the same notation is used for θ and ϕ (where ϕ are used in section 3.6 onward) when it is a random variable and when it is the actual value.

3.3 Prior to Posterior Analysis

Bayesian learning proceeds by combining information from observations expressed through the likelihood function based on the prior distribution – that is the investigator's existing state of knowledge before the observations are made – to obtain the posterior distribution. The mechanism of obtaining the posterior is the well-known Bayes' theorem. These interactions could be illustrated with a flow chart as shown in Figure 3.1 below.

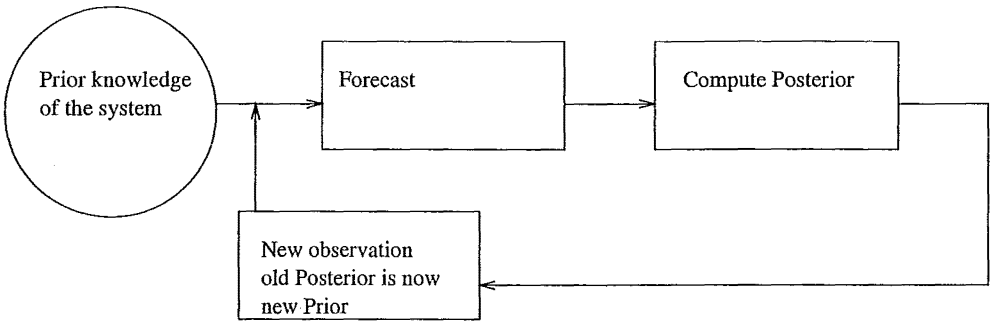

Figure 3.1 Prior to Posterior analysis

3.3.1 Prior Information

Suppose the prior distribution of the state vector for time t is assumed to be a normal distribution then it can be shown at any time t also to be normal with mean a_t and covariance R_t (see sections 3.3.2 to 3.3.5) as given below,

$$\theta_t \mid I_{t-1} \sim N[a_t, R_t] \tag{3.3}$$

Generally the univariate dynamic linear model will start with initial prior $(\theta_0 \mid I_{-1}) \sim N[a_0, R_0]$, for some assumed prior mean a_0 and variance R_0.

3.3.2 Forecasts

From the prior information, forecasts are generated using the equation (3.1). The forecast quantity Y_t is a linear combination of normally distributed variables, $\theta_t \mid I_{t-1}$ and ν_t, and is therefore also normally distributed. The forecast mean f_t and variance Q_t are obtained as below.

$$\begin{aligned} E[Y_t \mid I_{t-1}] &= E\left[\left(F_{t-1}'\theta_t + \nu_t\right) \mid I_{t-1}\right] \\ &= E\left[F_{t-1}'\theta_t \mid I_{t-1}\right] + E[\nu_t \mid I_{t-1}] \\ &= F_{t-1}'E[\theta_t \mid I_{t-1}] + E[\nu_t] \\ &= F_t'a_t \\ &= f_t, \end{aligned} \tag{3.4}$$

$$
\begin{aligned}
V[Y_t \mid I_{t-1}] &= V\left[\left(F_{t-1}'\theta_t + \nu_t\right) \mid I_{t-1}\right] \\
&= V\left[F_{t-1}'\theta_t \mid I_{t-1}\right] + V\left[\nu_t \mid I_{t-1}\right] \\
&= F_t' V[\theta_t \mid I_{t-1}] F_t + V[\nu_t] \\
&= F_t' R_t F_t + V_t \\
&= Q_t .
\end{aligned}
\tag{3.5}
$$

The forecast distribution for one step ahead therefore has the normal form $Y_t \mid I_{t-1} \sim N[f_t, Q_t]$. In deriving (3.5) we have assumed that the observation disturbance ν_t, is not correlated with the state, θ_t.

3.3.3 Likelihood

The model likelihood, a function of the model parameters, is the conditional forecast distribution evaluated at the observed value. It has the normal form

$$
\begin{aligned}
L(\theta_t \mid Y_t = y_t, V_t) &\propto p(Y_t = y_t \mid \theta_t, V_t) \\
&\sim N\left[F_t' \theta_t, V_t\right] .
\end{aligned}
\tag{3.6}
$$

3.3.4 Posterior Information

The prior information is combined with information in the observation (the likelihood) using Bayes' Theorem to yield the posterior distribution on the state,

$$p(\theta_t \mid I_{t-1}, y_t) = \frac{p(Y_t = y_t \mid \theta_t, V_t)\, p(\theta_t \mid I_{t-1})}{p(Y_t = y_t)}.$$

Notice that the denominator $p(Y_t = y_t)$ is not a function of the state θ_t and serves merely as a density normalizing constant. It is therefore typically ignored (since it is easily recovered when the scaled density is required). Thus Bayesian updating is expressed as a proportional form,

$$posterior \propto likelihood \times prior .$$

Therefore the state posterior density function is the product of two normal density functions as just seen in equations (3.6) and (3.3) respectively, yielding another normal density function.

$$\begin{aligned} p(\theta_t \mid I_{t-1}, y_t) &\propto \exp\left\{-\frac{1}{2}V_t^{-1}\left(y_t - F_t'\theta_t\right)^2\right\} \times \exp\left\{-\frac{1}{2}(\theta_t - a_t)'R_t^{-1}(\theta_t - a_t)\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\theta_t - m_t)'C_t^{-1}(\theta_t - m_t)\right\}, \end{aligned}$$

The mean m_t and variance C_t are obtained as

$$\begin{aligned} m_t &= E[\theta_t \mid I_{t-1}] = a_t + A_t e_t , \\ C_t &= V[\theta_t \mid I_{t-1}] = R_t - A_t A_t' Q_t , \\ A_t &= \frac{R_t F_t}{Q_t} , \\ e_t &= y_t - f_t . \end{aligned}$$

The posterior mean is adjusted from the prior value by a multiple of the one step ahead forecast error. The amount of that adjustment is determined by the quantity A_t, which is the regression matrix of the state vector θ_t on the observation Y_t conditional upon the history I_{t-1}. This regression matrix, or *adaptive factor* as it is called, is determined by the relative size of the state prior variance and the observation variance. This means that the larger the observation variance compared with the state prior variance, the smaller will be the adaptive factor. Therefore the underlying state estimate does not require substantial movement. Conversely, if the state prior variance is large compared with the observation variance, then the observation has a lot of relevant information for the state and adjustment from the prior to the posterior should properly reflect that position.

3.3.5 The Evolution Process

After the posterior distribution is calculated at time t, we move to time $t+1$. Thus given the posterior distribution for the state at time t as normally distributed with mean m_t and variance C_t, direct application of the system evolution equation leads to the prior for time $t+1$. Once again a linear combination of normally distributed quantities yields a normal distribution,

$$\theta_{t+1} \mid I_t \sim N[a_{t+1}, R_{t+1}],$$

where the moments are defined by

$$a_{t+1} = G_{t+1} m_t \ ,$$
$$R_{t+1} = G_{t+1} C_t G'_{t+1} + W_{t+1} \ .$$

We have now completed the cycle of prior to forecast to posterior to next prior. These stages characterize the routine on-line updating analysis of the model.

3.4 Component Forms

Many time series exhibit distinctive patterns such as trend and seasonal effects. In dynamic linear models we call these patterns "*component*". Such component forms can be included into the structure of the model. Here we outline the mathematical structure of those forms.

3.4.1 Polynomial Trend Components

The simplest trend model is the first order polynomial trend, or level. Observed series values are assumed to be stochastically distributed about a time varying value:

$$Y_t = \mu_t + v_t \ ,$$
$$\mu_t = \mu_{t-1} + w_t .$$

The system equation defines the level to be a simple random walk through time. We can identify this system in terms of defining dynamic linear models where the regression vector and system evolution matrix are constants,

$$F_t = 1, \quad G_t = 1.$$

A second order polynomial trend model allows for systematic growth or decline in level. The additional parameter quantifies the time-to-time change in level,

$$\begin{aligned} Y_t &= \mu_t + v_t, \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + w_{1t}, \\ \beta_t &= \beta_{t-1} + w_{2t}. \end{aligned}$$

The state vector comprises two elements, $\theta_t = (\mu_t, \beta_t)'$, the first representing the current level and the second representing the current rate of change in the level. Identifying the regression vector and system matrix is again straightforward,

$$F_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad G_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Higher order polynomial time trend models may be defined by generalizing and extending the first and second order models described here. Practically speaking, such higher order trends are seldom used.

3.4.2 Seasonal Components

Modeling seasonal patterns in time series requires a component form that is periodic. The most natural representation of such patterns is the seasonal factor

model, where a different parameter – a factor – is defined for each point in a cycle. Here is the seasonal factor used in this thesis.

Let *g(t)* be any real-valued function defined on the non-negative integers $t = 0, 1, \ldots$, where t is the time index. Note that we specifically take the function to be defined from time zero in order to conform with the usage of cyclical forecast function in dynamic linear models, as detailed below.

Definition 3.4.1

(1) $g(t)$ *is* ***cyclical*** *or* ***periodic*** *if, for some integer* $p \geq 1$, $g(t + np) = g(t)$ *for all integer* $t \geq 0$ and $n \geq 0$.

(2) Unless otherwise stated, the smallest integer p such that this is true is called the ***period*** *of g(.).*

(3) g(.) exhibits a single full ***cycle*** *in any time interval containing p consecutive time points, such as* $(t, t + p - 1)$, *for any* $t > 0$.

(4) The ***seasonal factor*** *of g(.) are the p values taken in any full cycle* $\psi_j = g(j), (j = 0, \ldots, p-1)$. *Notice that, for* $t > 0$, $g(t) = g(j)$ *where j is remainder after division of t by p, denoted by* $j = p \mid t$.

(5) *The* ***seasonal factor vector*** *at time t is simply the vector of seasonal factors permuted so that the first element is that for time t, namely*

$\psi_t = (\psi_j, \psi_{j+1}, \ldots, \psi_{p-1}, \psi_0, \ldots, \psi_{j-1})'$, *when the current seasonal factors factor is* ψ_j. *In particular, for any integer n and k = np,*

$\psi_k = (\psi_0, \ldots, \psi_{p-1})'$.

(6) In any cycle, the time point corresponding to the relevant seasonal factor ψ_j *is given a label* $M(j)$. *This label then defines the timing within each cycle, as, for example, with months within years where M(0) may be January, M(1) February, and so forth. Clearly the labels are cyclic with period p, and* $j = p \mid t$ *if, and only if,* $M(j) = t$.

The p seasonal factors describe a general pattern of period p; they may take arbitrary real values. For this reason, the seasonal pattern is termed *form-free.*

Definition 3.4.2

The $p \times 1$ *vector* E_p *and* $p \times p$ *permutation matrix P are defined as*

$$E_p = (1, 0, \ldots, 0)' ,$$

and

$$P = \begin{pmatrix} 0 & 1 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & . & . & . & 0 \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ 0 & 0 & 0 & . & . & . & 1 \\ 1 & 0 & 0 & . & . & . & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ 1 & \mathbf{0}' \end{pmatrix}.$$

It is clear to see that P is p-cyclic, satisfying the equation

$$P^{k+np} = P^k, \qquad (k = 1,\ldots, p),$$

for any $n \geq 0$. In particular, $P^{np} = \mathrm{I}$ for any such n. At anytime t, the current value of $g(.)$ is just ψ_j given by

$$g(t) = E_p' \psi_t \tag{3.7}$$

where $j = p \,|\, t$.

Using the permutation matrix, it is also clear that the seasonal factors rotate according to

$$\psi_t = P\psi_{t-1} \tag{3.8}$$

for all $t \geq 0$.

This relationship provides the initial step in constructing a dynamic linear model for pure seasonality. Suppose the desired forecast function is cyclical

with period p. Equations (3.7) and (3.8) imply that the forecast function has the form of that in a Time Series dynamic linear model with regression vector $F = E_p$ and system matrix $G = P$.

3.5 Block Structured Models

Component forms for trend, seasonality, and regression are the building blocks for constructing models of complex time series behavior. The linear additive structure of the dynamic linear model enables component models to be brought together in a straightforward manner. For example, a linear growth trend plus seasonal factor period two is a dynamic linear model with a regression vector and system matrix defined by

$$\theta_t = \begin{pmatrix} \theta_T \\ \theta_S \end{pmatrix}_t,$$

$$F_t = \begin{pmatrix} F_T \\ F_S \end{pmatrix}_t = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix},$$

$$G_t = \begin{pmatrix} G_T & 0 \\ 0 & G_S \end{pmatrix}_t = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Note that the regression vector simply stacks the individual regression vectors on top of each other; similarly the system state stacks up the component states. The system matrix is structured block diagonally. (Naturally, for all three quantities the stacking order must be the same, although the particular order

chosen does not matter.) In this way, models with any number of components may be constructed.

3.5.1 Block Discounting

Discounting is a practical solution to the problem of setting evolving disturbance variances. The state prior variance at any time is computed as a function of the posterior variance at the previous time determined by a discount factor, $\delta \in (0,1]$. The discount factor represents the amount of information loss attributed to temporal advancement,

$$V[\theta_t \mid I_{t-1}] = \delta^{-1} G V[\theta_{t-1} \mid I_{t-1}] G'.$$

Discounting variances this way is equivalent to setting the evolution variance as a proportion of the posterior variance,

$$W_t = \left(\frac{1}{\delta} - 1\right) G C_{t-1} G',$$

recalling that C_{t-1} denotes the posterior variance $V[\theta_{t-1} \mid I_{t-1}]$. In a model with multiple components, the recommend discount strategy is to proceed component by component. Separate discount factors are specified for each component and individual component evolution variance matrices computed. The overall state evolution variance matrix is then set to the block diagonal composition of these individual elements.

To illustrate, the trend plus seasonal example (see section 3.5) has two components. Defining δ_T and δ_S to be the discount factors for theses components, structure the evolution covariance matrix as

$$W_t = \begin{pmatrix} W_T & 0 \\ 0 & W_S \end{pmatrix}_t$$
$$= \begin{pmatrix} \left(\frac{1}{\delta_T} - 1\right) G_T C_{T\,t-1} G_T' & 0 \\ 0 & \left(\frac{1}{\delta_S} - 1\right) G_S C_{S\,t-1} G_S' \end{pmatrix},$$

where $C_{T\,t-1}$ is the posterior covariance (sub)matrix for the trend component at time (*t-1*), and $C_{S\,t-1}$ is similarly the posterior covariance for the seasonal component.

3.6 Variance Learning

The analysis detailed thus far has assumed a known observation variance sequence $\{V_t\}$. In most applications including forecasting the stock price the observation variance is not known sufficiently well to approximate it by a fixed value. A learning mechanism is necessary. A tractable analytical solution exists for this problem, an extension of the normal-gamma conjugate analysis for the standard linear model.

Working in terms of the precision, $\phi = V^{-1}$, define the constant unknown variance dynamic linear model (compare with 3.1 and 3.2) as

$$\text{Observation Equation:} \quad Y_t = F_t'\theta_t + \nu_t, \qquad \nu_t \sim N\left[0, \phi^{-1}\right],$$
$$\text{System Equation:} \quad \theta_t = G_t\theta_{t-1} + w_t, \qquad w_t \sim N\left[0, W_t^*\phi^{-1}\right].$$

The scaling of the system disturbance covariance w_t by the unknown observation variance is necessary for a conjugate analysis. It is just a scale factor and clearly we can set $W_t = W_t^*\phi^{-1}$ to recover the normal form of the system equation. In practice the discount formulation is used to determine W_t so the mathematical detail here need not be cause for concern.

