ANALYTIC MODELS OF ELECTRICITY LOAD

By

Xichuan Zhang

A THESIS SUBMITTED FOR

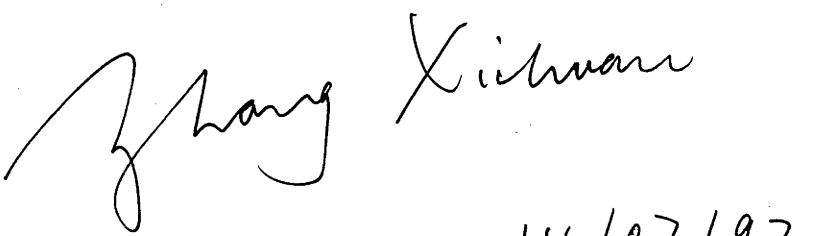
THE DEGREE OF DOCTOR OF PHILOSOPHY OF

THE AUSTRALIAN NATIONAL UNIVERSITY

July 1, 1992

Declaration

The results of this thesis are my own except where otherwise acknowledged.



14/07/92

Abstract

The main topic of this thesis is modelling and forecasting electricity load using regional data observed from every few minutes up to a daily basis (short-term load data). The various models presented and evaluated in this thesis utilize half-hourly and weekly load data from New Zealand and quarter-hourly and daily load information and three hourly meteorological data from Canberra, Australia.

Subset AR model selection procedures have been investigated in chapter 2. The concept of the projection modulus of a particular lag in a subset AR model is established so that a more efficient subset AR searching algorithm is created. Applying the new algorithm to computer simulated and real data shows that the new algorithm is more efficient in searching for the optimal model. The impact on the performance of the Kalman filter of initial conditions for the state vector in a state space model has been examined and a recursive estimation procedure has been established to estimate the initial conditions so that the Riccati difference equation converges quickly.

The various approaches to short term electricity load modelling and forecasting have been systematically reviewed in chapter 1. In chapter 4, a new additive model is introduced. The setup of this model enables us to handle properly the trend, the transitions between the profiles of weekdays and weekend days, the important periodicities, and the residual innovation series.

In chapter 5, several multiplicative models are built. Comparisons of established models with these demonstrate that the proposed additive model and a subset ARAR model are best at both within sample fitting and post sample predictions. Theoretical analysis and real forecasting practice also show that the performance of these two

models are influenced by the seasons. This effect derives directly from the influences of climate on the load.

In chapter 6, a non-linear model of the relationship between the electricity load and temperature is built to take the climatic influences into account. The climatic influences were carefully identified though the existence of outliers and variance heterogeneity. This non-linear model allows extraction of the weather sensitive and insensitive loads. The relation between temperature and other load characteristics are captured in diagrams of the load profiles and of the time of peak load, both of which are derived from this non-linear model.

In chapter 7, two dynamic models, an ARMAX model and a structural state space model, which use weather information have been established for daily electricity load. A sequential hypothesis testing procedure is also provided to find the optimal state space models. Further study areas are suggested in chapter 8.

Acknowledgments

I am grateful to my supervisor, Professor R.D. Terrell, for the guidance, encouragement and criticism he has given me throughout my studies at ANU. I would also like to thank my advisor, Professor E.J. Hannan and Dr. T. Breusch for various discussions of my thesis.

My final thanks must go to my wife, Xiuping Jia, and my parents for all their support.

Notation

Abbreviations

AIC	Akaike information criterion
AR	auto-regression model
ARMA	autoregressive moving average model
ARIMA	autoregressive integrated moving average model
BIC	Bayes information criterion
CAM	conventional additive model
CF	criterion function
ECM	error-correction mechanism
FFT	fast Fourier transform
GLS	generalized least squares
LM	Lagrange multiplier
LR	likelihood ratio
LSE	least squares estimate
MA	moving average
MLE	maximum likelihood estimate
MPI	most powerful invariant
MSM	multiplicative seasonal model
MXFFT	mixed radix fast Fourier transform
NID	normally and identically distributed
RLS	recursive least squares
RSS	residual sum squares

SF	stability function
STLF	short term load forecasts

Symbols

$\langle \alpha, \beta \rangle$	inner product of α and β
cov	covariance (matrix)
E	expectation
$\mathrm{mod}(m, n)$	remainder of m divided by n
$\delta(i, j)$	Kronecker delta
$\mathbf{H}(X)$	linear space spanned by X
$Sp(a_1, \cdots, a_n)$	linear space spanned by basic elements $a_1, \cdots, a_n$
$Tr(A)$	trace of matrix A
var	variance

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The last three decades has seen the development of a range of models for *short-term* electricity load, where "short-term" is used throughout this thesis when the observations on electricity load are measureed at intervals ranging from every a few minutes to every day. The modeling techniques have been largely derived from statistical and systems engineering practice. With the recent development of techniques in the statistical and systems engineering disciplines, it has been possible to model and forecast short-term electricity load with greater accuracy and enhanced computational speed.

Short-term load information may be collected in its form of system load, regional load, or bus load. Different purposes will require us to produce short-term forecasts with lead times which range from a few minutes to a few weeks. Total system load forecasting is valuable in making decisions on whether to introduce or shut off different plants and for scheduling the most economic allocation between different plants in a power network system. An accurate regional load forecast is needed to allocate the total system load between different plants or to arrange switching of load between the various generating systems. Bus load forecasts are also needed for some applications related to economic and system security problems. Although the contributions to the three types of load are different, they have many similar features. We concentrate on reviewing and studying regional load modelling and forecasting when data from two regional short-term loads are available.

1.1 Influential Factors of Regional Load

The regional load is the sum of all the individual demands from an area. In principle, one could establish the load pattern if each individual consumption pattern were known. Still, the demand or usage pattern of an individual load or customer has some degree of randomness and unpredictability. There is a diversity of individual usage patterns in any region. These factors make it ineffective to predict regional electricity demand levels by extrapolating the estimated individual usage patterns and aggregating over a region. Fortunately, however, the total of the individual loads produces a distinct consumption pattern which can be statistically predicted.

The regional load behavior is influenced by a number of factors. The major factors are categorized as follows and discussed in the next four subsections.

- economic
- time
- weather
- random effects.

1.1.1 Economic Factors

The economic environment in which a utility operates has a clear effect on the electricity demand consumption patterns. Factors, such as area demography, levels of industrial and/or commercial activities, nature and level of penetration and saturation of the appliance population, developments in the regulatory climate and more generally, economic trends have significant impacts on the regional load behaviour. The economic factors are not, however, explicitly represented in the short-term load forecasting models because of the longer term impact associated with them.

1.1.2 Time Factors

Three principal time factors – seasonal effects, weekly and daily cycles, casual, legal and religious holidays – play an important role in influencing the profile of load patterns. Certain changes in the load pattern occur gradually in response to seasonal variations such as the number of day-light hours and the changes in temperature. On the other hand, there are seasonal events which bring about abrupt but important structural modifications in the electricity consumption pattern. These are the shifts to and from Day-light Saving Time, changes in the rate structure(time-of-day or seasonal demand), start of the school year, and significant reductions of activities during vacation periods(e.g. Christmas-New Year period). The weekly- daily- periodicity of the load is a consequence of the work-rest pattern of the region population. The existence of statutory and religious holidays has the general effect of significantly lowering the load values to levels well below "normal".

1.1.3 Weather Factors

Meteorological conditions are responsible for significant variations in the load pattern. Most utilities have large components of weather-sensitive load, such as those due to space heating, air conditioning, and agricultural irrigation. In many systems, temperature is the most important weather variable in terms of its effects on the load. For any given day, the deviation of the temperature variable from a normal value may cause load changes of such a magnitude as to require major modifications in the pattern. The recent recorded history of temperature also affects the load profile, i.e. the "build up" effect. Humidity is a factor that may also affect the load in a manner associated with temperature, particularly in hot and humid areas. Thunderstorms also have a strong effect on the load due to the marked change in temperature that they can induce. Other factors that impact on load behavior are wind speed, precipitation, and cloud cover/light intensity.

1.1.4 Random Disturbances

A power system is continuously subject to minor random disturbances since the system load is a composite of a large number of diverse individual demands. There are also certain events such as widespread strikes, shutdown of industrial facilities and special television programs whose occurrence is known *a priori*, but whose effect on the load is uncertain.