3.6.1 Prior Information

At time t the prior information extends over both the state and the scale parameter,

$$\theta_t \mid I_{t-1}, \phi \sim N\left[a_t, R_t^*\phi^{-1}\right],$$
$$\phi \mid I_{t-1} \sim GAMMA\left[\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right].$$

Conditional on the scale, the prior information and the model specialise to the normal dynamic linear model already detailed (taking $R_t = R_t^*\phi^{-1}$). The parameters of the gamma prior on the scale represent the degrees of freedom, n_{t-1}, and sums of squared errors, d_{t-1}, with mean equal to the ratio of these

quantities, $\frac{n_{t-1}}{d_{t-1}}$. The significance of these definitions will become clear later in the analysis.

3.6.2 Forecasts

The conditional forecast distribution (the predictive distribution) is the normal form as seen in section 3.3.2, now with the scale made explicit,

$$Y_t \mid (I_{t-1}, \phi) \sim N\left[F_t' a_t, \phi^{-1} Q_t^*\right],$$

where the observation scale free forecast variance is

$$Q_t^* = 1 + F_t' R_t^* F_t.$$

From standard normal-gamma theory, unconditionally the forecast has a t distribution on n_{t-1} degrees of freedom,

$$Y_t \mid I_{t-1} \sim t_{n_{t-1}}[f_t, Q_t],$$

where the mean and scale parameter are given in familiar forms,

$$f_t = F_t' a_t,$$
$$Q_t = S_{t-1} + F_t' R_t^* S_{t-1} F_t.$$

Notice the similarity of this result with the known variance case. The point forecast is the same in both cases. The forecast scale has the same algebraic form with the exception that the unknown variance is estimated by its prior expected value, $S_{t-1} = \frac{d_{t-1}}{n_{t-1}}$.

3.6.3 Posterior Information

Obtaining the posterior distribution on the scale parameter is an easy application of Bayes' theorem since we can eliminate the state parameter by using the conditional predictive distribution just described and the marginal prior on the scale:

$$
\begin{aligned}
p(\phi \mid I_{t-1}, y_t) &\propto p(y_t \mid I_{t-1}, \phi)\, p(\phi \mid I_{t-1}) \\
&\propto \left(\phi^{-1} Q_t^*\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\phi^{-1} Q_t^*\right)^{-\frac{1}{2}} (y_t - f_t)^2\right\} \times \phi^{\frac{n_{t-1}}{2}-1} \exp\left\{-\frac{\phi}{2} d_{t-1}\right\} \\
&\propto \phi^{\frac{n_{t-1}-1}{2}} \exp\left\{-\frac{\phi}{2}\left[\frac{e_t^2}{Q_t^*} + d_{t-1}\right]\right\} \\
&\sim GAMMA\left[\frac{n_t}{2}, \frac{d_t}{2}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
n_t &= n_{t-1} + 1, \\
d_t &= d_{t-1} + \frac{e_t^2}{Q_t^*}.
\end{aligned}
$$

The scale posterior has the same probability distribution as the prior, a gamma distribution, but with updated parameters. The degrees of freedom parameter increases by one - an additional piece of information has been processed - and the sums of squares term is incremented by the square of the (scaled) forecast error. The posterior point estimate of the scale is just the ratio of these quantities, the sums of squares divided by the degrees of freedom, $S_t = \frac{d_t}{n_t}$.

The posterior for the state is most easily obtained by considering the joint distribution with the observation. Both conditional distributions, $(Y_t \mid I_{t-1}, \phi)$ and $(\theta_t \mid I_{t-1}, \phi)$, are normal, so their joint distribution is also normal, the covariance being easily identified from the observation equation

$$\left.\begin{matrix} Y_t \\ \theta_t \end{matrix}\right| I_{t-1}, \phi \sim N\left[\begin{pmatrix} f_t \\ a_t \end{pmatrix}, \begin{pmatrix} \phi^{-1}Q_t^* & F_t' R_t^* \phi^{-1} \\ R_t^* F_t \phi^{-1} & R_t^* \phi^{-1} \end{pmatrix}\right].$$

Using properties of the multivariate normal distribution, conditioning the state on the observed value yields a normal distribution, which is directly identified as

$$(\theta_t \mid I_{t-1}, Y_t = y_t, \phi) \sim N\left[m_t, C_t^* \phi^{-1}\right],$$

where the moments are updated from the prior values as already seen for the normal analysis, the difference here being that the scale conditioning is made explicit,

$$
\begin{aligned}
m_t &= a_t + R_t^* F_t \phi^{-1} \left(\phi^{-1} Q_t^*\right)^{-1} (y_t - f_t) \\
&= a_t + \frac{R_t^* F_t e_t}{Q_t^*},
\end{aligned}
$$

$$
\begin{aligned}
C_t^* \phi^{-1} &= R_t^* \phi^{-1} - R_t^* F_t \phi^{-1} \left(\phi^{-1} Q_t^*\right)^{-1} F_t' R_t^* \phi^{-1} \\
&= R_t^* \phi^{-1} - \frac{R_t^* F_t F_t' R_t^* \phi^{-1}}{Q_t^*}.
\end{aligned}
$$

The marginal posterior state distribution - removing the conditioning on the scale parameter - is obtained by integrating over ϕ in the joint distribution,

$$
\begin{aligned}
p(\theta_t \mid I_t) &= \int p(\theta_t, \phi \mid I_t)\, d\phi \\
&= \int p(\theta_t \mid I_t, \phi) p(\phi \mid I_t)\, d\phi \\
&\propto \int \left|C_t^* \phi^{-1}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta_t - m_t)' \left[C_t^* \phi^{-1}\right]^{-1} (\theta_t - m_t)\right\} \phi^{\frac{n_t}{2}-1} \exp\left\{-\frac{\phi d_t}{2}\right\} d\phi \\
&\int \phi^{\frac{(n_t+k)}{2}-1} \exp\left\{-\frac{\phi}{2}\left[(\theta_t - m_t)' C_t^{*-1} (\theta_t - m_t) + d_t\right]\right\} d\phi
\end{aligned}
$$

(where k is the dimension of the state vector, θ.) Writing $S_t = \frac{d_t}{n_t}$ (the point estimate of the scale parameter) the integral reduces to

$$
\left[n_t + (\theta_t - m_t)' \left(C_t^* S_t\right)^{-1} (\theta_t - m_t)\right]^{-\frac{(n_t+k)}{2}}.
$$

This is the form of a multivariate Student-t distribution (in k dimensions) on n_t degrees of freedom with mean m_t and scale $C_t = C_t^* S_t$.

3.6.4 Variance Discounting

The dynamic linear model stresses dynamic modeling through time, the stochastic element of the system equation providing for parametric change. Supporting arguments in favor of dynamic change apply equally strongly to the scale parameter as to the system state. The variance learning model analysed above does not allow for such a dynamic, the scale is explicitly determined as an unknown constant. That analysis may be extended to permit stochastic change in the scale using the information discounting strategy.

Posterior information on the scale at time $(t-1)$ is described by a gamma distribution,

$$\phi_{t-1} \mid I_{t-1} \sim GAMMA\left[\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right].$$

For the variance learning analysis to remain tractable the prior information on the scale for time t must stay within the gamma family of distributions. But while the distributional form cannot be changed, the parameterization may be altered as desired.

The system equation specifies a formal model for stochastic evolution of the state, but that is extremely difficult to do for the scale. The reason lies in the functional form of the gamma distribution - it does not have the convenient mathematical properties the normal distribution exhibits. However, a formal model of scale evolution is not strictly necessary. What matters is to have the

time t prior information suitably adjusted from the time $(t-1)$ posterior information so as to encompass temporal uncertainty. Discounting the precision is a natural way of achieving this aim as seen with the state. Applying the idea to the scale parameter leads to discounting both parameters of the gamma distribution. For a variance component discount factor δ define the prior information on the scale as

$$\phi_t \mid I_{t-1} \sim GAMMA\left[\frac{\delta n_{t-1}}{2}, \frac{\delta d_{t-1}}{2}\right].$$

The mean of the prior is preserved at the posterior value since it is the ratio of the two parameters, $\left(\frac{\delta n_{t-1}}{2}\right) / \left(\frac{\delta d_{t-1}}{2}\right) = \left(\frac{n_{t-1}}{2}\right) / \left(\frac{d_{t-1}}{2}\right)$, while the variance is inflated by a factor of δ^{-1}, just as was done for the state.

Unconditional forecast, state prior and posterior distributions are simple modifications of the undiscounted case,

$$\begin{aligned} Y_t \mid I_{t-1} &\sim t_{\bar{n}_t}[f_t, Q_t], \\ \theta_t \mid I_{t-1} &\sim t_{\bar{n}_t}[a_t, R_t], \\ \theta_t \mid I_t &\sim t_{\bar{n}_t}[m_t, C_t], \end{aligned}$$

where

$$
\begin{aligned}
f_t &= F_t' a_t \ , \\
Q_t &= \overline{S}_t + F_t' R_t F_t \ , \\
m_t &= a_t + \frac{R_t F_t e_t}{Q_t} \ , \\
C_t &= \left(\frac{S_t}{\overline{S}_t} \right) \left[R_t - R_t F_t F_t' R_t \right] / Q_t \\
&= \left(\frac{S_t}{\overline{S}_t} \right) R_t \left[I - F_t A_t' \right] \\
&= \left(\frac{S_t}{\overline{S}_t} \right) \left[R_t - A_t A_t' Q_t \right] \ , \\
\overline{n}_t &= \delta n_{t-1} \ , \\
\overline{d}_t &= \delta d_{t-1} \ , \\
\overline{S}_t &= \frac{\overline{n}_t}{\overline{d}_t} = S_{t-1} \ .
\end{aligned}
$$

There are four basic points of departure of this unknown variance analysis from the known variance analysis. In the forecast variance Q_t, the prior point estimate $\overline{S}_t$ replaces the previously known value V_t. The state posterior covariance is resealed at each time to reflect the latest information on the scale parameter (the factor $\frac{S_t}{\overline{S}_t}$ in the equation for C_t). The marginal distributions are now Student-*t* instead of normal. And of course there is the probabilistic description of the scale itself. The scale posterior is clearly the same form as in the no-discount model but with different parameter values,

$$
\phi_t \mid I_t \sim G\left[\frac{n_t}{2}, \ \frac{d_t}{2} \right] ,
$$

where

$$n_t = \overline{n}_t + 1$$
$$d_t = \overline{d}_t + \frac{S_{t-1}e_t^2}{Q_t}.$$

3.7 Trading Stock price under Bayesian dynamic linear model

The stock prices are analysed by block-structured model, which contains a second order polynomial trend. Also the variance is initially unknown thus require variance learning and discounting. Seasonality is not assumed in these models and for this purpose p is set to 1. However the seasonality component is left in order to make it easy to incorporate seasonality if necessary. The reason of not using a seasonal effect here is partly to keep the model simple and in any case stock prices do not often have a long-term seasonal effect.

The discount parameters δ for trend and variance are 0.9, 0.99 for all stocks. The prior and initial parameters of the model are shown in the table 3.1. Note that these parameters are automatically selected by the Bayesian Analysis of Time series (BATS) computer package develop by Andy Pole (contact ap@stat.duke.edu). Hence BATS is used here to analysed stocks under the Bayesian dynamic linear models. The author of BATS does not wish to share his method of selecting the prior since he recently used BATS in the industry.

Stock			AIRVA	BRY	CAH	LNN	TEL
θ_t	μ_t	a_0	50.49	121.10	269.00	41.78	147.50
		R_0^*	10.27	92.39	92.39	66.73	51.33
	β_t	a_0	0.3552	-0.0345	-0.6966	0.1286	0.5241
		R_0^*	0.0278	2.50	0.2502	0.1807	0.1390
	ψ_t	a_0	0	0	0	0	0
		R_0^*	0	0	0	0	0
d_0			0.0054	0.4825	0.0483	0.0348	0.0268
n_0			1	1	1	1	1

Table 3.1 The prior and initial parameters for each stock

The trading strategy simulation is summarized as follows.

1. Start with an initial amount (cash balance) of \$1000 and n the number of observations equal to 1.
2. Analyze each stock and predict the price for the $n^{th}+1$ observation.
3. If the predicted price in step 3 is less than the price for n^{th} observation and the cash balance is more than 0, buy the stock at the $n^{th}+1$ observation price using all cash available.
4. If the predicted price in step 3 is more than price for n^{th} observation and the number of stocks held is more than 0, sell all of the stock at the $n^{th}+1$ observation price.
5. If the situation does not satisfy conditions 4 or 5 do not trade.
6. Increase n by 1 and go to step 2.

The use of this strategy is discussed at the end of section 2 (see section 2.5). The following table 3.2, 3.3 and figures 3.2 to 3.6 are the results of the trading strategy above. Note that the results are for either cash balance or the amount of stocks held in dollars noted here as "worth" and in figures 3.1 to 3.5 worth is compared with the adjusted stock price.

The annual returns under Bayesian dynamic linear models are 9.02, -8.99, 1.99, 2.29 and 22.42 percents for Air New Zealand, Brierley, Carter Holt Harvey, Lion Nathan and Telecom respectively while the adjusted stock prices gives 6.13, -10.56, -0.86, -0.42, 20.73 percents respectively. Thus, the results show that the Bayesian dynamic linear model performed better than the adjusted stock price even though trading Brierley share under the given model and strategies is not profitable.

Stock	Mean	Standard Deviation	Maximum	Minimum	Final Cash Balance
AIRVA	1368.39	198.50	1945.95	924.49	1861.94
BRY	1166.02	315.02	1727.63	477.56	507.30
CAH	1352.99	257.46	1886.45	759.41	1152.10
LNN	1049.08	69.73	1293.49	874.46	1077.41
TEL	3080.40	1135.37	5506.86	1000.00	4290.89

Table 3.2 Summary statistics of the trading simulation results (worth).

Stock	Mean	Standard Deviation	Max.	Min.	Final Cash Balance
AIRVA	1606.09	371.00	2281.77	784.53	1535.91
BRY	1122.45	303.32	1541.67	312.50	447.92
CAH	1199.52	263.89	1664.00	460.00	940.00
LNN	867.68	128.82	1210.05	598.17	970.32
TEL	3065.42	1059.54	5000.00	994.90	3882.65

Table 3.2 Summary statistics of adjusted stock price.

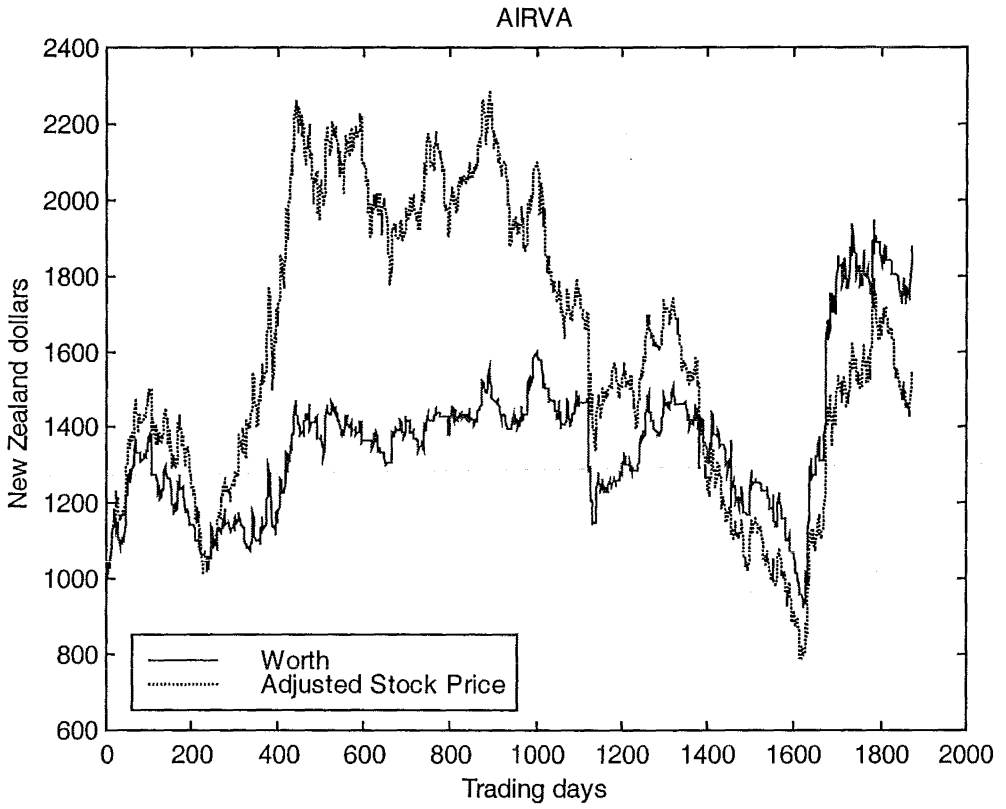

Figure 3.2 Simulation result for Air New Zealand Ltd. "A" (AIRVA)

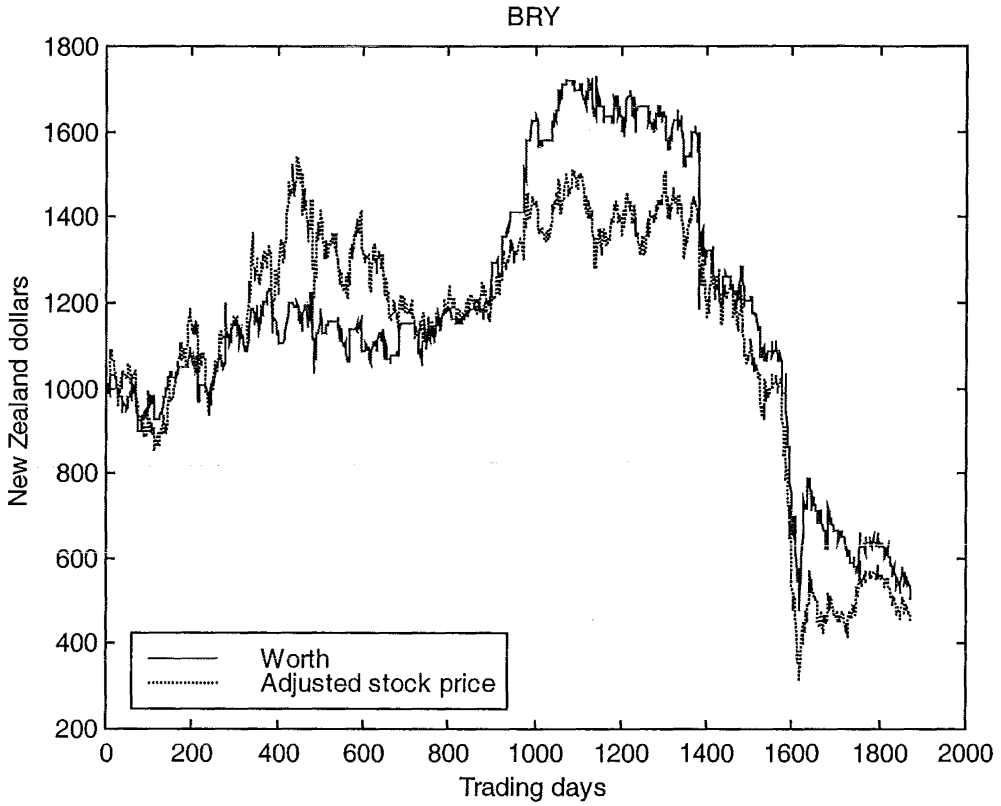

Figure 3.3 Simulation result for Brierley Investments Ltd. (BRY)

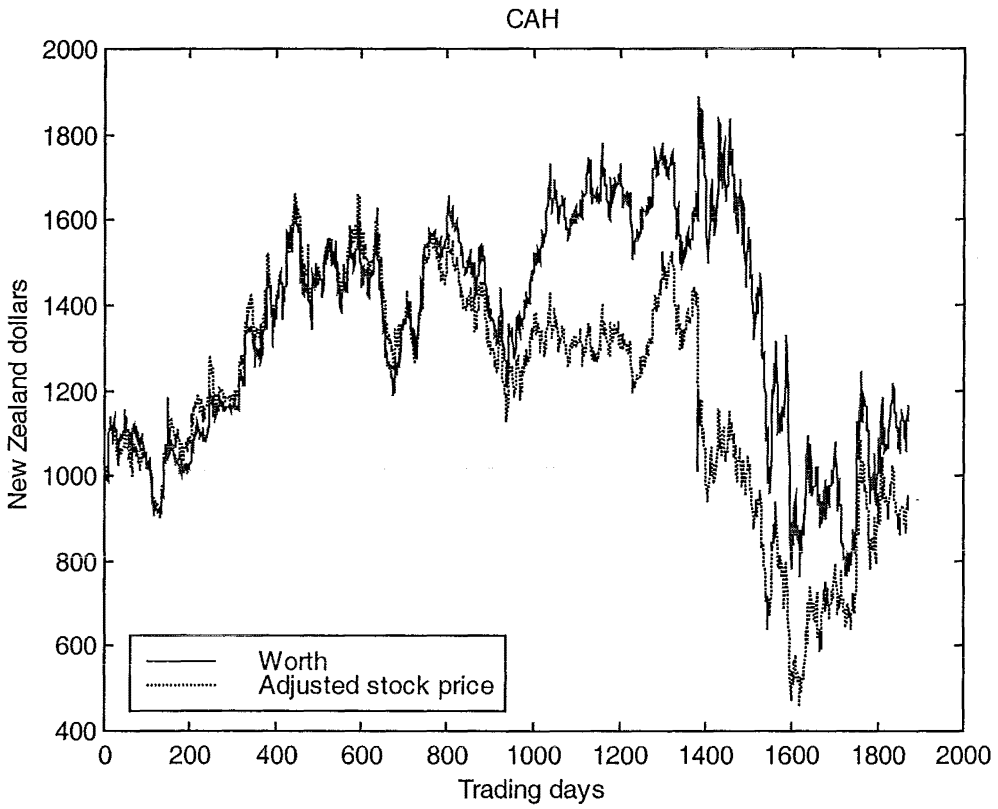

Figure 3.4 Simulation result for Carter Holt Harvey Ltd. (CAH)

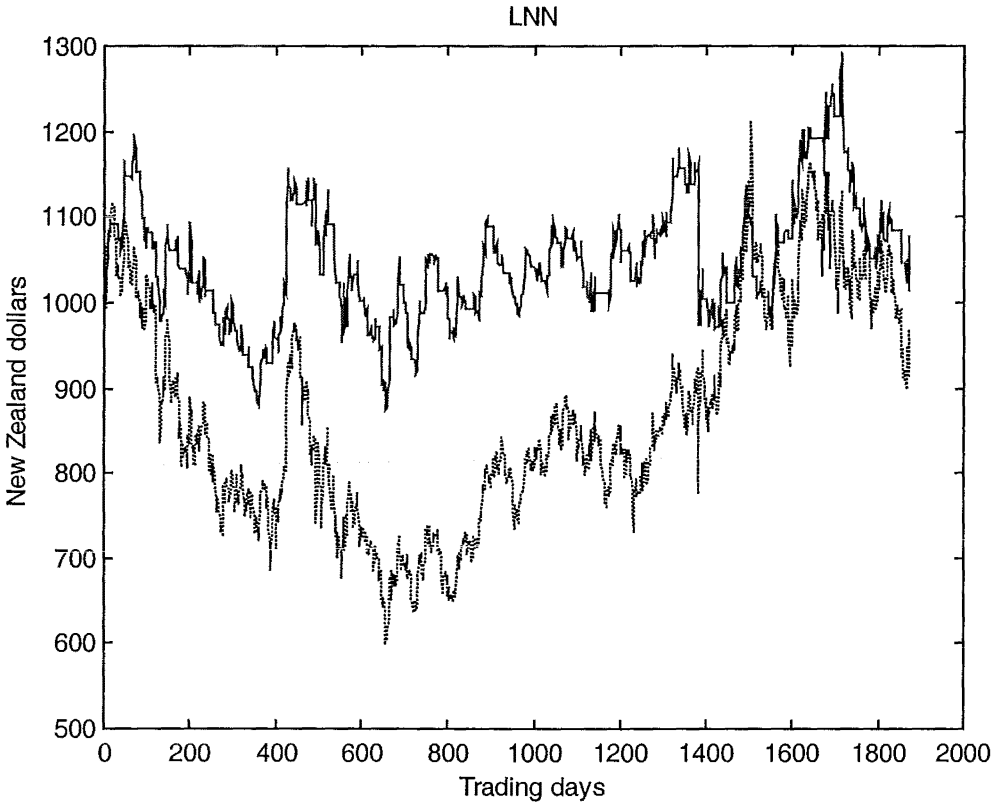

Figure 3.5 Simulation result for Lion Nathan Ltd. (LNN)

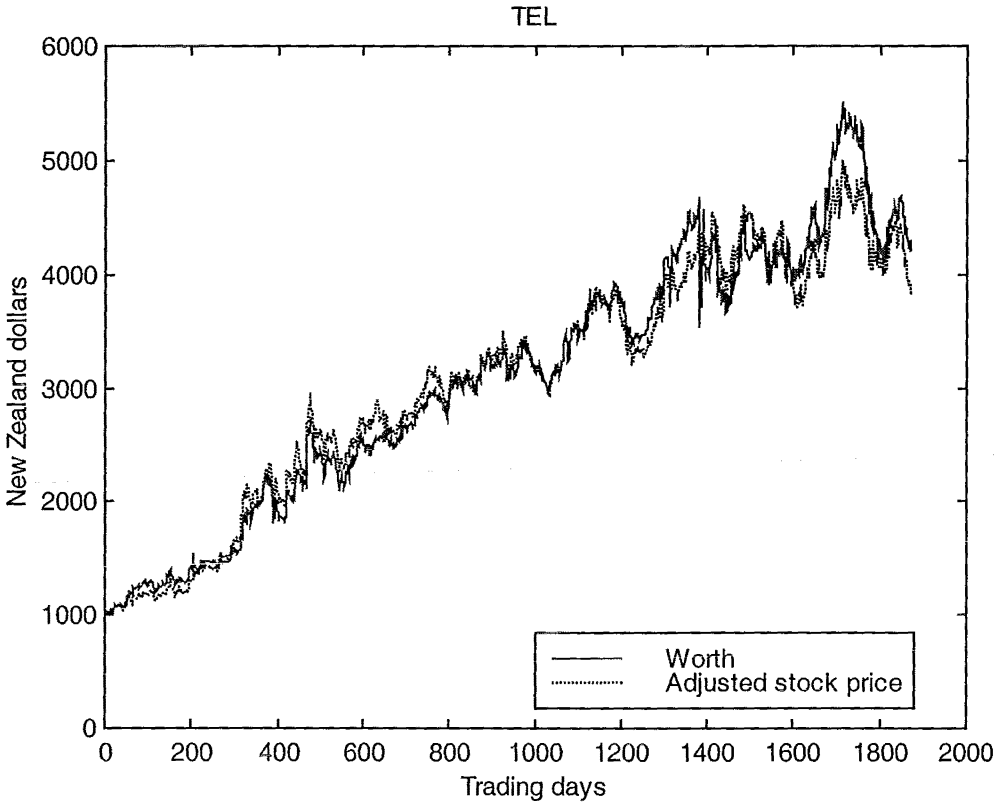

Figure 3.6 Simulation result for Telecom Corporation of New Zealand Ltd. (TEL)

Chapter 4

4 Fuzzy based neural networks

4.1 Introduction

Fuzzy neural networks are connecting links between fuzzy logic and neural networks. The goal of this connection is to combine the advantages of both techniques in order to process uncertain or vague information. Since the fuzzy neural networks are an integration of both fuzzy logic and neural networks, we will introduce here the basic concepts of fuzzy logic in section 2, and neural networks and fuzzy neural networks in section 3. Section 4 shows how fuzzy neural networks is applied to forecast stock prices. Section 5 looks at stock trading simulation using fuzzy neural networks.

4.2 Basics of Fuzzy Logic

L. Zadeh who published his first work on fuzzy sets in 1965 introduced the theory of fuzzy logic. The basic idea of fuzzy logic is to allow not only the values 1 and 0 corresponding to *true* and *false*, but instead the whole interval

[0,1] as degrees of truth. This leads to a radical extension of classical logic. Zadeh was not the first to introduce a multi-valued logic calculus. In the twenties a Polish mathematician J. Lukasiewicz (1920) had already developed multi-valued logic calculus.

Although critics continually state that all important problems can be dealt with by classical means, use of fuzzy logic is becoming more and more wide spread. In particular this is indicated by the success of fuzzy logic in the discipline of control techniques. In the beginning of the seventies, Zadeh (1972) introduced the concept of fuzzy logic control. In recent years, it has been shown that with special kind of fuzzy controllers each continuous function on a compact set could be approximated to any degree of accuracy.

4.2.1 Fuzzy Sets

At first, we will have a look at fuzzy sets. We obtain fuzzy sets by extending the membership predicate "$\in$ " to the interval [0, 1] instead of only using the classical truth-values 0 and 1. This means that a set can contain points with a certain *degree*. This degree of membership can be considered in different ways. On the one hand, the membership grade can be interpreted as grade of probability, on the other hand it can be regarded as grade of possibility. In

any case fuzzy set theory allows us to attach uncertainty with data and we use this to advantage in neural networks.

Definition 4.2.1 *Let X be a (classical) set (called* universe*) and* μ_A *a (classical) function:*

$$\mu_A : X \to [0,1].$$

Then we call the set of ordered pairs

$$A = \{(x, \mu_A(x)) \mid x \in X\}$$

a fuzzy set *on X ;* μ_A *is called* the membership function.

Hence we get a fuzzy set by extending the characteristic function of a classical (crisp) set to the interval [0,1]. This extension enables membership grades to have values between 0 and 1. A point x with membership grade $\mu_A(x) = \frac{1}{2}$ would only be "half" in a (classical interpreted) set A. Thus classical sets are also fuzzy sets, but their membership functions have only values in $\{0,1\}$.

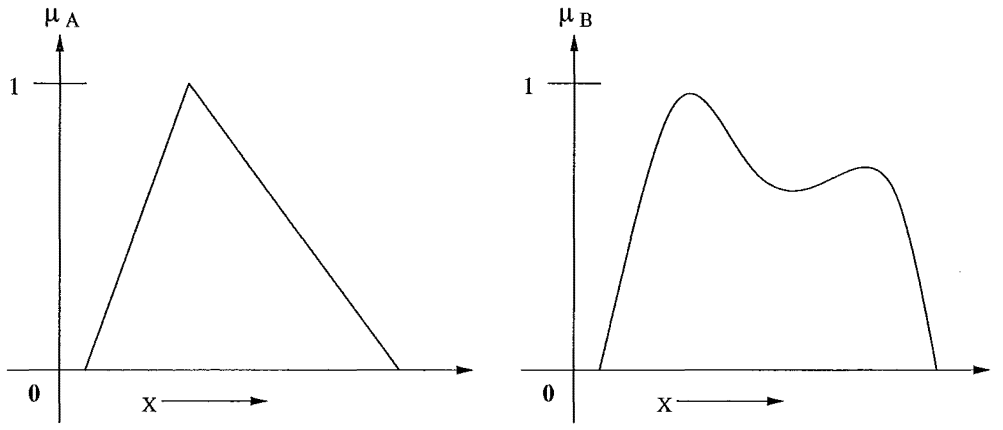

Figure 4.1 Membership functions of two fuzzy sets A and B

Definition 4.2.2 *The support of a fuzzy set A is the crisp set of all* $x \in X$ *such that* $\mu_A(x) \geq 0$

Definition 4.2.3 *The (crisp) set of all* $x \in X$ *that belong to the fuzzy set A at least with the degree* α *is called the* α -level set: $A_\alpha = \{x \in X \mid \mu_A(x) \geq \alpha\}$.