1.2 Classification of Short-Term Load Forecasts

Based on the forecasting purpose, the data available, and computational requirements, the forecasting models are classified in this thesis into two categories: (1) models using only past load data or (2) models using both weather and load data. Furthermore, the models in each category can be further sub-classified into (1) an off-line approach or (2) an on-line approach. As the load demand is a time-dependent random process, many statistical techniques have been applied to load data. These include multiple regression approaches, time series approaches in the time domain, such as exponential smoothing, Box and Jenkins type models and state space approach, etc. and in the frequency domain, such as spectral decomposition, FFT, etc. Excellent reviews on the application of these techniques to short-term load forecasting along with their limitations are reported in Gross and Galiana (1987) and Abu-El-Magd and Sinha (1982). Valuable bibliographies on load forecasting can be found in Sachdev et al. (1977), IEEE Committee Report (1980) and IEEE Committee Report (1981). In the following sections, we review the the use of time series and state space models for "short term" load forecasting, and therefore this coverage will not include all existing model types in the literature. In the more detailed discussion (see section 1.2.1), the emphasis will be on a two stage process of modelling which separates the load into a base load part and a stochastic part including the effect of climate. The characteristic features of the different methods, their merits and drawbacks are stressed. Finally in this introduction chapter, some proposed models and associated model identification

techniques are suggested for our study in the main part of this thesis.

1.2.1 Two Stage Time-Series Model Approach

Suppose the load $\{X(t)\}$ at each discrete sampling time t of the forecast period for duration T is represented by a time series $\{X(t), t = 1, \cdots, T\}$, the basic idea of this model is to divide the load into two parts in an additive form or in a multiplicative form as follows

$$X(t) = b(t) + z(t) \tag{1.1}$$

$$X(t) = b(t)z(t) \tag{1.2}$$

where $b(t)$ is a base load part which includes a trend component and an average day of the week pattern, but may or may not include the weather sensitive load and $z(t)$ is a stochastic load part. The base load is assumed to be of a recognizable pattern due to normal consumer energy use. The stochastic load is assumed to consist of weather sensitive and random components.

The advantage of the two stage modelling approach is that it simplifies the load modelling by decomposing the relevant time-frames and relevant exogenous factors.

First Stage

There are three major methods of extracting the base load from the observed data. The first method based on the additive form finds the base load which does not include the weather sensitive load. There are then a variety of ways of revealing the base part for different data observation intervals. In principle, this method is based on the shape of the load profile. For instance, we can average the data on basis of type of day in a week, time of a day in an iterative way to obtain the daily levels and the daily profiles which comprise the base load. Some other similar ways to extract the base load can be found in Farmer and Potton (1968), Gupta and Yamada (1972), Metteren and Son (1979), Lijesen and Rosing (1971), Holst and Jonsson (1984), Vemuri et al.

(1986), etc. Abou-Hussien et al. (1981) also developed an iterative way to include the weather sensitive load in this part.

The second method, which is also based on the additive form and is used by Christiaanes (1971), Galiana et al. (1974), Sharma and Mahalanabis (1974) and others, proposes a cyclic function of time with a period of a week to describe the weekly variations in hourly load. The function selected is of the form

$$b(t) = c + \sum_{i=1}^{m} (a_i \sin \omega_i t + b_i \cos \omega_i t) \tag{1.3}$$

that is a Fourier series with m frequencies and $\omega_i = 2\pi K_i/168$ where K_i is a positive integer less than 84, the Nyquist limit, (see Christiaanes (1971)). Moutter et al. (1986a) and Moutter et al. (1986b) argue that the frequencies ω_i may not harmonically related to daily period, and proposed a procedure using the FFT to locate the most likely frequencies.

The third method is based on Box and Jenkins ARIMA schemes, which is in the multiplicative form, to transform the stochastic process of $X(t)$ into a stationary time series by means of a pre-whitening filter as follows

$$\phi(B^s)\nabla^d \nabla_s^D X(t) = \theta(B^s) z(t) \tag{1.4}$$

where s is the length of a one week period; $\phi(B^s)$ and $\theta(B^s)$ are polynomials in B^s; $z(t)$ is assumed to be a zero-mean stationary series.

This method was first used by Stanton et al. (1969) and Gupta (1971) to forecast the medium and long-range load demand of a power system. Vemuri et al. (1973) and many others followed the same approach. It is to be noted that the determination of the order of the filter relies on methods of hypothesis testing and on order determinative statistical criteria used in each investigation.

Second Stage

For the second method in the first stage (see equation (1.3)), the forecast of the base load part is made by exponential smoothing developed by Brown (1965). An extensive analysis has been made by Panuska and Koutchouk (1975) to select the

fitting functions and to suggest reasonable values for the smoothing constant. A drawback of this method is that the accuracy of the forecasts depends heavily on the smoothing constant when the stochastic load is assumed to be white noise and because the latter assumption may not be realistic.

For general cases, the stochastic load part $z(t)$ is assumed to be a stationary process, which is also dependent on the weather condition changes if the weather sensitive component is not included in the base load part.

Spectral Decomposition: Farmer and Potton (1968) use the spectral decomposition approach to expand the stochastic load part into a Karhunen-Loeve spectral decomposition form

$$z(t) = \sum_{k=1}^{K} a_k \lambda_k^{1/2} \vartheta_k(t) + e(t) \tag{1.5}$$

where the eigenvalues λ_k and $\vartheta_k(t)$ are determined by the Karhunen–Loeve integral equation

$$\int_0^T R_z(t,\tau)\vartheta_k(\tau)d\tau = \lambda_k \vartheta_k(t) \tag{1.6}$$

where $R_z(t,\tau)$ is the autocorrelation function of $\{z(t)\}$. The mean-square error takes the form

$$\delta^2 = e^2(t) = R_z(t,t) - \sum_{k=1}^{K} \lambda_k \vartheta_k^2(t) \tag{1.7}$$

By means of this relation, the number of terms, K, in the expansion may be selected to give the required accuracy in the representation of the residual load $z(t)$. The forecast at time t is given by a following recursive algorithm. If the stochastic load is observed at time $t_1, \cdots, t_n$, and if $\hat{x}_{n-1}(t)$ is the most probable value of the load at time t then the most probable value of the load at t is given by

$$\hat{x}_n(t) = \hat{x}_{n-1}(t) + \frac{S_{n-1}(t,t_n)[x(t_n) - \hat{x}(t_n)]}{S_n(t_n,t_n) + \delta^2(t_n)} \tag{1.8}$$

where $S_{n-1}(t,t')$ is a covariance matrix defined by

$$S_n(t,t') = S_{n-1}(t,t') - \frac{S_{n-1}(t,t_n)S_{n-1}(t',t_n)}{S_{n-1}(t_n,t_n) + \delta^2(t_n)} \tag{1.9}$$

with an initial value for the above recurrence equation $x_0(t) = 0$,
$S_0(t,t') = \sum_{k=1}^{K} \lambda_k \vartheta_k(t)\vartheta_k(t')$.

Lijesen and Rosing (1971) argue that Farmer's approach ignored the effect of weather conditions on the stochastic load part, and used an approach similar to that of Farmer but was particularly concerned with the effect of weather conditions on the a_k coefficient of (1.5). A functional relationship is determined between these coefficients and the weather variables so that the coefficient can be estimated based on the weather forecast.

As Farmer has pointed out, the drawback of this approach is that the accuracy of prediction achieved on-line falls far short of that indicated by the off-line work and the forecasting function costs too much in computing time and in computer capacity, mainly for storage. Furthermore, this method is sensitive to the values of the coefficients a_k which cannot be updated optimally in the on-line situation.

Stochastic Time-Series Approach: A common method of modelling the stochastic load part is the Box and Jenkins ARMA model Box and Jenkins (1976). A definitive work on hourly "short-term" load is Holst and Jonsson (1984) in which the stochastic load part is assumed to satisfy,

$$\begin{aligned} &\nabla_{168}\nabla_{24}\nabla X(t) = z(t) && \text{first stage} \\ &z(t) = (1 - c_1 B)(1 - c_2 B^{24})(1 - c_3^{168})e(t) && \text{second stage} \end{aligned} \tag{1.10}$$

where $e(t)$ is zero mean white noise with unknown variance.

A problem with the ARMA models is that weather information is not included explicitly to explain the weather sensitive component of the load. This drawback can be remedied by using a transfer function model developed by Box and Jankins when the weather sensitive component is not included in the base load part

$$\begin{aligned} &z(t) = \frac{\omega(B)}{\eta(B)} W(t-k) + v(t) \\ &\theta(B)v(t) = \phi(B)e(t) \end{aligned} \tag{1.11}$$

where $W(t)$ is the input weather variable; $\omega(B)$, $\eta(B)$, $\theta(B)$, $\phi(B)$ are polynomial function of B and $e(t)$ is a zero-mean white noise with unknown variance.

Hagan and Klein (1977) claim this model is slightly better than the ARMA model in forecasting the load data they used. The reason for the small gain in forecasting is that the transfer function model assumes that the weather variable $W(t)$ is linearly related to the load although there is evidence that load demand and temperature are not linearly related. Irisarri et al. (1982) suggest using a nonlinear relation between temperature and load, recommended by Galiana et al. (1974), to transform temperature data into a weather sensitive load variable which will replace $W(t)$ in the transfer function model (1.11). Campo and Ruiz (1987) argued that humidity along with temperature also plays an important role in affecting the load. They introduced a *Temperature Humidity Index* (THI) to replace $W(t)$. Hagan and Behr (1987) suggest that the relation between load and temperature can be approximately modeled by a polynomial function and they transform temperature (or a temperature humidity index) into a weather sensitive load variable to become an input for the transfer function model. Lu et al. (1989) noticed that the nonlinear relation between temperature and load is time variant. They suggest using a recursive scheme to update a polynomial function and transform temperature into weather sensitive load as an input of the transfer function model.