$A_\alpha^* = \{x \in X \mid \mu_A(x) > \alpha\}$ *is called* the strong α -level set.

Classical subset-hood can be extended to fuzzy sets, too. We define:

Definition 4.2.4 *Let A and B be fuzzy sets on the universe X. Then A is said to be a fuzzy subset of B* $(A \subset B)$ *if* $\mu_A(x) \leq \mu_B(x)$ *for all* $x \in X$.

If we restrict the possible membership grades to the values 0 and 1, this definition meets the definition of crisp subset-hood.

Next, we introduce the term fuzzy power set:

Definition 4.2.5 *Let X denote a universe and let A be a fuzzy set on X. Then* $\tilde{P}(A) := \{B | B \text{ is fuzzy set on } X \wedge \mu_B(x) \leq \mu_A(x) \forall\, x \in X\}$ *is called* the fuzzy power set *of A.*

$\tilde{P}(A)$ is the set of all fuzzy subsets of A. If we identify the universe X with a fuzzy set whose membership values are 1, we can interpret $\tilde{P}(X)$ as the set of all fuzzy sets on X.

4.2.2 Fuzzy Numbers

Fuzzy numbers are a certain kind of fuzzy sets on $\mathbb{R}$ (real numbers) with particular properties. Intuitively everybody knows the concept of fuzzy numbers from everyday life. When we say, "It's about twenty degrees Celsius warm" this means that there is a high probability for the temperature

to be twenty degrees Celsius. However, it is also possible (although less probable) that the temperature is about 19 or 21 degrees.

Fuzzy numbers are an appropriate tool for modeling values attached with uncertainty or inaccuracy. The fuzzy power set of $\mathbb{R}$ is denoted by $\tilde{P}(\mathbb{R})$. We are interested in a particular type of fuzzy numbers, which we call triangular fuzzy numbers. For this work, it is sufficient to know that the membership functions of triangular fuzzy numbers are triangular-shaped (see Figure 4.1) and therefore can be represented by only three real values: the modal value and the left and right "fuzziness". The notation for a triangular fuzzy number is as follows: $A=(m,\alpha,\beta)_{trian}$, where $\alpha,\beta>0$. α and β are called the *left* and *right fuzziness*, respectively. $A.\lambda = m-\alpha$ is called the *left* and $A.\rho = m+\beta$ as the right edge of A.

Definition 4.2.6 *A fuzzy number A is said to be triangular if:*

$$\mu_A(x)=\begin{cases} 0 & , \ x\le m-\alpha \\ \dfrac{x-m+\alpha}{\alpha} & , \ m-\alpha<x\le m \\ \dfrac{m-x+\beta}{\beta} & , \ m<x\le m+\beta \end{cases}$$

The set of all triangular fuzzy numbers on $\mathbb{R}$ is called $\text{Trian}(\mathbb{R})$.

By declaring triangular fuzzy numbers we have created an analogy to "traditional" numbers. In contrast to traditional (crisp) numbers, fuzzy numbers exhibit fuzziness, which is described by linear *reference functions,* which decrease monotonically from the modal value to the left and to the right hand. Now we need operations on fuzzy numbers, which are analogous to the operations we have for crisp numbers. One possible way of achieving this aim is using the extension principle, which allows us to fuzzify classical concepts. Note that a crisp number can be treated as a fuzzy number for example $x=(x,0,0)_{trian}$ where $x\in\mathbb{R}$.

4.2.3 The Extension Principle

Definition 4.2.7 (Extension Principle by Zadeh) *Let* $X = X_1 \times X_2 \times ... \times X_n$ *be the non-empty Cartesian product of the universes* $X_1,...,X_n$ *and* $A_1,...,A_n$ *be fuzzy sets on* $X_1,...,X_n$. *Let* $f : X \to Y$ *be a (classical) function from X to a universe Y. Then the fuzzy set B defined by*

$$B = \{(y, \mu_B(y)) \mid y = f(x_1,..., x_n), (x_1,..., x_n) \in X\}$$

where

$$\mu_B(y) = \sup_{(x_1,...,x_n) \in f^{-1}(y)} \min\{\mu_{A1}(x_1), ..., \mu_{An}(x_n)\} \quad , \quad f^{-1}(y) \neq 0$$

is said to be the f-extension of $A_1,..., A_n$. $f^{-1}(y)$ *denotes the set of antecedents* $f^{-1}(\{y\})$ *of y.*

Notation: $B = f(A_1,..., A_n)$.

The extension principle allows us to apply classical functions directly to fuzzy sets by achieving the membership grade of a crisp function value from the membership grades of the values of the set of its antecedents. For each value in the set of antecedents, the minimum of the membership grades of its components is computed. The desired membership grades of the images is the supremum of these minima. The great advantage of the extension principle is the possibility of extending crisp mathematical concepts to fuzzy sets.

In the following section, we describe the extension of the operations which are used in a neural network to operations on triangular fuzzy numbers. This generalisation is based on the extension principle. Unfortunately the use of triangular fuzzy numbers causes the loss of some algebraic properties of classical numbers. On the other hand this restriction is very important for decreasing the computational expense within a fuzzified neural network because all operations on triangular fuzzy numbers can be implemented by operations on three real numbers only. However triangular fuzzy numbers do not have a group structure relative to addition or multiplication, and the product of two triangular fuzzy numbers is no longer triangular. For this reason we will, when necessary, approximate all of the obtained membership functions by triangular linear functions in order to restrict all operations to triangular fuzzy functions.

4.2.4 Fuzzification of the Basic Operations

We next define the addition and the multiplication of triangular fuzzy numbers by means of the extension principle.

Definition 4.2.8 *Let* $A = (m_1, \alpha_1, \beta_1)_{trian}$ *and* $B = (m_2, \alpha_2, \beta_2)_{trian}$ *be two triangular fuzzy numbers. Then the triangular fuzzy number* $C =$

$(m_1 + m_2, \alpha_1 + \alpha_2, \beta_1 + \beta_2)_{trian}$ *is called the fuzzy sum of A and B. Notation: C* $= A \oplus B$.

We obtain this definition directly by applying the extension principle to the classical function $f(x, y) = x + y$ and no approximation is needed.

Using the extension principle in order to fuzzify the multiplication is not sufficient

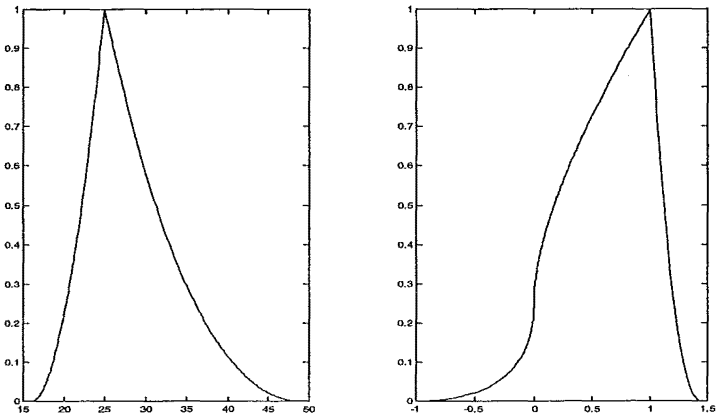

Figure 4.2 Various products of triangular fuzzy numbers

because the product C = $A \otimes B$ of two triangular fuzzy numbers has no longer linear but curved reference functions (see Figure 4.2).

In order to decrease the computational expense within neural networks we simply substitute the obtained membership functions by linear functions. This corresponds to a linear approximation with three supporting points. Hence, we define:

Definition 4.2.9 *Let* $A = (m_1, \alpha_1, \beta_1)_{trian}$ *and* $B = (m_2, \alpha_2, \beta_2)_{trian}$ *be two triangular fuzzy numbers. Then we call the triangular fuzzy number C*

$= (m, \alpha, \beta)_{trian}$ *the approximated fuzzy product of A and B. The values m,* α *and* β *are computed by:*

$$
\begin{aligned}
m &= m_1 \cdot m_2, \\
\alpha &= m - \min\{A.\lambda \cdot B.\lambda,\ A.\lambda \cdot B.\rho,\ A.\rho \cdot B.\lambda,\ A.\rho \cdot B.\rho\}, \\
\beta &= \max\{A.\lambda \cdot B.\lambda,\ A.\lambda \cdot B.\rho,\ A.\rho \cdot B.\lambda,\ A.\rho \cdot B.\rho\} - m.
\end{aligned}
$$

Notation: $C = A \tilde{\otimes} B$.

We get this definition directly by applying the extension principle to the classical function $f(x, y) = x \cdot y$ and approximating the obtained membership functions by a linear function (detail in Laarhoven and Pedrycz (1983)).

Arithmetic on triangular fuzzy numbers is very much like interval arithmetic. To compute a new modal value we only need the modal values of the affected triangular fuzzy numbers, whereas the new edges are obtained directly by interval arithmetic.

The fuzzy addition and the approximated fuzzy multiplication defined above are commutative and associative as a result of the extension principle.

Another important property of these operations, which can also be drawn from the extension principle, is that monotonic behaviour is preserved.

We have covered here the necessary concepts of fuzzy sets, which we need for fuzzy neural network. Next we look first at traditional neural networks and then we outline fuzzy neural networks by using the fuzzy operations introduced here.

4.3 Neural Networks

Artificial neural networks, which are based on biological structures, were already investigated in the forties and fifties. Especially well known are the model by McCulloch and Pitts (1943) and the perceptron, which was proposed by Frank Rosenblatt (1961). The negative results occurring during the investigation of networks of perceptrons caused a stagnation of neurocomputing in the sixties and seventies. The most significant problem was the lack of a learning rule for multilayered networks. The backpropagation algorithm, which was developed in the beginning of the seventies, the model introduced by Hopfield (1982) and the investigation of

associative memories in the eighties entailed intensive research in various disciplines up to the present.

The revival of interest in artificial neural networks can also be explained by their versatile applicability. In contrast to traditional computational models there is no need for an algorithm which is especially designed for the actual problem. Instead, the neural network is capable of finding the right configuration far solving the actual problem simply by the use of training data and a learning law. The success of neural models is, in particular, in disciplines where traditional computational models have difficulties, e.g. speech recognition.

4.3.1 Definition of Neural Networks

Definition 4.3.1 *A formal neuron v is a Quadruple (x, w, a, S) which consists of a vector of input variables* $x = (x_1,...,x_n)^T$, *a weight vector* $w = (w_1,...,w_t)^T$, *an integration function a(x, w) and an activation function S(a).*

Definition 4.3.2 *A* neural network *is a directed graph G = (K, E), which consists of the set of nodes* $K = \{v_1,...,v_q \mid v_i \text{ is formal neuron}\}$ (were q is the

total number of nodes in a neural network) *of all formal neurons and the set E of links between all formal neurons.*

The vector $w = (w_1,...,w_n)$ of *all* weights used in the neural net is called the *weight vector* of the net. If $w_i \in X$ for all $i \in \{1, . . ., N\}$ then X^N is called the *weight space* of the net.

An important point of these definitions is the locality of the information processed by the formal neurons. Each neuron gets information only by its incoming connections. It is possible for a formal neuron to possess local memory, but there is no global information interchange. This allows massively parallel information processing.

Formal neurons within a neural network are normally arranged in layers. If the input connections of a layer gain information only from the output of the preceding layer, the net is said to be a *feedforward network:*

Definition 4.3.3 *A neural network of units (sets of formal neurons), which is defined according to definition 4.3.2, is called a* feedforward network, *if the directed graph given by the net is free of cycles.*

The units mentioned in definition 4.3.3 may consist of subnets, which do not necessarily have a feedforward structure, and it should be noted that many authors have such structures. The neural nets used here will be in all cases layer by layer interconnected with feedforward networks.

Figure 4.3 shows the basic structure of a formal neuron and a layer by layer fully connected feedforward network consisting of two input units and three output neurons. The presented network is called two-layered because the input units are not real formal neurons but only transmit input values to the next layer. Therefore, the input links of these units are not weighted. The transfer function (which can be defined by $T(x,w)=S(a(x,w))$) of an input unit is the identity function.

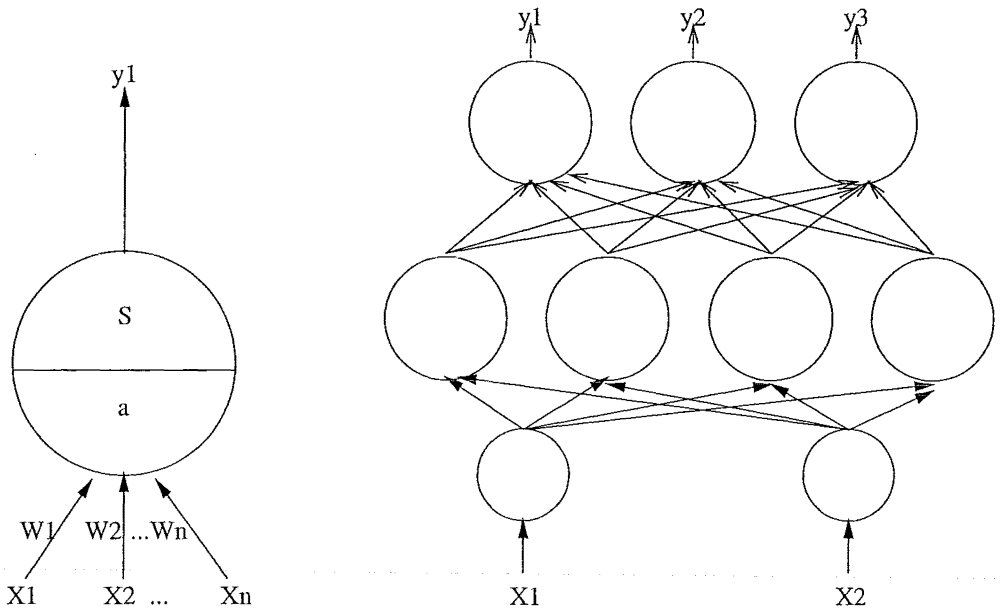

Figure 4.3: Basic structure of a formal neuron (left) and a multilayer feedforward network (right) showing outputs y_i for inputs x_i.

4.3.1.1 Backpropagation Networks

The most well known feedforward net is the backpropagation network. The learning problem of the backpropagation net consists of finding output values for provided input data, which fit the desired output values best. Normally these input/output data represent a function. If one defines an *error function* for such a net, which measures the error in the actual output data, one can regard the learning problem of a backpropagation net as a problem of optimisation. The aim is to minimise the error of the network. For this

reason the error function is regarded as a function on the weight space of the net. A function often used in this context is the sigmoid function.

Definition 4.3.4 *The function* $s_c : \mathbb{R} \to [0, 1]$ *defined by*

$$s_c(x) = \frac{1}{1+e^{-cx}}$$

is called a sigmoid function.

In traditional backpropagation networks the activation function of a formal neuron is the sigmoid function, because it is differentiable on its whole domain. The derivative of the sigmoid can be written by using the sigmoid function itself. This fact makes this function especially suitable for being used within a neural net, which is trained by means of, steepest descent. A backpropagation network can be defined by:

Definition 4.3.5 *A feedforward network, which is defined according to definition 4.3.2, is called a* backpropagation network, *if the integration function of the formal neurons is the weighted sum* $a(x,w) = \sum_{i=1}^{n} x_i w_i$ *and the activation function is the sigmoid function.*

Although the true feedforward network uses the sigmoid activation function many have made exceptions to this rule and in the fuzzy networks used here the logarithmic function (see section 4.4.4) is used as the activation function.

The error function of a backpropagation network is next defined by:

Definition 4.3.6 *The function*

$$E(w) = \frac{1}{2}\sum_{i=1}^{p} \|t_i - y_i\|^2$$

is called the error function on the weight space. In this definition $y_1,\ldots,y_p$ *denote the p output vectors of the neural network, which correspond to the p-input vectors of the training set,* $t_1,\ldots,t_p$ *denote the desired output vectors and w is the weight vector of the neural network.*

The *training set* is a set of given input/output data, which represents the function the net is meant to learn. The error of the net is then computed on the set of these training data:

Definition 4.3.7 *The finite set T = {(x, t)} of pairs (x, t),* $x \in \mathbb{R}^m$ *and* $t \in \mathbb{R}^n$,

on which we compute the error function of a backpropagation network consisting of m input units and n output neurons, is called training set.

In order to check the result of the training process it is necessary to have another set of input/output data of the function, which is to be learned. This set should have an empty intersection with the original training set. This restriction is important because the pairs of training data are just supporting points of an unknown function. If these supporting points are chosen awkwardly the net will perhaps learn something completely different.

Another problem, which can occur during the training of a backpropagation network is called *over-training*. This means that within the training process the approximation error of the net relative to the *training data* decreases, but at a certain point the error relative to data that do not occur in the training set starts to increase again. In these cases another set of data, which is called the *training test set*, can be used to test the performance of the net during the training. However with fuzzy neural networks, this does not tend to be a problem (see section 4.3.2).

4.3.1.2 The Backpropagation Learning Rule

The backpropagation algorithm is based upon the method of steepest descent. The goal is to minimize the error function of the net on the weight space by finding the weight vector w which minimizes $E(w)$. Generally the following learning rule is used:

$$w(t+1) = w(t) - \eta \nabla E .$$

η denotes a given learning rate and ∇E is the gradient of the error function with respect to w. For a single weight w_{ij}^{k} in the k^{th} layer ($k \geq 1$ because the input layer has no weights). We compute the new weights by the following equation:

$$w_{ij}^{k}(t+1) = w_{ij}^{k}(t) - \eta \frac{\partial E}{\partial w_{ij}^{k}(t)} .$$

One learning step of the backpropagation algorithm can be divided into four phases:

- forward pass
- computation of the net error

- backward passing of the error signal
- modification of the weights

where the last three are generally grouped together and called the backward pass.

During the forward pass, the net receives an input datum of the training set. The activation of the net is then computed. After that, the error of the net in comparison with the desired output datum is computed. This error is passed back through the net in order to compute the required weight modifications. These weight modifications are carried out in the last step.

To compute the error of the net according to definition 4.3.6 we have to present all of the data of the training set to the net, because the error is computed on the whole training set. In general these are two possible methods for weight correction.

1. The corrections for each weight are computed and stored. After presenting the last input datum the weight modification is carried out. This variation of the algorithm is called *batch method* or *off-line-training.*

2. If only one pair (x, t) of the training set is used to compute an error value and the weight modification is carried out directly basing only on the error signal of this pair, this is called *on-line-training*. In this case if the training data are chosen uniformly at random the direction of the weight correction oscillates around the direction of the negative gradient. Therefore the

direction roughly follows that of the negative gradient. Hence, this version is also called the *stochastic version*, and this is the method that is used here.

4.3.2 Fuzzy Neural Networks

The definition of a fuzzy neural network used here is very similar to the definition of a backpropagation network. The difference consists in the use of triangular fuzzy numbers instead of real numbers and extended operations instead of classical operations.

Using fuzzy neural networks instead of crisp networks has some advantages:

1. The use of fuzziness allows us to process data provided with uncertainty. In this respect, the class of functions that can be approximated increases.

2. If we think of fuzzy input data as an extension of traditional crisp data we are able to cover the whole input space with the supports of only some training data thus diminishing the risk of over-training, which always exists if the traditional backpropagation learning law is used.

3. Covering the input space enables us to gain information about the reaction of the net on yet unknown input data.

4. The fuzziness of the weights can be interpreted as a kind of "softening" of the error surface. This effect can be used to skip local minima in which the backpropagation learning law may get trapped.

Of course the use of fuzzy neural nets has some disadvantages, too. The computational expense within a fuzzy neural network exceeds the costs of a comparable backpropagation net by far (storing a triangular fuzzy number requires three real values).

4.3.2.1 Definition of Fuzzy Neural Network

As we have fuzzified the $\oplus$ and $\otimes$ operations we have in fact fuzzified the operations needed for the integration function of a formal neuron already, we only have to deal with an applicable extension of the activation function.

Usually the activation function S of the formal neurons in a backpropagation network is the sigmoid function s_c. Therefore we have to extend this function in order to process triangular fuzzy numbers.

Definition 4.3.8 *Let* $A = (m, \alpha, \beta)_{trian}$ *be a triangular fuzzy number. Then we call the function* $\tilde{s}_c$ *defined by*

$\tilde{s}_c(A) = (s_c(m),\ s_c(m) - s_c(m-\alpha),\ s_c(m+\beta) - s_c(m))_{trian}$ *for* $c \in [0, \infty]$ the approximated extension of the sigmoid function.

The definition for $c \in [-\infty, 0]$ can be obtained analogously.

We obtain this definition directly by applying the extension principle to the classical function $f(x) = s_c(x)$ and approximating the obtained reference functions by linear functions as done in section 4.2.4.

Now we can define a fuzzy neural networks as follows:

Definition 4.3.9 *A feedforward network, which is defined according to definition 4.3.2, is called a fuzzy neural network, if the integration function of the formal neurons is the weighted sum* $\tilde{a}(x, w) = \sum_{i=1}^{n} x_i \tilde{\otimes} w_i$ *and the activation function of the formal neurons is the approximated sigmoid function* $\tilde{s}_c$.

The weighted sum in this definition denotes the addition defined according to the extension principle in section 4.2. A neural network, which is defined according to definition 4.3.9 highly resembles a backpropagation net. The difference is that it implements a function from $[\text{Trian}(\mathbb{R})]^m$ to $[\text{Trian}(\mathbb{R})]^n$ instead of implementing a function from $\mathbb{R}^m$ to $\mathbb{R}^n$.

4.4 Stock price simulator using fuzzy neural networks

The particular fuzzy neural network used here, developed by Hobbs and Bourbakis (1995), is designed around the assumption that "the price of a stock should generally change proportionally with (or against) certain market indexes". The neural network, after analyzing these market variables, predicts the true price for a particular stock *S*, and decides whether it is under-priced or over-priced. Although the overall market might swing up one day, and down the next, the neural network model should be right when the market stays constant; the up and down swings should offset each other on the other days. Although only one market variable (the NZSE 40 index) is used in this thesis, the networks can handle more than one market variable. Thus, the following sections will show the general case of the networks where there are more than one market variable.

Figure 4.4 shows the network structure and the following sections gives details of the processes involved in the fuzzy neural network used here.

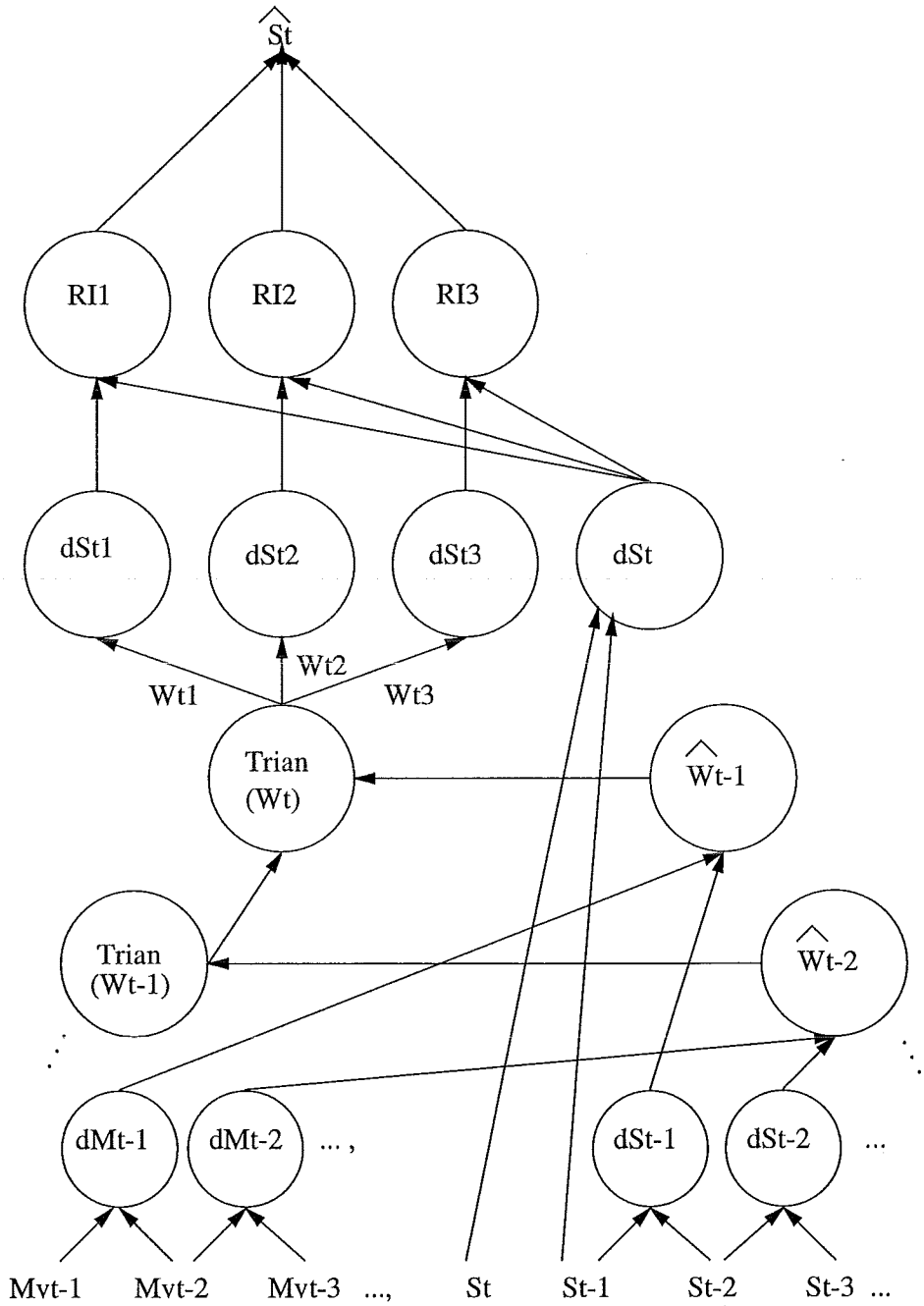

Figure 4.4 The structure of the fuzzy neural network used.

4.4.1 Preprocessing input

The input consists of a set of N market variables defined as $Mv_i, (i = 1,..., N)$, and a stock price S, which we wish to simulate. The market variables can be either daily stock prices (other than S) or market indexes. Again, we note that in this thesis only the NZSE 40 index is used as the market variable. The first step is to calculate the relative change of the input, so that different market variables and the stock can be compared in terms of relative growth [Kung (1993)]. The transformation used is

$$dMv_i^t = \ln\left(\frac{Mv_i^t}{Mv_i^{(t-1)}}\right), \quad i = 1, ..., N$$

where, Mv_i^t denote the market variable number i at time t, and dMv_i^t the logarithmic relative change of Mv_i^t.

The logarithmic relative change of stock price S at time t, dS^t is calculated in a similar manner. The transformation from stock price S at time t to dS^t is required because the sensitivity of stock S with respect to a market variable Mv_i is needed in the next step of the network which is the calculation of the ideal weights.

4.4.2 Ideal Weight (Sensitivity)

The prediction of the neural fuzzy networks model is based on the assumption that the price of stock S is related to the various market variables chosen. The networks are designed to learn this relationship and use it in establishing movements of the stock price that are either under or over priced. Each market variable Mv_i has a different sensitivity with respect to stock S. This sensitivity (the ideal weight) indicates whether Mv_i moves with or against S, and to what degree (see figure 4.5). The formula for the ideal weight is:

$$\hat{w}_i^t = \frac{dS^t}{dMv_i^t}, \quad i = 1, ..., N.$$

Where $\hat{w}_i^t$ denotes the ideal weight for market variable i at time t.

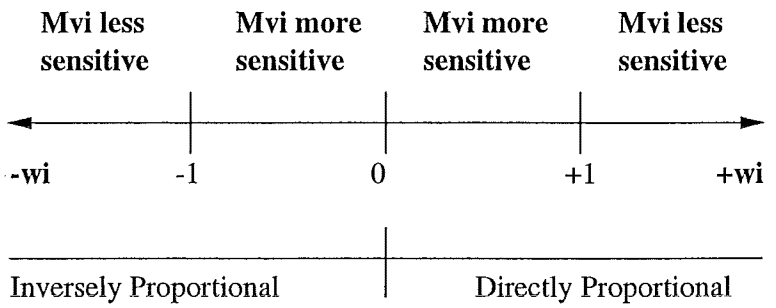

Figure 4.5 Sensitivity of stock S with respect to a market variable Mv_i .

4.4.3 Net weight and fuzzy learning rate

Before the neuron for a particular market variable i calculates its "prediction," the net weight is revised by taking a weighted average of the previous net weights and the previous ideal weights. The net weight at time t for a particular market variable i is

$$w_i^t = \eta \otimes w_i^{t-1} \oplus (1 \oplus \eta) \otimes \hat{w}_i^{t-1}$$

where η is the learning rate, w_i^{t-1} is the previous net weight and $\hat{w}_i^{t-1}$ is the previous ideal weight. Since different learning rates alter the magnitude of a market variable's previous performance, it is beneficial to use several learning rates. Thus, a single market variable can provide short-term

predictions as well as longer term ones. Therefore η is assigned to be a triangulate fuzzy number.

4.4.4 Activation Function and Prediction

At each iteration, the neuron fuzzy network model calculates N components from the overall network prediction, where N is the number of market variables used. To obtain a single prediction, all these components are averaged based on each neuron's reliability index. This reliability index RI_i is a Gaussian activation function of the average absolute error $(error_i)$, as shown in figure 4.6, and it is describe by

$$RI_i = C_1 \exp\left\{-error_i C_2^{-1}\right\} .$$

Since C_1 is effectively normalized by the weight average, any arbitrary value, for example 10, will work. However, C_2 defines the steepness of the curve, and thus defines how important a component is. For example, the component with 1% average error is more important than the component with 10% average error. If a 20-fold will be chosen more important a $C_2 = 0.05$ is used.