Another equivalent model representation is the ARMAX model with the form

$$\theta(B)z(t) = \vartheta(B)W(t) + \phi(B)e(t) \tag{1.12}$$

where $W(t)$ is the input weather variable; $\theta(B)$, $\vartheta(B)$ and $\phi(B)$ are polynomial functions of B and $e(t)$ is white noise with zero mean and unknown variance.

$W(t)$ can be replaced by a nonlinearly transformed weather variable, which is linearly related to the load, as mentioned above, to gain better fitting and forecasting (see Galiana et al. (1974) for example).

Although an additional model for the stochastic load part in the second stage can improve forecast accuracy if the model is properly specified, the problem with this approach is that the order of ARMA, transfer function, or ARMAX models are not easy to determine on-line. If there is evidence that the model order is changing with time, then the order should be re-estimated over a reasonable time span. Even if

the order is determined, one cannot automatically identify the model parsimoniously when the stochastic load part appears to need a "high" order. In most practical circumstances, an empirical model, which is of a parsimonious form but may not be adequate, is imposed. Some on-line recursive algorithms such as *on-line maximum likelihood estimation* (see Hagan and Klein (1978)) etc. are developed to update the parameters of the assumed model form but not the model order.

Rajurkar and Nissen (1985) use a *Date-Dependent Systems* approach developed by Pandit and Wu (1983) to obtain a model automatically, where weather information is not involved. The model identification procedure is carried out by successively fitting models to the data in higher order, n, of ARMA$(n, n-1)$ form, until there is no significant gain in fitting error as judged by certain criteria. In our opinion, this approach can lead to over-parameterization because of the restricted model form.

1.2.2 State Space Model Approach

Many researchers prefer to use a state space model for load demand. because they can then use powerful the Kalman filtering (see Kalman (1960) and Kalman and Bucy (1961)) to obtain on-line optimum forecasts which are otherwise difficult to realize by most time series and spectral decomposition approaches. A state space model has a general form as follows

$$\begin{cases} x(t+1) = A(\theta,t)x(t) + B(\theta,t)u(t) + \xi(t) & \text{Transition equation} \\ y(t) = C(\theta,t)x(t) + D(\theta,t)v(t) + \epsilon(t) & \text{Observation equation} \end{cases} \tag{1.13}$$

where (i) $\{x(t), t \geq 1\}$ is a sequence of $m \times 1$ state vectors; $y(t)$ is an observed series; $A(\theta,t)$, $B(\theta,t)$, $C(\theta,t)$, $D(\theta,t)$ are called system matrices. $u(t)$ and $v(t)$ are series exogenous to the system. (ii) $\forall t \in T^+$; $\xi(t)$ and $\epsilon(t)$ are respectively $m \times 1$ and 1×1 Gaussian random disturbances and they are conditionally independent of $x(0)$ and $\{y(\tau) | \tau < t\}$. $\forall t,\ s \in T^+$, both $\xi(t)$ and $\epsilon(t)$ have zero mean and with variance-covariance matrices.

$$E\left\{ \begin{pmatrix} \xi(s) \\ \epsilon(s) \end{pmatrix} (\xi(t), \epsilon(t)) \right\} = \begin{pmatrix} Q & S \\ S' & R \end{pmatrix} \delta(t-s) \tag{1.14}$$

and R and Q are constant definite and semi-definite matrices respectively.

A preliminary study of the load prediction problem through state space modelling and the Kalman filtering has been reported by Toyoda et al. (1970a) and Toyoda et al. (1970b). In hourly or daily load forecasting, the daily periodical load pattern and the effect of the weather condition on load are involved. Thus, they suggest a model with following state space form

$$\begin{cases} \begin{pmatrix} X(t+1) \\ \Delta(t+1) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & \alpha(t) \end{pmatrix} \begin{pmatrix} X(t) \\ \Delta(t) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \beta(t) & r(t) \end{pmatrix} \begin{pmatrix} T(t) \\ H(t) \end{pmatrix} \\ \qquad\qquad + \begin{pmatrix} v_1(t+1) \\ v_2(t+1) \end{pmatrix} \\ \\ y(t) = (S(t)\ 1) \begin{pmatrix} X(t) \\ \Delta(t) \end{pmatrix} + e(t) \end{cases} \tag{1.15}$$

where $X(t)$ is the pseudo-daily peak, or daily average load, $\Delta(t)$ is the load fluctuation because of weather conditions, temperature $T(t)$ and humidity $H(t)$, and $S(t)$ is the coefficient of daily standard load pattern ($S(t) \approx S(t+D)$, $D = 24$ hours). $\alpha(t)$, $\beta(t)$, $r(t)$, and $S(t)$ can be estimated using past observations.

A state space model also has a most important property that is that any time series stochastic model, such as ARIMA, ARMAX, exponential smoothing, etc., has a state space representation (see Kailath (1980) Hannan and Deistler (1988) for details). Therefore, some researchers convert some well established time series models mentioned in the last sub-section into state space models, then use the Kalman filtering techniques to estimate model parameters, and obtain on-line estimation and forecasting.

Singh et al. (1977) converted an AR model for the load demand into a state space model representation. Sharma and Mahalanabis (1974) used the Kalman filter to update the coefficients of a harmonic based model suggested by Christiaanes (1971). Galiana et al. (1974) also consider a model, in which the base load is modeled by a harmonic model (1.3) and the stochastic part modelled by an ARX model, in a

state space model form. Irisarri et al. (1982), and Campo and Ruiz (1987) converted a model which comprises a base part which modelled particularly time-of-day and day-of-week variation, and a stochastic part modelled by an ARX in a state space form.

The advantages of using a state space model come not only from the realization of on-line estimation and forecasting through the Kalman filtering but also in the dynamic linkage between the base load part and the stochastic load part in a state space framework. The base load now can be updated adaptively through a Kalman filter gain which is affected by the stochastic load (difference between observed load value and the base load) and the weather sensitive load.

A difficult problem in state space modelling is to determine the order of a model. Many researchers specify the order on the basis of their experience. For univariate time series, Akaike (1975) and Aoki (1987) developed a methodology where the form of the system matrices of a state space model are totally unknown. By using the Hankel matrix of the time series, and its principal components, this approach can determine the order of the state space model and then the model matrices in a canonical form. The merit of this approach is that the form of a state space model (system matrices) need not be pre-specified. The main drawback is that this approach may not be suitable for a high order system because every element of the system matrices represents an unknown parameter and there are therefore too many parameters to be estimated. In addition, the state vector generated by this approach has no physical interpretation. It is also difficult to include the weather variables in such a model for electricity load series.

To overcome this drawback, Harvey and Todd (1983) and Gersch and Kitagawa (1983) developed a *structural state space* modelling approach. This approach assumes that a time series consists of several components, such as trend, seasonal, and disturbance, etc. in an additive form, and each component has its own *micro* state space representation. This approach dynamically puts these *micro* representations into a main state space model framework which can reflect the interactive effects between

the different components.

Once the structure of a state space model has been properly specified, the most difficult problem in estimating the model parameters is the specification or estimation of the initial state vector and corresponding covariance matrix, and the covariance matrices of disturbance terms because they substantially affect the Kalman filtering (value of the Kalman filter gain) and then consequently the parameter estimation. Many researchers fail to address this problem or assume they are known *a priori*, and rely on the Kalman filter to converge to the covariance matrix of the state vector. The covariance matrices of disturbance terms, which affect the convergence speed and the adaptability of the model, are simply ignored or specified by experience without adjustment in a statistical manner.

1.3 Summary

Choosing a proper model form is the most important and crucial part of the load forecasting procedure. In other words the process of choosing a class of models which represents the load data is more important than the techniques, which may affect the parameter estimation, used for estimating the model parameters and minimizing a certain criterion.

As we reviewed above, most existing models in the literature assume that the weather sensitive load lies in the stochastic part. We believe that dividing the load into weather insensitive and weather sensitive components is a very fruitful idea. This simplifies the modelling effort for each component and allows more flexibility in representing and interpreting the load demand. A dynamic relation between the load and the weather insensitive, and between the load and the weather sensitive components should be considered according to their "energy" distribution (that is state) to the load. Furthermore, the accuracy of the forecast can be improved by making the model adaptive to the unknown changes.

The structure of this thesis is divided into two parts. The first part consists of

chapter 2 and 3. Some theoretical work related to subset AR selection procedure and a new subset AR selection procedure is presented in chapter 2. In chapter 3, for state space models, the effect of an initial state covariance matrix on its convergence speed is studied and we establish a new procedure to estimate the initial state vector and its covariance. Associated sensitivity analysis is also presented in chapter 3. The development in this part are placed in a general context but will be used in the later chapters for the practical electricity load forecasting models. The second part includes chapters 4, 5, 6 and 7 where we concentrate on practical application of short-term load modelling and forecasting.

In the *non-weather variable model* category, we develop a new base load modelling procedure which takes account of the day-of-week effect in a proper way in chapter 4, and the stochastic load is modelled by a subset AR model which is selected by the procedure developed in chapter 2. This new procedure overcomes the over-parameterization problems, and exhibits some new features in searching for an optimal subset AR model. The comparison of the forecasting performance for the proposed model with other popular models for New Zealand half-hourly short-term electricity load data is presented in chapter 5.

In the *weather variable model* category, we first establish a non-linear statistical functional relation between the load and weather conditions and then proceed with model estimation. The accuracy and stability of the model is also studied in chapter 6. From this proposed function, we extract a weather sensitive variable which is linearly related to the load, and regard it as the weather sensitive load component. In chapter 7, we develop a structural state space model in which the proposed weather sensitive variable is employed as an exogenous variable in a linear system model of the daily load data for the Canberra region in Australia. When identifying the proposed model, some theoretical work established in chapter 3 is employed for our new practical procedures which also addresses the initial estimation of the state vector and its covariance matrix, and the covariance matrix of the disturbance terms. A sequence of hypothesis tests is conducted to specify the final model form and the structure

of the disturbance covariance matrix. The on-line sequential updating procedure will not be discussed in this thesis because it is straight forward and would increase unduly the size of this thesis. We present our conclusions and discussion of suggested directions for future research in chapter 8.

Chapter 2

On the Selection of Subset AR Model

2.1 Introduction

In the analysis of stationary time series, the use of autoregressive(AR) models has played a pivotal role. Amongst linear time series models, autoregressive models are the simplest to estimate and may easily be used for forecasting purposes.

A zero-mean stationary stochastic process $\{X(t)\}$ is said to be generated by an autoregressive model of order k, denoted by AR(k), if it satisfies the stochastic difference equation

$$X(t) + a(1)X(t-1) + \cdots + a(k)X(t-k) = \epsilon(t) \tag{2.1}$$

where $\{\epsilon(t)\}$ is a Gaussian white noise process with variance σ^2. Using the backward shift operator, the equation (2.1) may be written as

$$\Phi(B)X(t) = \epsilon(t)$$

where

$$\Phi(z) = 1 + a(1)z + \cdots + a(k)z^k$$

The condition for the time series $\{X(t)\}$, satisfying equation (2.1), to be stationary is that all the roots of $\Phi(z) = 0$ lie outside the unit circle.

In general, when a model of the form (2.1) is fitted to a set of observations on a stationary time series $\{X(t)\}$, the fitted model will include all the terms $\{X(t-i); i =$

$1, 2, \cdots, k\}$. In many situations, in particular where there may be evidence that a time series may have some form of seasonal behaviour, this may lead to models which include many more parameters than are strictly necessary to describe its behaviour. It is often desirable to use models of the form (2.1) where some of the $\{a(i)\}$ are set equal to zero. Such models are referred to as subset autoregressive time series models.

A major problem in fitting autoregressive models, even of full order, has always been the choice of the order of the model. Consequently, many authors, Akaike (1974), Box and Jenkins (1976), Hannan (1970), Parzen (1974), to mention a few, have paid special attention to this problem. The criteria for the choice of the AR model order is not discussed here. We suppose the optimum order, k, of the AR model is known.

The principal idea of an optimum subset AR model selection procedure is to allocate the best subset AR models for all sizes at the first stage and then to use a model criterion to select the optimum subset AR model from those best models. Residual variance is a common statistic used to measure the goodness of fit of a candidate model. Therefore, it is also a criterion used to find the best subset AR model of a particular size. Without considering the computing efficiency, one can always find the best subset AR model of a specified size by comparing the residual variances among the subset AR models with all possible lag combinations and selecting the one with smallest residual variance.

A more efficient procedure to select an optimum subset AR model was developed by adapting Hocking and Leslie's algorithm (see Hocking and Leslie (1967)) for an optimum subset regression model. McClave (1975) has extended Hocking and Leslie's algorithm to time series models to give a method of fitting subset autoregressive models. This method provides a fast procedure for selecting the best subset without having to evaluate all possible lag combinations for a specified size.

In section 2.2 of this chapter, we first analyse the properties of a subset AR model, and the effect of a removed lag on the remaining lags which compose a subset AR model with reduced size by 1, and then construct a new statistic to measure the goodness of fit of a candidate subset AR model which place emphasis on the

representability of the removed lag by the remaining lags. By using this new statistic, we develop a new algorithm to search for the optimum subset AR model in section 2.3. Some numerical examples presented in section 2.4 prove that the proposed procedure is considerably more efficient than McClave's algorithm in the sense that less subset AR models need to be checked to find the optimum subset AR model for a given data set.

2.2 Properties of Subset AR Models

Before discussing the selection procedure for subset AR models, we examine the properties of a subset AR model. Suppose, in a subset AR model with size m and included lags $\{i_1, i_2, \cdots, i_m\}$ and maximum lag k, the estimated AR coefficients are the solution of the following subset Yule-Walker equation

$$R_{(i_1,i_2,\cdots,i_m)}\ a_{(i_1,i_2,\cdots,i_m)} = r_{(i_1,i_2,\cdots,i_m)} \tag{2.2}$$

where

$$R_{(i_1,i_2,\cdots,i_m)} = (r_{(i_1,i_2,\cdots,i_m)}(i,j))_{k\times k}$$

$$R_{(i_1,i_2,\cdots,i_m)}(i,j) = \begin{cases} r(i,j) & i,j \in (i_1,i_2,\cdots,i_m) \\ 1 & i=j, \text{ and } i \notin (i_1,i_2,\cdots,i_m) \\ 0 & \text{otherwise} \end{cases}$$

$$r_{(i_1,i_2,\cdots,i_m)} = (r_{(i_1,i_2,\cdots,i_m)}(i))_{k\times 1}$$

$$r_{(i_1,i_2,\cdots,i_m)}(i) = \begin{cases} -r(i) & i \in (i_1,i_2,\cdots,i_m) \\ 0 & i \notin (i_1,i_2,\cdots,i_m) \end{cases}$$

where $r(i) = r(-i)$ is lag k autocorrelation and $r(i,j) = r(i-j)$.

The above subset Yule-Walker equations can be regarded as a transformation arising from $R_{(i_1,i_2,\cdots,i_m)}$ applied to $a_{(i_1,i_2,\cdots,i_m)}$ to produce $r_{(i_1,i_2,\cdots,i_m)}$. $R_{(i_1,i_2,\cdots,i_m)}$ forces the AR coefficients of the lags complementing $(i_1, i_2, \cdots, i_m)$ to be set equal to zero, i.e. $R_{(i_1,i_2,\cdots,i_m)}$ projects the effects of the lags which are complementary to $(i_1, i_2, \cdots, i_m)$ onto the lags $(i_1, i_2, \cdots, i_m)$. The increase in residual variance of the reduced size

subset AR model is caused by the effect of the lags which are complementary to the lag set $(i_1, i_2, \cdots, i_m)$ that cannot be represented by the $(i_1, i_2, \cdots, i_m)$ lags. In other words, the deletion effect can be expressed by the increase in residual variance. The optimum subset AR with size m is chosen as a subset of size m, among the $\begin{pmatrix} k \\ m \end{pmatrix}$ candidates, with the minimum deletion effect arising from the loss of the set complimentary to the included set $(i_1, i_2, \cdots, i_m)$. Therefore, the increase in residual variance due to the exclusion of the lag set which is complimentary to $(i_1, i_2, \cdots, i_m)$ in a AR model serves as the criterion for the optimal choice of a subset AR model of size m.

In general, the residual variance of the subset AR is

$$W_{(i_1,i_2,\cdots,i_m)} = C(0)[1 - r'_{(i_1,i_2,\cdots,i_m)} R^{-1}_{(i_1,i_2,\cdots,i_m)} r_{(i_1,i_2,\cdots,i_m)}] \tag{2.3}$$

where $C(0)$ is the variance of the raw data set.

The term $r'_{(i_1,i_2,\cdots,i_m)} R^{-1}_{(i_1,i_2,\cdots,i_m)} r_{(i_1,i_2,\cdots,i_m)}$ is the measure of R^2 for the subset AR fitting and represents the goodness of fit for the subset AR model.