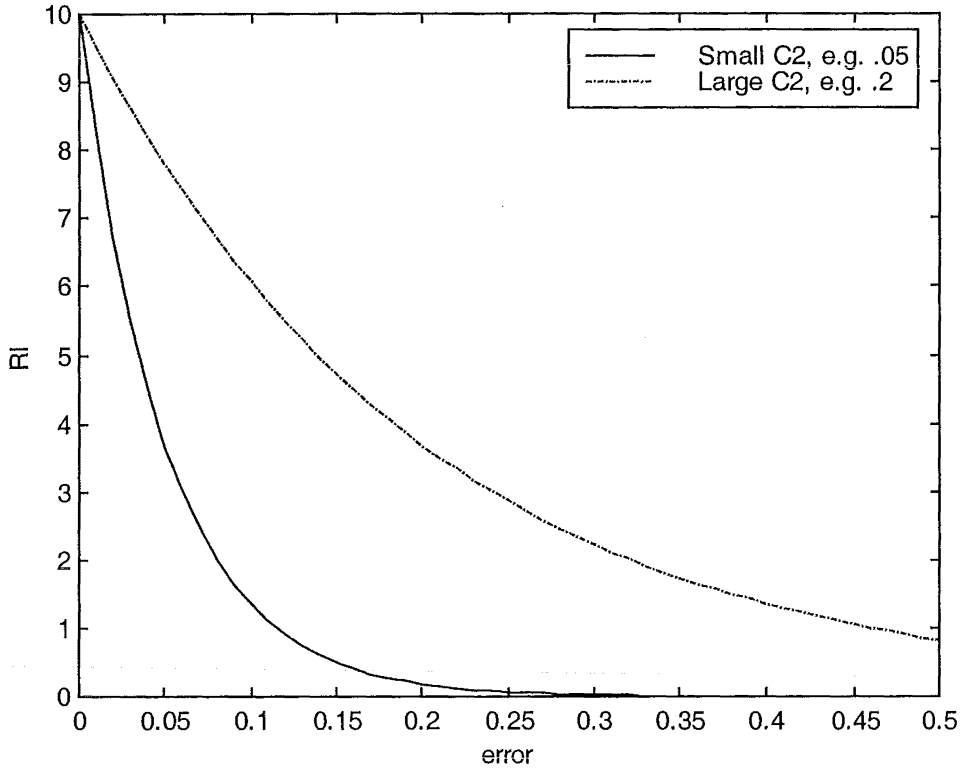

Figure 4.6. Gaussian activation function

For a particular market variable i (note again that the NZSE 40 is the only market variable used here) the fuzzy $w_t = (m, a, b)_{trian}$ is converted to the three crisp numbers $w_{t1} = m - a$, $w_{t2} = m$ and $w_{t3} = m + b$. These weights are combined with dMv_i^t to gives us the predictions $d\hat{S}_{t1}$, $d\hat{S}_{t2}$, $d\hat{S}_{t3}$ by the equation $d\hat{S}_{tk} = dMv_i^t w_{tk}$, for $k = 1,\ 2,\ 3$. Finally, these three predictions are averaged based on the reliability index to give the final prediction $\hat{S}_t$ for the stock price as discussed below. (See also figure 4.4).

4.5 Trading Stock price under neural fuzzy networks

The trading strategy simulation is summarized as follows.

1. Start with an initial amount (cash balance) of \$1000 and n the number of observations equal to two.
2. Predict the logarithmic relative change of price for the n^{th} observation.
3. If the predicted price in step 2 is less than the actual logarithmic relative change of price for n^{th} observation and the cash balance is more than 0, buy the stock at the $n^{th}+1$ observation price using all cash available.
4. If the predicted price in step 2 is more than the actual logarithmic relative change of price for n^{th} observation and the number of stocks held is more than 0, sell all of the stock at the $n^{th}+1$ observation price.
5. If the situation does not satisfy conditions 3 or 4 do not trade.
6. Increase n by 1 and go to step 2.

The use of this strategy is discussed at the end of section 2 (see section 2.5). The following table 4.1, 4.2 and figures 4.7 to 4.11 are the results of the trading strategy above. Note that the results are for either cash balance or the amount of stocks held in dollars noted here as "worth" and in figures 4.7 to 4.11 worth is compared with adjusted stock price.

The annual returns for trading under neural fuzzy networks are 15.35, -4.91, 12.96, 20.52 and 32.31 percents for Air New Zealand, Brierley, Carter Holt Harvey, Lion Nathan and Telecom respectively while the adjusted stock prices gives 6.13, -10.56, -0.86, -0.42, 20.73 percents respectively. Thus, the results show that the Bayesian dynamic linear model performed much better than the adjusted stock price even though trading Brierley share under the given model and strategies is not profitable.

Stock	Mean	Standard Deviation	Maximum	Minimum	Final Cash Balance
AIRVA	1868.24	519.52	3278.46	893.78	2795.49
BRY	1135.21	180.37	1504.98	631.85	695.81
CAH	1663.12	325.69	2487.60	949.80	2403.99
LNN	2090.02	845.11	3928.87	1000.00	3834.20
TEL	3764.85	1942.83	8034.43	989.95	7506.39

Table 4.1 Summary statistics of the trading simulation results (worth).

Stock	Mean	Standard Deviation	Max.	Min.	Final Cash Balance
AIRVA	1606.09	371.00	2281.77	784.53	1535.91
BRY	1122.45	303.32	1541.67	312.50	447.92
CAH	1199.52	263.89	1664.00	460.00	940.00
LNN	867.68	128.82	1210.05	598.17	970.32
TEL	3065.42	1059.54	5000.00	994.90	3882.65

Table 4.2 Summary statistics of adjusted stock price.

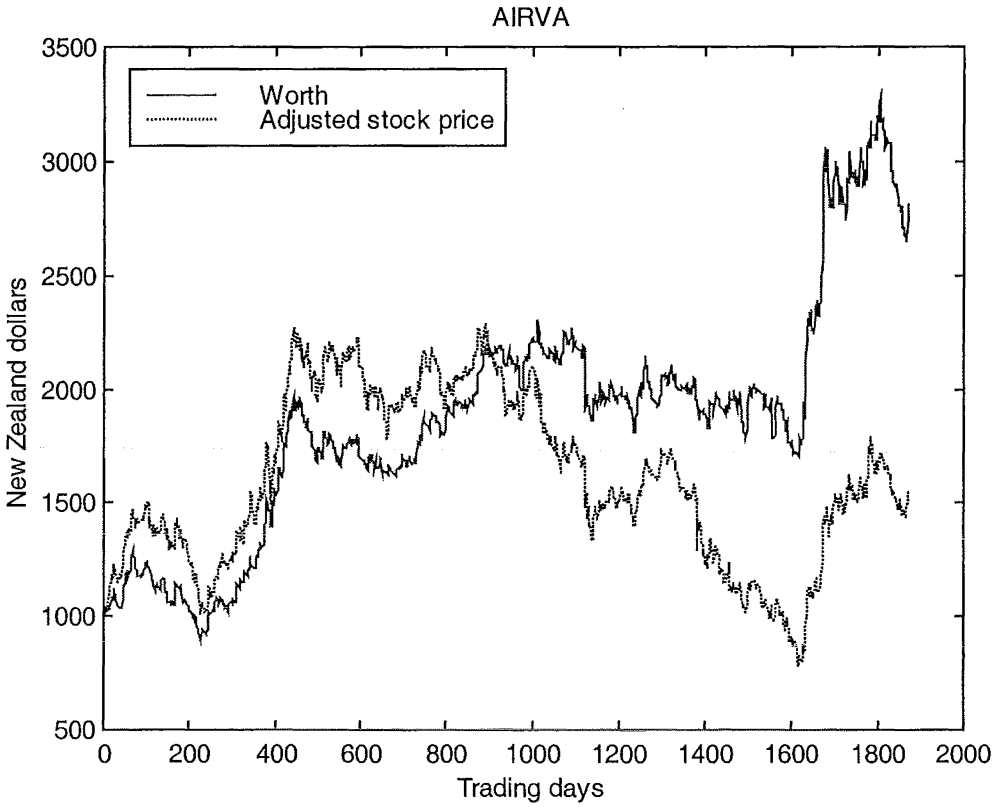

Figure 4.7 Simulation result for Air New Zealand Ltd. "A" (AIRVA)

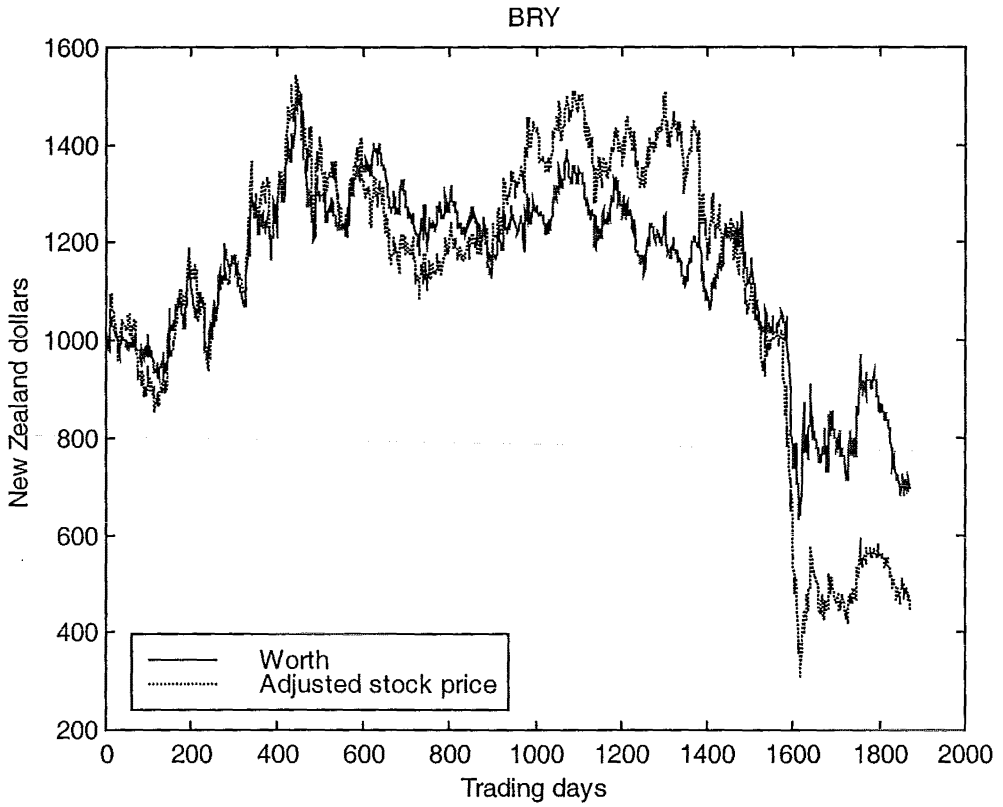

Figure 4.8 Simulation result for Brierley Investments Ltd. (BRY)

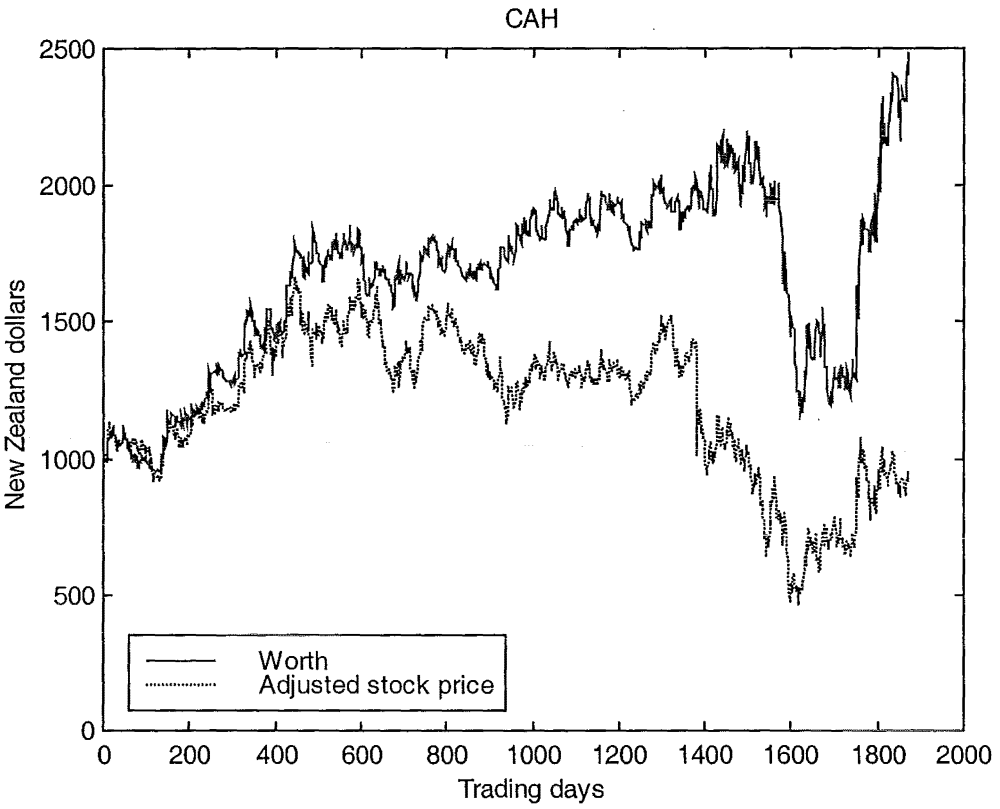

Figure 4.9 Simulation result for Carter Holt Harvey Ltd (CAH)

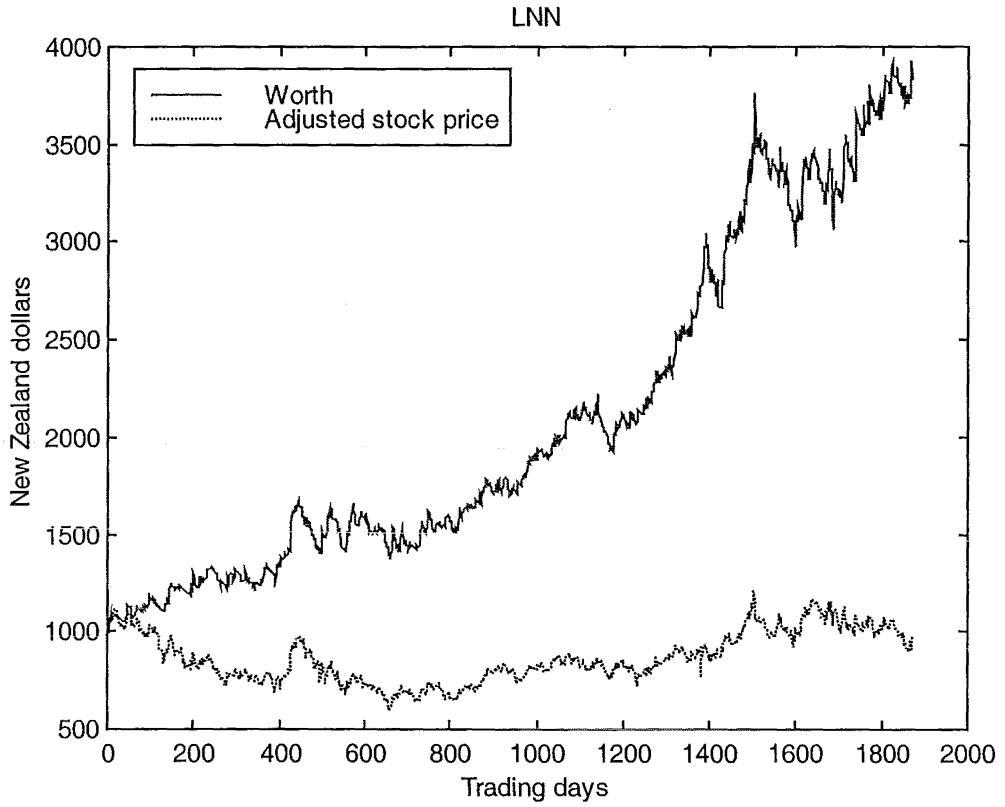

Figure 4.10 Simulation result for Lion Nathan Ltd. (LNN)

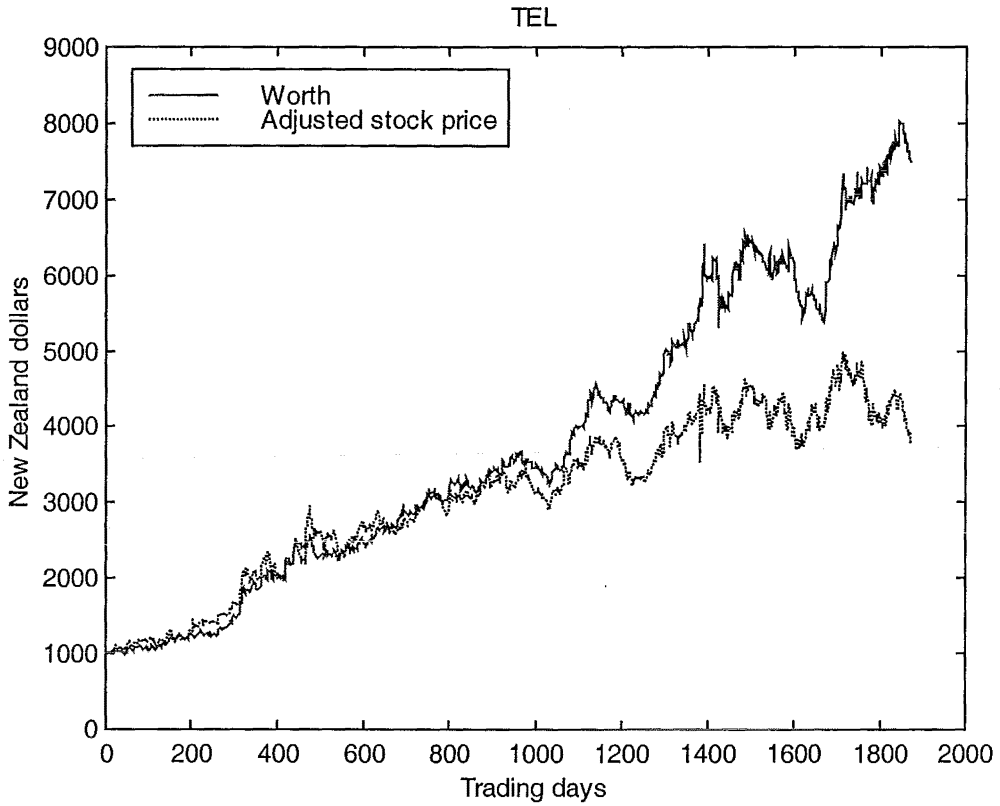

Figure 4.11 Simulation result for Telecom Corporation of New Zealand Ltd. (TEL)

Chapter 5

5 Summary of models used for forecasting stock market prices

5.1 Introduction

For the past forty years or so there has been a growing interest on forecasting the stock market prices. Many different models and assumptions have been made. In this chapter some of these models will be discussed but it is impractical to list all the methods that have been used.