Now, we examine the effects of reducing the subset AR size by one on the corresponding R^2. Without loss of generality, we suppose that the AR coefficient on the i_mth lag is set to zero. The R^2 of the reduced AR subset model is

$$r'_{(i_1,i_2,\cdots,i_{m-1})} R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} r_{(i_1,i_2,\cdots,i_{m-1})}$$

We define

$$Q_{(i_m)} = \begin{cases} r(i,j) & i \neq j,\ i = i_m,\ \text{or } j = i_m \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Since $Q_{(i_m)}$ is a symmetric matrix with rank 2, there exists a full rank matrix,C, with dimension $k \times 2$ which satisfies

$$Q_{(i_m)} = CBC'$$

where B is a full rank matrix with dimension 2×2.

It is easy to verify that one of the choices for C and B is

$$C(i,1) = \begin{cases} \frac{r(i_m - i)}{2} & i \in (i_1, i_2, \cdots, i_{m-1}) \\ 1 & i = i_m \\ 0 & i \notin (i_1, i_2, \cdots, i_{m-1}) \end{cases}$$

$$C(i,2) = \begin{cases} -C(i,1) & i \neq i_m \\ 1 & i = i_m \end{cases}$$

for $i = 1, 2, \cdots, k$.

$$B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

We know that the coefficients of the reduced subset AR model are

$$a_{(i_1,i_2,\cdots,i_{m-1})} = R^{-1}_{(i_1,i_2,\cdots,i_{m-1})}\, r_{(i_1,i_2,\cdots,i_{m-1})}$$

and

$$\begin{aligned} R_{(i_1,i_2,\cdots,i_{m-1})} &= R_{(i_1,i_2,\cdots,i_{m-1},i_m)} - Q_{(i_m)} \\ &= R_{(i_1,i_2,\cdots,i_{m-1},i_m)} - CBC' \end{aligned}$$

From the matrix inversion lemma,

Lemma 2.1

$$(A + CBC')^{-1} = A^{-1} - A^{-1}C'[B^{-1} + CA^{-1}C']^{-1}CA^{-1}$$

for any matrices A, B, C if A^{-1}, B^{-1} and $(A + CBC')^{-1}$ exist. □

We, therefore, have the following expression for $R^{-1}_{(i_1,i_2,\cdots,i_{m-1})}$

$$R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} = R^{-1}_{(i_1,i_2,\cdots,i_{m-1},i_m)} - H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})} \tag{2.5}$$

where

$$H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})} = R^{-1}_{(i_1,i_2,\cdots,i_m)}\, C[-B^{-1} + CR^{-1}_{(i_1,i_2,\cdots,i_m)}C']^{-1}C\, R^{-1}_{(i_1,i_2,\cdots,i_m)} \tag{2.6}$$

Therefore,

$$\begin{aligned} a_{(i_1,i_2,\cdots,i_{m-1})} &= R^{-1}_{(i_1,i_2,\cdots,i_m)}\, r_{(i_1,i_2,\cdots,i_{m-1})} - H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})} r_{(i_1,i_2,\cdots,i_{m-1})} \\ &= a_{(i_1,i_2,\cdots,i_m)} - \{R^{-1}_{(i_1,i_2,\cdots,i_m)} r_{(i_m)} + H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})} r_{(i_1,i_2,\cdots,i_{m-1})}\} \end{aligned} \quad (2.7)$$

From the above equation, it can be seen that the term

$$\{R^{-1}_{(i_1,i_2,\cdots,i_m)} r_{(i_m)} + H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})} r_{(i_1,i_2,\cdots,i_{m-1})}\}$$

is actually the i_mth lag contribution to the coefficients of the subset AR model with lag set $(i_1, i_2, \cdots, i_m)$. Therefore, the effect of the reduction of the subset AR size by one is obtained as the contribution of the removed lag i_m and we see that contribution is subtracted from the coefficients of the subset AR model before the i_mth lag is removed. From equation (2.7), the modification of the value of the reduced subset AR coefficients is given by,

$$\begin{aligned} a_{(i_1,i_2,\cdots,i_{m-1})}(i) &= a_{(i_1,i_2,\cdots,i_m)}(i) + R^{-1}_{(i_1,i_2,\cdots,i_m)}(i, i_m) r(i_m) \\ &\quad - \sum_{l=1}^{k} H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}(i, l) r_{(i_1,i_2,\cdots,i_{m-1})}(l) \end{aligned} \quad (2.8)$$

where $i \in (i_1, i_2, \cdots, i_{m-1})$ and from $a_{(i_1,i_2,\cdots,i_{m-1})}(i_m) = 0$. We have,

$$\begin{aligned} &a_{(i_1,i_2,\cdots,i_m)}(i_m) = \\ &\quad -R^{-1}_{(i_1,i_2,\cdots,i_m)}(i_m, i_m) r(i_m) + \sum_{l=1}^{k} H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}(i_m, l) r_{(i_1,i_2,\cdots,i_{m-1})}(l) \end{aligned} \quad (2.9)$$

and

$$\sum_{l=1}^{k} H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}(j, l) r_{(i_1,i_2,\cdots,i_{m-1})}(l) = 0 \quad (2.10)$$

where $j \notin (i_1, i_2, \cdots, i_m)$.

The R^2 of the reduced subset AR model is

$$\begin{aligned} &r'_{(i_1,i_2,\cdots,i_{m-1})} R^{-1}_{(i_1,i_2,\cdots,i_{m-1})}\, r_{(i_1,i_2,\cdots,i_{m-1})} \\ &\quad = r'_{(i_1,i_2,\cdots,i_{m-1})} R^{-1}_{(i_1,i_2,\cdots,i_m)}\, r_{(i_1,i_2,\cdots,i_{m-1})} \\ &\qquad - r'_{(i_1,i_2,\cdots,i_{m-1})} H^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}\, r_{(i_1,i_2,\cdots,i_{m-1})} \end{aligned}$$

The complexity of the effects of AR size reduction by one makes theoretical analysis and practical evaluation of the modification of the subset AR coefficients and the goodness of fit measure, R^2, very difficult since the inversion of $R_{(i_1,i_2,\cdots,i_m)}$ is involved. Many authors in the subset AR literature, such as McClave (1975), successfully used numerical means to avoid calculating $R^{-1}_{(i_1,i_2,\cdots,i_m)}$. Although those numerical algorithms are very efficient in evaluating the increase in residual variance (equivalent to R^2) due to a reduction of AR size, the difficulties for theoretical analysis remain.

2.2.1 The Property of Subset AR Models in Hilbert Space

The above analysis can be combined with the theory of linear system analysis. So that now, we consider $(\{x(t)\},\{x(t-1)\},\cdots,\{x(t-k)\},\{\epsilon(t)\})$ which generate a linear space $\mathbf{P}_k$. Bearing in mind that the basic elements $\{x(t)\}$ are random variables, the dimension of the space $\mathbf{P}_k$ is infinite and its basic elements are neither orthogonal to nor independent of each other which is different from the multiple regression model where the variates are often assumed independent. For simplicity, we define $\Theta(l) = \{x(t-l)\}$ $l = 0,1,2,\cdots,k$, $\epsilon = \{\epsilon(t)\}$ and an inner product for any two elements, $\alpha,\beta \in \mathbf{P}_k$ as

$$\langle \alpha,\beta\rangle = \mathbf{cov}(\alpha,\beta) \tag{2.11}$$

and a distance as defined by

$$d(\alpha,\beta) = \langle \alpha-\beta,\alpha-\beta\rangle \tag{2.12}$$

It is easy to verify that the $\mathbf{P}_k$ is a "random" Hilbert space. The basic element $\Theta(0)$ is not independent of the remaining basic elements. i.e. there exists a numerical vector $a_k = (a_k(1),a_k(2),\cdots,a_k(k))$ which satisfies

$$\Theta(0) = a_k(1)\Theta(1) + a_k(2)\Theta(2) + \cdots + a_k(k)\Theta(k) + \epsilon \tag{2.13}$$

Forming appropriate inner products with $\Theta(i), i = 1,2,\cdots,k$ on both sides of the above equation, we have the Yule-Walker equations,

$$R_k a_k = r_k \tag{2.14}$$

where $R_k = (r(i,j))_{k\times k}$, $r(i,j) = C(|i-j|)/C(0), r_k(i) = C(i)/C(0)$, $i,j = 1,2,\cdots,k$. We can, without confusion, delete the subscript k so that $R_k = R$, $a_k = a$, $r_k = r$

To fit a subset AR model of fixed subset size m is to choose a subset of lags $(i_1, i_2, \cdots, i_m)$ and restrict to zero the AR coefficients associated with lags that compliment $(i_1, i_2, \cdots, i_m)$. On the other hand, the basic elements of $\mathbf{P}_k$ corresponding to those AR coefficients restricted to zero are forced to be linearly represented by $\mathbf{P}_m(i_1, i_2, \cdots, i_m)$ which is a linear random subspace generated by $\Theta(i_1), \cdots, \Theta(i_m)$. i.e. there exists a matrix, C, with dimension $k\times(m+1)$ satisfying the relation

$$\begin{pmatrix} \Theta(1) \\ \Theta(2) \\ \vdots \\ \Theta(k) \\ \epsilon \end{pmatrix} = C \begin{pmatrix} \Theta(i_1) \\ \Theta(i_2) \\ \vdots \\ \Theta(i_m) \\ \epsilon_m \end{pmatrix}$$

Therefore,

$$\begin{aligned} \Theta(0) &= a(\Theta(1), \cdots, \Theta(k), \epsilon)' \\ &= aC(\Theta(i_1), \cdots, \Theta(i_m), \epsilon) \\ &= \tilde{a}(i_1)\Theta(i_1) + \cdots + \tilde{a}(i_m)\Theta(i_m) + c_m\epsilon_m \end{aligned}$$

Forming inner products with $\Theta(i_l)$, $l = 1,2,\cdots,m$ on both sides of the above equation, we have

$$\begin{pmatrix} r(i_1) \\ r(i_2) \\ \vdots \\ r(i_m) \end{pmatrix} = \begin{pmatrix} 1 & \cdots & r(i_1 - i_m) \\ r(i_2 - i_1) & \cdots & r(i_2 - i_m) \\ \cdots & \cdots & \cdots \\ r(i_m - i_1) & \cdots & 1 \end{pmatrix} \begin{pmatrix} \tilde{a}(i_1) \\ \tilde{a}(i_2) \\ \cdots \\ \tilde{a}(i_m) \end{pmatrix} \tag{2.15}$$

The above equation is equivalent to

$$r_{(i_1,\cdots,i_m)} = R_{(i_1,\cdots,i_m)}\tilde{a}_{(i_1,\cdots,i_m)}$$

where

$$r_{(i_1,\cdots,i_m)}(i) = \begin{cases} r(i) & i \in (i_1, i_2, \cdots, i_m) \\ 0 & i \notin (i_1, i_2, \cdots, i_m) \end{cases}$$

and

$$R_{(i_1,\cdots,i_m)}(i,j) = \begin{cases} r(i-j) & i,j \in (i_1,i_2,\cdots,i_m) \\ 1 & i=j \text{ and } i,j \notin (i_1,i_2,\cdots,i_m) \\ 0 & \text{otherwise} \end{cases}$$

which is exactly the subset Yule-Walker equation (2.2).

The optimum subset AR with size m arises when we choose a lag subset, $(i_1,\cdots,i_m)$, from the lag set $(1,\cdots,k)$ which achieves the minimum value of $c_m^2\langle\epsilon_m,\epsilon_m\rangle$ from the $\begin{pmatrix} k \\ m \end{pmatrix}$ candidate models. On the other hand, it implies that the basic elements of $\mathbf{P}_k$ corresponding to the complement of $(i_1,\cdots,i_m)$ can be "almost" linearly represented by $\mathbf{P}_m(i_1,\cdots,i_m)$.

Now, we examine the effects of reducing the size by 1 in a subset AR model. Suppose we have a size m subset AR Model with maximum lag k,

$$X(t) = \sum_{j=1}^{m} a_{(i_1,i_2,\cdots,i_m)}(i_j)X(t-i_j) + \epsilon(t) \tag{2.16}$$

Equivalently, the model (2.16) has an expression in $\mathbf{P}_m(i_1,i_2,\cdots,i_m)$ of the random Hilbert space $\mathbf{P}_k$

$$\Theta(0) = \sum_{j=1}^{m} a_{(i_1,i_2,\cdots,i_m)}(i_j)\Theta(i_j) + c_m\epsilon_m \tag{2.17}$$

There is a linkage between the estimates of a subset AR model of size m and a subset AR model of size $(m-1)$ which is obtained by setting one coefficient to zero. Without loss of generality, we restrict $a_{(i_1,i_2,\cdots,i_m)}(i_m)$ to zero to form a size $(m-1)$ subset AR model. The linkage is dependent upon an auxiliary relation, i.e. $\Theta(i_m)$ projects on $\mathbf{P}_{m-1}(i_1,i_2,\cdots,i_{m-1})$, and is shown as follows

$$\Theta(i_m) = \Theta(i_m)|_{\mathbf{P}_{m-1}(i_1,i_2,\cdots,i_{m-1})} + e_{i_m} \tag{2.18}$$

where

$$\Theta(i_m)|_{\mathbf{P}_{m-1}(i_1,i_2,\cdots,i_{m-1})} = \sum_{j=1}^{m-1} b(i_j)\Theta(i_j) \tag{2.19}$$

and

$$e_{i_m} \perp \mathbf{P}_{m-1}(i_1,i_2,\cdots,i_{m-1}), \quad e_{i_n} \perp \epsilon_m$$

Therefore, the linear expression linking $\Theta(0)$ and $\mathbf{P}_{m-1}(i_1, i_2, \cdots, i_{m-1})$ is

$$\Theta(0) = \sum_{j=1}^{m-1} a_{(i_1,i_2,\cdots,i_{m-1})}(i_j)\Theta(i_j) + \epsilon_{m-1} \tag{2.20}$$

where $\epsilon_{m-1} \perp \mathbf{P}_{m-1}(i_1, i_2, \cdots, i_{m-1})$ and we have,

$$\begin{aligned}\Theta(0) &= \sum_{j=1}^{m-1} \left\{a_{(i_1,i_2,\cdots,i_m)}(i_j) + a_{(i_1,i_2,\cdots,i_m)}(i_m)b(i_j)\right\}\Theta(i_j) \\ &\quad + \left\{a_{(i_1,i_2,\cdots,i_m)}(i_m)e_{i_m} + c_m\epsilon_m\right\}\end{aligned} \tag{2.21}$$

From equations (2.20) and (2.21), we clearly see that the coefficients of the subset AR model of size $(m-1)$ and the coefficients of the subset AR model of size m have the following relation

$$a_{(i_1,i_2,\cdots,i_{m-1})}(i_j) = a_{(i_1,i_2,\cdots,i_m)}(i_j) + a_{(i_1,i_2,\cdots,i_m)}(i_m)b(i_j) \tag{2.22}$$

where $j = 1, \cdots, (m-1)$.

Therefore, having obtained the coefficients of the subset AR model of size m, estimates of the coefficients of the subset AR model of size $(m-1)$ require only the estimates of the coefficients of the regression of $\Theta(i_m)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$.

Furthermore, the equation (2.18) can be expressed as

$$\Theta(i_m) = \sum_{j=1}^{m-1} b(i_j)\Theta(i_j) + e_{i_m} \tag{2.23}$$

Taking inner products with $\Theta(i_j)$, $(j = 1, 2, \cdots, m-1)$ on both sides of equation (2.23) yields

$$\begin{pmatrix} r(i_m - i_1) \\ \vdots \\ r(i_m - i_{m-2}) \\ r(i_m - i_{m-1}) \end{pmatrix} = \begin{pmatrix} 1 & \cdots & r(i_1 - i_m) \\ r(i_2 - i_1) & \cdots & r(i_2 - i_m) \\ \cdots & \cdots & \cdots \\ r(i_{m-1} - i_1) & \cdots & 1 \end{pmatrix} \begin{pmatrix} b(i_1) \\ b(i_2) \\ \cdots \\ b(i_{m-1}) \end{pmatrix} \tag{2.24}$$

The equation (2.24) is equivalent to the following equation

$$r^{(i_m)}_{(i_1,\cdots,i_{m-1})} = R_{(i_1,\cdots,i_{m-1})}b_{(i_1,\cdots,i_{m-1})} \tag{2.25}$$

where

$$r^{(i_m)}_{(i_1,\cdots,i_{m-1})}(j) = \begin{cases} r(i_m - j) & j \in (i_1,\cdots,i_{m-1}) \\ 0 & \text{otherwise} \end{cases}$$

$$b_{(i_1,\cdots,i_{m-1})} = \begin{cases} b(j) & j \in (i_1,\cdots,i_{m-1}) \\ 0 & \text{otherwise} \end{cases}$$

The linear regression coefficients of lag i_m on lag j is $b(j)$ where $j \in (i_1,\cdots,i_{m-1})$ which is the solution of equation (2.25).

From equation (2.21), the associated increase in residual variance is, therefore,

$$\begin{aligned} & \sigma^2_{(i_1,\cdots,i_{m-1})}(i_m) \\ &= \langle a_{(i_1,\cdots,i_m)}(i_m)e_{i_m}, a_{(i_1,\cdots,i_m)}(i_m)e_{i_m}\rangle \\ &= a^2_{(i_1,\cdots,i_m)}(i_m)\langle e_{i_m}, e_{i_m}\rangle \\ &= a^2_{(i_1,\cdots,i_m)}(i_m)\langle \Theta(i_m) - \sum_{j=1}^{m-1} b(i_j)\Theta(i_j), \Theta(i_m) - \sum_{j=1}^{m-1} b(i_j)\Theta(i_j)\rangle \\ &= a^2_{(i_1,\cdots,i_m)}(i_m)C(0)(1 - [r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]' R^{-1}_{(i_1,\cdots,i_{m-1})}[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]) \\ &= A_{(i_1,\cdots,i_m)}(i_m)E_{(i_1,\cdots,i_m)}(i_m) \end{aligned} \tag{2.26}$$

where $A_{(i_1,\cdots,i_m)}(i_m) = a^2(i_1,\cdots,i_m)(i_m)$,

$$E_{(i_1,\cdots,i_m)}(i_m) = C(0)(1 - [r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]' R^{-1}_{(i_1,\cdots,i_{m-1})}[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}])$$

using equation (2.25).