Section 5.2 is a summary of the most widely accepted hypothesis about market prices, which is the Martingale hypothesis, which states that forecasting is not profitable. Section 5.3 discuss the use of other related time series and variables that some forecasters suggest as being helpful in predicting the stock market prices or in deciding on which stock to invest in. Section 5.4 reviews the regime-switching model that switch the model using an indicator variable, which can be applied to some interesting characteristics in stock prices to improve forecastability. Such characteristics are low volatility, earning and

size portfolio, seasonal effects, price reversal and extra-ordinary returns that will be discussed in section 5.4.1 to 5.4.5 respectively. Section 5.5 looks at disaggregation methods where cross-section regressions with time-varying coefficients are examined. Section 5.6 considers models that use causal variables that have significant effect on stock market prices. Section 5.7 looks at heteroskedasticity of the absolute stock price return, which can be modeled by Generalized Autoregressive Conditional Heteroskedasticity (GARCH) type models.

5.2 Martingale Hypothesis

The most common hypothesis about stock prices is the martingale hypothesis, which states that stock prices are a martingale process i.e.

$$E[P_{t+1} \mid I_t] = P_t,$$

where I_t is any information set which includes the prices $P_{t-j}, j \geq 0$. This hypothesis can be describe in economist terms as the Efficient Market Hypothesis. There are various forms of this hypothesis and an example is the one given by Jensen (1978):

"A market is efficient with respect to information set I_t if it is impossible to make economic profits by trading on the basis of this information set." By "economic profits" is meant the risk-adjusted returns "net of all cost". An obvious difficulty with this hypothesis is that it is unclear how to measure risk.

A survey by Fama (1970) concluded that the stock market prices are a martingale process. The information sets used by Fama were I_{1t} (lagged prices or lags logged prices) and I_{2t} (I_{1t} plus a few explanatory variables such as earnings and dividends). The data periods were usually daily and monthly.

Also a survey by Granger (1972) concluded "Almost without exception empirical studies ... " support a model for $p_t = \log P_t$ of the form $\Delta p_{t+1} = \theta \Delta p_t + 1_{t-1} + \varepsilon_{t+1}$, where θ is near zero, 1_t contributes only to the very low frequencies and ε_t is zero mean white noise. This model is closely approximates to the Efficient Market Hypothesis since $E[\Delta p_{t+1}]$ is almost always equal to zero. Many forecasters today still hold this view. However others feel that profitable forecasting is possible, here are some of them.

5.3 Use of related series and risk measurement

Many researchers in the late 1980's saw a new development in forecastability, where new methods, new data sets, longer series, different time periods and new explanatory variable were used. An important reference to these is Guimarães, Kingsman and Taylor (1989). The objective of this reference is to survey some of the stock forecasting works and make suggestions to forecasters working on other related series to the stock price P_t, such as

D_t = dividend for period t,

R_t = return = $(P_t + D_t + P_{t-1}) / P_{t-1}$,

r_t = return on a "risk free" investment

[The return on 30-day U.S. Treasury bills is often used in empirical work.],

$R_t - r_t$ = excess return,

β = non-diversifiable risk level of the stock,

Note that the capital asset pricing model (CAPM) introduced by Black, Jensen and Scholes (1972) is often used to approximate β the non-diversifiable risk level of the stock. The capital asset pricing model is describe by

$$R_t - r_t = \beta\,(Rm_t - r_t) + \varepsilon_t\,,$$

where Rm_t is the return on the market portfolio based on all assets and ε_t is the random error. Share brokers often use CAPM to decide whether they should invest in certain stock.

Empirically the Center for Research on Securities Prices (CRSP) at the University of Chicago provides an Rm_t, based on the value-weighted transactions of all stocks listed on the New York and American Stock exchanges. Also some studies called $Rm_t - r_t$ in CAPM "*market return*" which can be measured empirically by the Standard and Poor's 500 index.

5.4 Regime switching models

Another approach by some researchers was to look at switching models. These take into account that under different situations the forecastability is different

and so used a *regime switching model* to allow for this. We say that x_t is a regime-switching model if x_t is generated by:

$$x_t = \begin{cases} \alpha_1 + \gamma_1 x_{t-1} + \varepsilon_t & \text{if } z_t \text{ in } A \\ \alpha_2 + \gamma_2 x_{t-1} + \varepsilon_t & \text{if } z_t \text{ not in } A \end{cases}$$

where z_t is an auxiliary indicator variable, A is the switching variable , γ and α are constants and ε_t is the random error. For example, if z_t is a lagged value of x_t one has the switching threshold autoregressive model (STAR) discussed in detail in Tong (1990). However z_t can be more generals, and five examples of this follow. The variance of the residual ε_t may also vary with regime.

5.4.1 Switching with Low Volatility

LeBaron (1990) suggest the following switching model based on his observation that when stock price volatility is low the price tends to be more forecastable.

He uses R_t, the weekly returns of the Standard and Poor 500 index from 1946 to 1985, giving about 2,000 observations. The indicator variable used is the recent volatility

$$z_t = \hat{\sigma}_t^2 = \sum_{i=0}^{10} R_{t-i}^2$$

and the switching variable A is the lowest one-fifth quantile of the observed $\hat{\sigma}_t$ value in the first half of the sample. The regime-switching model was estimated using the first half of the sample and post-sample true one-step forecasts were evaluated over the second half. The model found is

$$R_t = \begin{cases} \alpha + 0.18R_{t-1} + \varepsilon_t & \text{if } z_t \le A \\ \alpha + \varepsilon_t & \text{otherwise,} \end{cases}$$

For the low volatility regime he finds a 3.1 percent improvement in forecast mean squared error over white noise with non-zero mean. He first takes α to be constant across regimes; but relaxing this assumption did not result in improved forecasts. He also tried different volatility regimes but no improvement was found.

5.4.2 Switching according to Earnings and size portfolios

Unlike the previous study that use the regime switching model to switch on a single stock price, Keim (1989) looks at the whole stock market. He formed portfolios based on the market value of equity (size) and the ratio of earning to price (E/P) and then calculated monthly returns (in percentages) using all the quoted stocks on either the New York or American Stock Exchanges from 1951 to 1986.

Each March 31st all stocks were ranked on the total market value of the equity (price × number of shares). The bottom 10% were put into the first (or smallest) portfolio, the next 10% in the second portfolio and so on up to the shares in the top 10% ranked giving the "largest" portfolio. The portfolios were changed annually and average monthly returns calculated. Similarly, the portfolios were formed form highest E/P values to lowest (positive) values. Shares of companies with negative earnings went into a separate portfolio.

He showed that the portfolio with the highest E/P ratio and smallest size has both a high average return and a beta value (the non-diversifiable risk for the given stock) only slightly above that of a randomly selected portfolio. The result was found to hold for both non-January months and for January,

although returns in January were much higher. Somewhat similar results have been found for stock on other, non-U.S. exchanges.

The results are consistent with a regime-switching model with regime determined by the size and E/P variables at the start of the year. However, as ranking is used, these variables for a single stock are related to the actual values of the variables for all other stocks.

5.4.3 Switching by seasonal effects

Several of seasonal effects have been suggested but the strongest and the most widely documented is the January effect.

For example Keim (1989) found that the portfolio using the highest E/P values and the smallest size gave an average return of 7.46 (standard error 1.41) over Januarys but only 1.39 (0.27) in other months. A second example is the observation that the small capitalization companies (bottom 20% of companies ranked by market value of equity) out-performed the S&P index by 5.5 percent in January for the year 1926 to 1986. These small firms earned inferior returns in only seven out of the 61 years. Other examples are given in Ikenbury and Lakonishok (1989).

The evidence suggests that the means of returns have regime changes with an indicator variable that takes a value of unity in January and zero in other months.

5.4.4 Switching by price reversal

Several studies have found that shares that perform poorly over one period are likely to perform well over a subsequent period, thus giving price change reversals see DeBondt (1989).

For example, Dyl and Maxfield (1987) selected 200 trading days at random in the period January 1974 to January 1984 and on each day three NYSE or AMEX stocks with the greatest percentage of price loss (on average -12%) were noted. Over the next ten trading days, these stocks earned a risk-adjusted return of 3.6 percent. The three highest gainers lost an average 1.8-% over the same period. Other studies find similar evidence for daily, weekly and even monthly returns. Transaction costs will be fairly heavy and a strategy based on these results will probably be risky.

However, Lehman (1990) considered a portfolio whose weights depended on the return of a security the previous week minus the overall return, with positive weights on previous losers and negative weights on previous winners. The portfolio was found to consistently produce positive profits over the next week, with very few losing periods and so with small risk. Transaction costs were again substantial but worthwhile profits were achieved for transaction costs at a level appropriate for larger traders. Thus, after allowing for risk and costs, a portfolio based on price reversal was found to be clearly profitable.

Long-term price reversals have also been documented. For example, Dark and Kato (1986) found in the Japanese market that for the year 1964 to 1980, the three-year returns for declined portfolios of extreme previous losers exceed the comparable returns of extreme previous winners by an average 70 percent. In this case the indicator variable is the extreme relative loss value of the share. As before, the apparent forecastability leads to a simple investment strategy, but knowledge is required of the value taken by some variables based on all stocks in some market.

5.4.5 Switch between "ordinary" returns and extra-ordinary returns

It is well known that the stock markets occasionally experience extraordinary movements, as occurred in October 1987, for example. Friedman and Laibson (1989) point out that these large movements are of overpowering importance and may obscure simple patterns in the data. They consider the Standard and Poor 500 quarterly excess returns (over treasury bills) for the first quarter of 1954 to the last quarter of 1988. After removal of just four extreme values, chosen by using a Poisson model the remaining data fits an AR(1) model with significant lag coefficient of 0.207 resulting in an R^2 value of 0.036. The two regimes are thus the "ordinary" excess returns, which seem to be forecastable, and the extra-ordinary returns, which are not, from the lagged data at least.

5.5 Disaggregation method

Even from the early work on forecasting stock market prices the aggregates of stock prices such as the Dow Jones or, Standard and Poor indices, [Fama(1970), Granger(1972)] or portfolios of a random selection of stocks or some small group of individual stocks were used.

Since then the availability of fast computers and of daily data for all securities on the New York and American stock exchanges, for example, allows examination of all the stocks. The situation allows cross-section regressions with time-varying coefficients, which could detect regularities that were not previously available. Jegadeesh (1990) for example, uses monthly data to fit cross-section models of the form

$$R_{it} - \overline{R}_{it} = a_{0t} + \sum_{j=1}^{12} a_{jt} R_{i,t-j} + a_{13,t} R_{i,t-24} + a_{14,t} R_{i,t-36} + \mu_{it}$$

for each month where R_{it} is the return on security i in month t and $\overline{R}_{it}$ is the average return over a long period (four or six years) which exclude the previous three years. Thus, a lagged average relationship is considered with coefficients changing each month. The subscript i is included to emphasize that this operation is carried out across the cross section. Many of the average coefficients $a_{j,t}$ were significantly different from zero, particularly at lags one and twelve, but other average coefficients were also significant, including at lags 24 and 36. A few examples are shown, with t-values in brackets.

Table 6.1 Cross-Sectional Regression Estimate

	$\bar{a}_1$	$\bar{a}_{12}$	$\bar{a}_{14}$	R_c^2
All months	-0.09(18)	0.034(9)	0.019(6.5)	0.108
January	-0.23(9)	0.08(5)	0.034(2.6)	0.178
Feb. to Dec.	-0.08(17)	0.03(8)	0.017(6)	0.102

Source: Jegadeesh (1990).

The result also shows that there is some time varying structure in the data since the R_c^2 values (the adjusted multiple coefficient of determination) are around 10%. Also as noticed earlier, January has more forecastability than other months.

5.6 Use of causal variables

Most of the studies discussed so far have considered forecasting of prices from just previous prices but it is also obviously sensible to search for other variables that provide some forecastability. The typical regression model is

$$\Delta p_t = \text{constant} + \underline{\beta}' \underline{X}_{t-1} + \varepsilon_t ,$$

where $\underline{X}_t$ is a vector of plausible explanatory, or causal variable, with a variety of lags considered. For example Darrat (1990) considered a monthly price index from the Toronto Stock Exchange for the period January 1972 to February 1987 and achieved a relationship:

$$\Delta p_t = \underset{(6)}{2.3}\Delta V_{t-1} - \underset{(5.3)}{0.25}\Delta P_{t-1} + \underset{(4.2)}{0.35}\Delta L_{t-10} - \underset{(3.0)}{0.015}\Delta C_{t-3},$$

$$R^2 = 0.46\text{, Durbin \& Watson = } 2.01,$$

where p_t, V, P, L and C represent the monthly price index, volatility of interest rates, production index, long term interest rate and cyclical adjusted budget deficit respectively. Note that only significant terms are shown and the modulus of *t-values* is in parentheses underneath. Many other variables were considered but not found to be significant, including changes of short-term rates, inflation rate, base money and the US - Canadian exchange rate, all lagged once. An apparently high significance R^2 value is obtained but no out-of-sample forecastability is investigated.

Pesaran and Timmerman (1990) also consider simple models that produce useful forecastability and they also conduct a careful evaluation of the model.

Here is an example of their model where the dependent variable Y_t is the quarterly excess return on the Standard and Poor 500 portfolio, the period considered from the first quarter of 1960 to the last quarter of 1988.

$$\begin{aligned} Y_t = &-0.097 + \underset{(5.7)}{17.2} D_{t-2} - \underset{(2.8)}{1.59} I_{t-3} - \underset{(6.2)}{0.03} TE_{t-1} \\ &+ \underset{(4.6)}{0.025} TB_{t-2} + \underset{(5.5)}{0.066} \Delta B_{t-1} + \varepsilon_t, \end{aligned} \tag{5.1}$$

$$R_c^2 = 0.364, \text{Durbin \& Watson} = 2.02,$$

where *D, I, TE, TB* and *B* represent dividend yield, inflation rate, T-bill end, T-bill begin and twelve month bond state respectively. Here dividend yield at time *t* is

$$\frac{\text{dividend on S \& P index } (t-1)}{\text{price of S \& P index } (t)}.$$

T-bill "end" means it is measured at the end of the third month of the quarter, "begin" indicates that it is measured at the end of the first month of the quarter. Therefor the two *T*-bill terms in the equation are effectively the change in the *T*-bill interest rate from one month to the next, plus one at the end of the quarter. Since a reasonable R_c^2 is found, the model can potentially be used for forecasting. (It should be noted that R_c^2 increased to 0.6 or so for annual data.)