There are m subset AR models of size $(m-1)$ which can be chosen from the lag set of $(i_1,\cdots,i_m)$. The lags $(i_1,\cdots,i_{m-1})$ are chosen as the optimum subset AR of size $(m-1)$ only if $\sigma^2_{(i_1,\cdots,i_{m-1})}(i_m)$ is the minimum among those m candidates. The increase in residual variance due to the setting of $a_{(i_1,\cdots,i_{m-1})}(i_m)$ to zero depends on both the value of $a^2_{(i_1,\cdots,i_m)}(i_m)$, which is estimated before lag i_m is removed, and the value of $[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]' R^{-1}_{(i_1,\cdots,i_{m-1})}]\ [r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]$.

Since we have,

$$R_{(i_1,\cdots,i_m)} = \begin{pmatrix} R_{(i_1,\cdots,i_{m-1})} & r^{(i_m)}_{(i_1,\cdots,i_{m-1})} \\ (r^{(i_m)}_{(i_1,\cdots,i_{m-1})})' & 1 \end{pmatrix} \tag{2.27}$$

and suppose we define the inverse as,

$$R^{-1}_{(i_1,\cdots,i_m)} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \tag{2.28}$$

Then utilizing the following matrix lemma:

Lemma 2.2 *If square matrix A is non-singular and*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

A_{11} *and* A_{22} *are nonsingular, then,*

$$A^{-1} = \begin{pmatrix} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & -(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix}$$

□

we have

$$B_{11} = \{R_{(i_1,\cdots,i_{m-1})} - r^{(i_m)}_{(i_1,\cdots,i_{m-1})}[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]'\}^{-1} \tag{2.29}$$

$$B_{22} = \{1 - [r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]'R^{-1}_{(i_1,i_2,\cdots,i_{m-1})}r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}\}^{-1} \tag{2.30}$$

and then, from lemma 2.1, we have

$$\begin{aligned} R^{-1}_{(i_1,\cdots,i_{m-1})} &= \{B_{11}^{-1} + r^{(i_m)}_{(i_1,\cdots,i_{m-1})}[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]'\}^{-1} \\ &= B_{11} - B_{11}r^{(i_m)}_{(i_1,\cdots,i_{m-1})}\{I + [r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]'B_{11}r^{(i_m)}_{(i_1,\cdots,i_{m-1})}\}^{-1} \\ &\quad \times [r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]'B_{11} \end{aligned} \tag{2.31}$$

Therefore, if we pre-multiply by $[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]'$ and post-multiply by $r^{(i_m)}_{(i_1,\cdots,i_{m-1})}$ on both sides of the above equation, we have

$$[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]'R^{-1}_{(i_1,\cdots,i_{m-1})}[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}] = S(i_m) - \frac{S^2(i_m)}{1+S(i_m)} = \frac{S(i_m)}{1+S(i_m)} \tag{2.32}$$

where $S(i_m) = [r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]'B_{11}r^{(i_m)}_{(i_1,\cdots,i_{m-1})}$

It is then easy to verify that

$$[r^{(i_m)}_{(i_1,\cdots,i_{m-1})}]' B_{11} r^{(i_m)}_{(i_1,\cdots,i_{m-1})} = [(r^{(i_m)}_{(i_1,\cdots,i_{m-1})}, 0)]' R^{-1}_{(i_1,\cdots,i_m)} (r^{(i_m)}_{(i_1,\cdots,i_{m-1})}, 0) \tag{2.33}$$

so that

$$S(i_m) = [\tilde{r}^{(i_m)}_{(i_1,\cdots,i_m)}]' R^{-1}_{(i_1,\cdots,i_m)} \tilde{r}^{(i_m)}_{(i_1,\cdots,i_m)} \tag{2.34}$$

where $\tilde{r}^{(i_m)}_{(i_1,\cdots,i_m)} = \begin{pmatrix} r^{(i_m)}_{(i_1,\cdots,i_{m-1})} \\ 0 \end{pmatrix}$.

Therefore, we have the following theorem:

Theorem 2.1 *Suppose the AR coefficients of a size m subset AR model* $X(t) + a_{(i_1,\cdots,i_m)}(i_1)X(t-i_1) + \cdots + a_{(i_1,\cdots,i_m)}(i_m)X(t-i_m) = \epsilon_m(t)$ *are known. Equivalently, we have the subset Yule-Walker equations,*

$$R_{(i_1,i_2,\cdots,i_m)} a_{(i_1,i_2,\cdots,i_m)} = r_{(i_1,i_2,\cdots,i_m)} \tag{2.35}$$

where

$$R_{(i_1,i_2,\cdots,i_m)} = (R_{(i_1,i_2,\cdots,i_m)}(i,j))_{k\times k}$$

$$R_{(i_1,i_2,\cdots,i_m)}(i,j) = \begin{cases} r(i-j) & i,j \in (i_1,i_2,\cdots,i_m) \\ 1 & i=j, \text{ and } i \notin (i_1,i_2,\cdots,i_m) \\ 0 & \text{otherwise} \end{cases}$$

$$r_{(i_1,i_2,\cdots,i_m)} = (r_{(i_1,i_2,\cdots,i_m)}(i))_{k\times 1}$$

$$r_{(i_1,i_2,\cdots,i_m)}(i) = \begin{cases} -r(i) & i \in (i_1,i_2,\cdots,i_m) \\ 0 & i \notin (i_1,i_2,\cdots,i_m) \end{cases}$$

and where $r(i) = r(-i)$ *is the lag k autocorrelation and* $r(i,j) = r(i-j)$.

If the matrix $R^{-1}_{(i_1,i_2,\cdots,i_m)}$ *is known, the increase in residual variance* $\sigma^2_{(i_1,i_2,\cdots,i_m)}(i_j)$ *due to the removal of any lag* $i_j \in (i_1,i_2,\cdots,i_m)$ *satisfies*

$$\sigma^2_{(i_1,i_2,\cdots,i_m)}(i_j) = C(0) a^2_{(i_1,i_2,\cdots,i_m)}(i_j)\left(1 - \frac{S(i_j)}{1+S(i_j)}\right) \tag{2.36}$$

where

$$S(i_j) = [\tilde{r}^{(i_j)}_{(i_1,\cdots,i_m)}]' R^{-1}_{(i_1,\cdots,i_m)} \tilde{r}^{(i_j)}_{(i_1,\cdots,i_m)}$$

$$\tilde{r}^{(i_j)}_{(i_1,\cdots,i_m)} = \begin{cases} r(j-i_j) & j \in (i_1,\cdots,i_{m-1}) \text{ and } j \neq i_j \\ 0 & \text{otherwise} \end{cases}$$

□

It can be seen clearly that $S(i_m)$ tends to be large when the absolute value of $r(i_m - j)$, $j \in (i_1,\cdots,i_{m-1})$ is large, and $\frac{S(i_m)}{1+S(i_m)}$ is a monotonic increasing function of $S(i_m)$, where

$$\lim_{S(i_m)\to+\infty} \frac{S(i_m)}{1+S(i_m)} = 1$$

and $S(i_m) \to +\infty$ if and only if $R_{(i_1,\cdots,i_m)}$ is singular.

In general, the term $(1-\frac{S(i_j)}{1+S(i_j)})$ in the increase in residual variance $\sigma^2_{(i_1,i_2,\cdots,i_m)}(i_j)$, due to removing lag i_j in equation (2.36), represents the error in the linear representation of lag i_j by lags $(i_1,\cdots,i_{j-1},i_{j+1},\cdots,i_m)$.

By lemma 2.2 and equation (2.30), we have

$$R^{-1}_{(i_1,\cdots,i_m)}(i_m,i_m) = (1-[r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]' R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})})^{-1} \tag{2.37}$$

and from equation (2.26), we have following proposition

Proposition 2.1 *If the matrix $R^{-1}_{(i_1,\cdots,i_m)}$ is known, the increase in residual variance $\sigma^2_{(i_1,\cdots,i_m)}(i_j)$ due to the removal of lag $i_j \in (i_1,i_2,\cdots,i_m)$ can be expressed as,*

$$\sigma^2_{(i_1,i_2,\cdots,i_m)}(i_j) = C(0)\frac{a^2_{(i_1,i_2,\cdots,i_m)}(i_j)}{R^{-1}_{(i_1,\cdots,i_m)}(i_j,i_j)} \tag{2.38}$$

□

However, how the coefficient $a_{(i_1,i_2,\cdots,i_m)}(i_j)$ of the removed lag i_j affects the variance increase is not clear. In the next section, without loss of generality and for the sack of simplicity, we suppose the removed general lag i_j is i_m.