There are some experiments with non-linear lagged dependent variables which produced some increase in R_c^2 to about 0.39, but this more complicate model was not further evaluated. A simple switching portfolio trading rule was considered:

(i) Buy the S&P 500 if the excess return was predicted to be positive according to equation (5.1), with the equation being sequentially re-estimated. Thus only data available at the time of the forecast was used in making the forecast.

(ii) If the predictor in equation (5.1) was negative, invest in *T*-bills.

The following table shows the rate of returns achieved by either a "buy-and-hold" market portfolio, or the switching portfolio obtained from the above trading rule or by just buying *T*-bills. As the switching rule involves occasional buying and selling, possibly, two levels of transaction costs are considered $\frac{1}{2}\%$ and 1%.

	Investment strategy			
	Market	Switching		*T*-bill
Transaction costs	0	0.5%	1%	0
Interest rate of returns	9.51	13.30	12.39	6.34
Standard deviation of returns	8.23	5.43	5.41	0.70
Wealth at end of period[a]	1394	3736	2961	595

[a] The period considered from 1960.I to 1988.IV and the wealth accumulates from an investment of $100 in December 1959.

Source: Pesaran and Timmerman (1990).

Although the results presented slightly favour the switching portfolio as zero transaction costs are assumed for the alternative investments, the trading rule based on equation (5.1) is seen to produce greater returns than the market (S&P 500) portfolio. They conclude that dividend / price ratios and interest rates have good long-run forecasting abilities for stock price index returns.

5.7 Return in stock price and GARCH models

If R_t is a return in stock price it has been found that R_t^2 is forecastable and $|R_t|$ even more so. Taylor (1986) finds evidence for this using U.S. share prices and Kariya, Tsukuda and Maru (1990) get similar results for Japanese stocks.

Ding, Granger and Engle (1993) also studied these properties using 17055 observations from the Standard and Poor 500 and found that it is possible to characterize $|R_t|^d$ for positive values of d to be "long-memory", with quite high autocorrelations for long lags and these properties are strongest for $d = 1$ or near 1. For example, if R_t is the daily return from the U.S. Standard and Poor index the autocorrelations for R_t are generally very small, the autocorrelations for R_t^2 are consistently above 0.1 up to lag 100 and for $|R_t|$ are above 0.35 up to lag 100. The so called "long memory" properties are examined for several other long speculative asset returns series by Ding, and Granger (1996). Some of the series they studied are the Japanese stock market index Nikkei, foreign exchange rate returns for the Deutschmark with the US

dollar, individual stock returns for Chevron, and minute-by-minute stock returns for a Japanese food company, Ajinomoto.

One possible explanation for the large positive autocorrelation between $|R_t|$ and $|R_{t+\tau}|$ or $|R_t|^d$ and $|R_{t+\tau}|^d$ is the heteroskedasticity of the data, i.e. the variance or conditional variance is changing over time. One family of nonlinear time series models that is able to capture some aspect of the time varying volatility structure is Engle's ARCH (AutoRegressive Conditional Heteroskedasticity) model Engle (1982). The ARCH model is defined as a data generating process for a random variable which has a conditional normal distribution with conditional variance a linear function of lagged squared residuals. More formally, the ARCH(p) model is defined as follows:

$$
\begin{aligned}
&R_t = \mu + \varepsilon_t, \\
&\varepsilon_t = s_t e_t, \qquad\qquad e_t \sim N(0,1), \\
&s_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 .
\end{aligned}
$$

where s_t^2 is the conditional variance, α_0 and α_is are constants.

Since its introduction by Engle (1982), the ARCH model has been widely used to model time-varying volatility and volatility that frequently has shocks

signal. Many modifications and extensions of the original ARCH model have also appeared in the literature.

For example, in order to capture the long memory property of the conditional variance process , Bollerslev (1986) introduce the GARCH(p,q) process (Generalized Autoregressive Conditional Heteroskedasticity), which defines the conditional variance equation as follows:

$$s_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j s_{t-j}^2 .$$

Taylor (1986) suggests modeling conditional standard deviation instead of conditional variance. Schwert (1989), following the argument of Davidian and Carroll (1987) modeled the conditional standard deviation as a linear function of lagged absolute residuals. The Taylor/Schwert GARCH(p,q) process defines the conditional standard deviation equation as follows:

$$s_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \left|\varepsilon_{t-i}\right| + \sum_{j=1}^{q} \beta_j s_{t-j} .$$

According to Ding, Granger, and Engle (1993) both Bollerslev's GARCH and Taylor/Schwert's model with appropriate parameters can produce similar

correlation patterns to the correlation pattern of R_t. It should be noted that most literature use Bollerslev's model when they mention GARCH processes.

It is also commonly found in empirical research that GARCH (1,1) with the estimated value of $\alpha_1 + \beta_1$ that is close to one represent R_t quite well. For example, Taylor (1986) estimated GARCH (1,1) models for 40 different financial time series. The results show that for all but six of the 40 series the estimated value of $\alpha_1 + \beta_1$ is greater than or equal to 0.97. In Ding, Granger, and Engle (1993), the estimated value of $\alpha_1 + \beta_1$ for daily S&P 500 returns is 0.997. This regularity is widely considered to be a characteristic of volatility persistence.

The Integrated GARCH(1,1) model, which restricts $\alpha_1 + \beta_1 = 1$, is then introduced by Engle and Bollerslev (1986) to model long-run volatility persistence. The IGARCH(1,1) model is always related to the random walk process in mean (random walk with drift). However, Nelson (1990) shows that IGARCH(1,1) process without drift is strictly stationary or goes to zero almost surely; the latter even when it is not covariance-stationary.

Further study by Ding and Granger (1996) shows that the pattern of the sample autocorrelation for various speculative returns is quite different from that of

the theoretical autocorrelation functions given by a GARCH(1,1) or IGARCH(1,1) processes. It is quite clear from the sample autocorrelation that there are different volatility components that will dominate different time periods. Some volatility components may have a very big short-run effect, but die out very quickly. Some of them may have a relatively smaller short-run effect, but last for long time periods. Ding and Granger (1996) suggest the Long Memory ARCH process, which is defined as follows:

$$\varepsilon_t = s_t e_t, \quad e_t \sim \text{i.i.d. D}(0,1),$$
$$s_t^2 = s^2(1-\mu) + \mu \sum_{k=1}^{\infty} \frac{B(p+k-1, q+1)}{B(p,q)} \frac{\varepsilon_{t-k}^2}{E|e_t|^2}.$$

where $B(p,q)$ is a *Beta* distribution parameters p,q and D(0,1) is any distribution with mean 0 and variance 1.

The estimated results show that this model gives a much better description for the data.

Also Hans Franses, van Dijk (1996) found that some stock market indices returns are skewed to the left i.e. there are more negative than positive outlying observations. The actual symmetric GARCH model cannot cope with such

skewness and, hence, one can expect that forecasts and forecast error variances from a GARCH model may be biased for skewed time series. Several models that deal with the problem were tested on weekly observed indices for the stock market in Germany (DAX), The Netherlands (EOE), Spain (MAD), Italy (MIL) and Sweden (VEC). Their results show that the Quadratic GARCH(1,1) model proposed by Engle and Ng (1993) can significantly improve on the linear GARCH model and the no-change cases (martingale model) when the forecasting models are calibrated on data which exclude such extreme events as the 1987 stock market crash. The QGARCH(1,1) model can be seen to be similar to GARCH(1,1) model except

$$s_t^2 = \alpha_0 + \alpha_1\left(\varepsilon_{t-1} - \gamma\right)^2 + \beta s_{t-1}^2$$

where $\alpha_0, \alpha_1, \beta > 0, \alpha_1 + \beta < 1$ and γ take a positive value.

Chapter 6

6 Conclusion

As the results from Chapters 2 to 4 show, it is possible to have profitable trading for all the stocks tried using ARIMA models, Bayesian dynamics linear models and fuzzy neural networks with similar trading strategies. Overall, the fuzzy neuron networks models out-performed other models for most stocks except Brierley Investment stocks where the ARIMA models performed better. The Bayesian dynamic linear models' performances are poor but close to ARIMA performances except for Telecom Corporation stocks where the Bayesian models out-performed the ARIMA models.

While the considerations of transaction cost have been ignored in this thesis, with the technology of online trading the traders now can trade stocks via the internet with much less transaction cost required so that the cost of trading will not considerable affect the conclusions.

The summary of different forecasting techniques used by other researchers in chapter 5 has shown that stock market price may have a structure but more research is needed to identify the structure. Whereas a simple strategy has been used here, other strategies should also be investigated and should be expected to give better results.

Thus, this thesis has shown that by using the past observations of stock prices, the above models and trading strategies, profitable forecasting is possible.

Reference:

Black, F., Jensen, M. and Scholes, M. (1972) The Capital Asset Pricing Model: Some Empirical test, ed., Michael Jensen, *Studies in the Theory of Capital Markets*. Praeger: New York.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307-327.

Box, G. E. P. and Jenkins, G. M. (1970) *Time Series Analysis: Forecasting and Control*, Holden-Day: San Francisco.

Dark, F. H. and K. Kato (1986) *Stock market over-reaction in the Japanese stock market*, Working paper, Iowa State University.

Darrat, A. F. (1990) Stock returns, money and fiscal deficits, *Journal of Financial and Quantitative Analysis* **25**, 387-398.

Davidian, M. and R. J. Carroll (1987) Variance function estimation, *Journal of American Statistical Association* **82**, No. 400, 1079-1091.

DeBondt, W. F. M. (1984) Stock price reversals and over-reaction to news events: A survey of theory and evidence, in: Guimarães et al. (1989).

Ding, Zhuanxin, C. W. J. Granger (1996) Modeling volatility persistence of speculative returns: A new approach, *Journal of Econometrics* **73**, 185-215.

Ding, Zhuanxin, C. W. J. Granger and R.F. Engle (1993) A long memory property of stock market returns and a new model, *Journal of Empirical Finance* **1**, 83-106.

Dyl, E. A. and K. Maxfield (1987) *Dose the stock market over-react?* Working paper, University of Arizona.

Engle, R. F. (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, *Econometrica* **50**, 987-1008.

Engle, R. F. and Ng, V. (1993) Measuring and testing the impact of news on volatility, *Journal of Finance* **48**, 1749-1778.

Engle, R. F. and T. Bollerslev (1986) Modeling the persistence of conditional variances, *Econometric Reviews* **5**, 1-50, 81-87.

Fama, E. F. (1970) Efficient capital markets: A review of theory and empirical work, *Journal of Finance* **25**, 383-417.

Friedman, B. M. and D. I. Laibson (1989) *Economic implications of extraordinary movements in stock prices*, Working paper, Economics Department, Harvard University.

G. E. P. Box and G. M. Jenkins (1962) Some Statistical Aspects of Adaptive Optimization and Control, *Journal of the Royal Statistical Society* **B 24** (2), 279-331.

Granger, C. W. J. (1972) Empirical studies of capital markets: A survey, in: eds., G. Szego and K. Shell, *Mathematical Methods in Investment and Finance*. North-Holland: Amsterdam.

Guimarães, R. M. C., B. G. Kingsman and S. J. Taylor (1989) *A Reappraisal of the Efficiency of Financial Markets*. Springer-Verlag: Berlin.

Harrison, P. J. and Stevens, C. F. (1976) Bayesian forecasting, *Journal of Royal Statistics Society* **B 38**, 205-228.

Ho, Y. C. and Lee, R. C. K. (1964) A Bayesian approach to problems in stochastic estimation and control, *IEEE Transactions on Automatic Control* **AC-9**, 333-339.

Hobbs, A. and Bourbakis, N. G. (1995) A neurofuzzy arbitrage simulator for stock investing, *Proc. IEEE/IAFE 1995 Conf. Computational Intelligence Financial Engineering*, New York, Apr. 9-11.

Hopfield J. (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities, *Proceedings of the National Academy of Sciences* **79**, 2554-2558.

Ikenberry, D. and J. Lakonishok (1989) Seasonal anomalies in financial markets: A survey, in: Guimarães et al. (1989).

Jazwinski, A. H. (1970) *Stochastic Processes and Filtering Theory*, Academic: New York.

Jegadeesh, N. (1990) Evidence of predictable behaviour of security returns, *Journal of finance* **45**, 881-898.

Jensen, M. C. (1978) Some anomalous evidence regarding market efficiency, *Journal of Financial Economics* **6**, 95-101.

Kalman, R. E. (1960) A new approach to linear filtering and prediction problems, *Transactions of the American Society of Mechanical Engineers, Journal of Basic Engineering* **82**, 35-45.

Kalman, R. E. and Bucy, R. S. (1961) New results in linear filtering and prediction theory, *Transactions of the American Society of Mechanical Engineers, Journal of Basic Engineering* **83**, 95-108.

Kariya, T., T. Tsukuda and J. Maru (1990) *Testing the random walk hypothesis for Japanese stock prices in S. Taylor models*, Working paper 90-94, Graduate School of Business, University of Chicago.

Keim, D. B. (1989) Earning yield and size effect: Unconditional and conditional estimates, in: Guimarães et al. (1989).

Laarhoven, P.J.M. and Pedrycz, W.A. (1983) Fuzzy extension of Saaty's priority theory, Fuzzy Sets and Systems **11**, 229-241.

LeBaron, B. (1990) *Forecasting improvements using a volatility index*, Working paper, Economics Departments, University of Wisconsin.

Lehman, B. N. (1990) Fad, martingales, and market efficiency, *Quarterly journal of Economics* **105** (1), 1-28.

Lukasiewicz, J. (1920) O logice trójwartociowej, *Ruch Filozoficzny* **5**, 170f.

Nelson, D. B. (1990) Stationarity and persistence in the GARCH(1,1) model, *Econometric Theory* **6**, 318-334.

P. H. Franses and D. van Dijk (1996) Forecasting Stock Market Volatility Using (Non-Linear) Garch Models, *Journal of Forecasting* **15**, 229-235.

Pesaran, M. H. and A. G. Timmerman (1990) *The statistical and economic significance of the predictability of excess returns on common stocks*, Program in Applied Econometrics Discussion paper # 26, University of California, Los Angeles.

Pole, A., West, M. and Harrison J. (1994) *Applied Bayesian Forecasting and Time Series Analysis*. Chapman & Hall: New York.

Rosenblatt, F. (1961) *Principles of Neurodynamics*, Spartan Press: Washington D.C.

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

Schwert, W. (1990) Stock volatility and the crash of '87, *Review of Financial Studies* **3**, No. 1, 77-102.

Taylor, S. (1986) *Modelling Financial Time Series*. Wiley: Chichester.

Thiele, T. N. (1880) *Sur la compensation de quelques erreurs quasi-systematiques par la methode des moindres carrees*, Reitzel: Copenhagen.

Tong, H. (1990) *Non-linear Time Series*. Oxford University Press: Oxford.

Tsay, R. S. and G. C. Tiao (1985) Use of canonical analysis in time series model identification, *Biometrika* **72**, 299-315.

West, M. and Harrison, J. (1989) *Bayesian Forecasting and Dynamic Models*. Springer: New York.

Yule, G. (1927) On a Method of Investigating Periodicity in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers, *Philosophical Transactions of the Royal Society of London* **A 226**, 265-298.

Zadeh, L. (1972) A rationale for Fuzzy Control, *Journal of Dynamic Systems, Measurement, Control* **94** (6), 3-4.