2.2.2 Residual Variance and Projection Modulus

In this section, we examine first how the coefficient of a removed lag affects the increase in residual variance, and then derive the concept of the projection modulus due to removing a lag.

From the matrix block representation of $R_{(i_1,i_2,\cdots,i_{n_1})}$ and $R^{-1}_{(i_1,i_2,\cdots,i_m)}$ in equation (2.27) and (2.28), the coefficient, $a_{(i_1,i_2,\cdots,i_m)}(i_m)$, of lag i_m is given as,

$$a_{(i_1,i_2,\cdots,i_m)}(i_m) = B_{12}r_{(i_1,i_2,\cdots,i_{m-1})} + B_{22}r(i_m) \tag{2.39}$$

and also from lemma 2.2, we have

$$\begin{aligned} B_{12} &= -\{1 - [r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]' R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}\}^{-1} \\ &\quad \times [r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]' R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} \end{aligned} \tag{2.40}$$

By substituting the above equation and equation (2.30) into equation (2.39), we have

$$a_{(i_1,i_2,\cdots,i_m)}(i_m) = \frac{r(i_m) - [r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]' R^{-1}_{(i_1,i_2,\cdots,i_{m-1})}\, r_{(i_1,i_2,\cdots,i_{m-1})}}{1 - [r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]' R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}} \tag{2.41}$$

The representation of the subset AR model with lags $(i_1, i_2, \cdots, i_{m-1})$ in Hilbert space is given by

$$\Theta(0) = \sum_{j=1}^{m-1} a_{(i_1,i_2,\cdots,i_{m-1})}(i_j)\Theta(i_j) + e^*_{m-1} \tag{2.42}$$

where $e^*_{m-1} \perp \Theta(i_j),\ j = 1, \cdots, m-1$.

Taking inner products with $\Theta(i_m)$ on both sides of the above equation, we have

$$\langle \Theta(i_m), e^*_{m-1} \rangle = r(i_m) - [r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]' R^{-1}_{(i_1,i_2,\cdots,i_{m-1})}\, r_{(i_1,i_2,\cdots,i_{m-1})} \tag{2.43}$$

It can be seen that the right hand side of the above equation (2.43) is the numerator of equation (2.41).

Similarly, projecting $\Theta(i_m)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$ (see equation (2.18) and (2.19)), we have equation (2.23), i.e.

$$\Theta(i_m) = \sum_{j=1}^{m-1} b(i_j)\Theta(i_j) + e_{i_m} \tag{2.44}$$

Taking inner products, using $\Theta(i_m)$, on both sides of the above equation and from equation (2.25), we have

$$\langle \Theta(i_m), e_{m-1} \rangle = 1 - [r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]' R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} r^{(i_n)}_{(i_1,i_2,\cdots,i_{m-1})} \tag{2.45}$$

The right hand side of the above equation (2.45) is the denominator of the equation (2.41), and by noting that $e^*_{m-1} \perp (\Theta(i_1), \cdots, \Theta(i_{m-1}))$ and $e_{m-1} \perp (\Theta(i_1), \cdots, \Theta(i_{m-1}))$ therefore, the coefficient of lag i_m in the subset AR model with lags $(i_1, \cdots, i_m)$ can be expressed as

$$a_{(i_1,i_2,\cdots,i_m)}(i_m) = \frac{\langle \Theta(i_m), e^*_{m-1} \rangle}{\langle \Theta(i_m), e_{m-1} \rangle} = \frac{\langle e_{m-1}, e^*_{m-1} \rangle}{\langle e_{m-1}, e_{m-1} \rangle} \tag{2.46}$$

where e^*_{m-1} is the error of linear regression of $\Theta(0)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$; and e_{m-1} is the error of linear regression $\Theta(i_m)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$. The interpretation of the above equation is quite obvious and is that coefficient $a_{(i_1,i_2,\cdots,i_m)}(i_m)$ is directly proportional to the projection of $\Theta(i_m)$ on the error of the linear regression $\Theta(0)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$ because the projection, $\langle \Theta(i_m), e^*_{m-1} \rangle$, measures how much of the error e^*_{m-1} can be represented by $\Theta(i_m)$; the coefficient $a_{(i_1,i_2,\cdots,i_m)}(i_m)$ is inversely proportional to the projection of $\Theta(i_m)$ on the error of the linear regression of $\Theta(i_m)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$, because the projection, $\langle \Theta(i_m), e_{m-1} \rangle$, measures how much of $\Theta(i_m)$ cannot be linearly represented by $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$. It seems odd that the better is the linear regression of $\Theta(i_m)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$, the larger is the magnitude of the coefficient, $a_{(i_1,i_2,\cdots,i_m)}(i_m)$. However, from examining the structure of the error e^*_{m-1} in equation (2.21) and (2.42), we know that

$$e^*_{m-1} = a_{(i_1,i_2,\cdots,i_m)}(i_m) e_{m-1} + c\epsilon_m \tag{2.47}$$

where c is a constant and $\epsilon_m \perp (\Theta(i_1), \cdots, \Theta(i_m))$, i.e. $\langle \epsilon_m, \Theta(i_j) \rangle = 0$, $j = 1, \cdots, m$.

Hence, we may write

$$\frac{\langle \Theta(i_m), e^*_{m-1} \rangle}{\langle \Theta(i_m), e_{m-1} \rangle} = \frac{a_{(i_1,i_2,\cdots,i_m)}(i_m) \langle \Theta(i_m), e_{m-1} \rangle + c \langle \Theta(i_m), \epsilon_m \rangle}{\langle \Theta(i_m), e_{m-1} \rangle} = a_{(i_1,i_2,\cdots,i_m)}(i_m)$$

This expression, i.e. $\langle \Theta(i_m), e_{m-1} \rangle$, which indicates how well the lag i_m is linearly represented by the lags $(i_1, \cdots, i_{m-1})$, is only a scaling factor for the value of $a_{(i_1,i_2,\cdots,i_m)}(i_m)$. It would be misleading as to the true meaning of the coefficient $a_{(i_1,i_2,\cdots,i_m)}(i_m)$ if we simply record that this coefficient is directly proportional to $\langle \Theta(i_m), e^*_{m-1} \rangle$ and inversely proportional to $\langle \Theta(i_m), e_{m-1} \rangle$ without recognizing the effect of the relation between e^*_{m-1} and e_{m-1} (see equation (2.47)).

However, the new coefficients for the regression of $\Theta(0)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$ can be constructed from the regression of $\Theta(i_m)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$ and the coefficients obtained in the regression of $\Theta(0)$ on the lags $(\Theta(i_1), \cdots, \Theta(i_m))$.

We recall that equation (2.7) in section 2.2 links the new coefficients resulting from the regression when lag i_m is deleted and the old coefficients which arise from the regression on the "full" set of lags $(\Theta(i_1), \cdots, \Theta(i_m))$. Now, equation (2.21) illustrates that the coefficients in the reduced regression (i.e. with lag i_m deleted) may be shown to arise from the original coefficients (from the "full" set) and those created from regression of $\Theta(i_m)$ on $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$.

The increase in residual variance produced by removing lag i_m is affected by how well the lags $(i_1, \cdots, i_{m-1})$ linearly explain the lag i_m and is given by multiplication of two components $A_{(i_1,\cdots,i_m)}(i_m)(= a^2_{(i_1,\cdots,i_m)}(i_m))$ and $E_{(i_1,\cdots,i_m)}(i_m)(= C(0)\langle\Theta(i_m), e_{m-1}\rangle)$. If $A_{(i_1,\cdots,i_m)}(i_m)$ is very small, this indicates that lag i_m produces a very small increase in residual variance when it is removed no matter how badly lag i_m is represented by lags $(i_1, \cdots, i_{m-1})$ because the value of $E_{(i_1,\cdots,i_m)}(i_m)$ is bounded by $C(0)$, a finite value. The value of $E_{(i_1,\cdots,i_m)}(i_m)$ represents how much lag i_m adds in explanation apart from lags $(\Theta(i_1), \cdots, \Theta(i_{m-1}))$. On the other hand, if $E_{(i_1,\cdots,i_m)}(i_m)$ is very small, this implies that lag i_m can almost be represented by lags $(i_1, \cdots, i_{m-1})$, and so, the contribution to the increase in residual variance by removing lag i_m will also be small if the coefficient $a_{(i_1,\cdots,i_m)}(i_m)$ has a bounded value.

From the above explanation of the coefficient of lag i_m, $a_{(i_1,\cdots,i_m)}(i_m)$, and the increase in residual variance due to the removal of lag i_m, it can be seen that $\langle\Theta(i_m), e^*_{m-1}\rangle$ measures how much information remains in lag i_m which cannot be described by the size $(m-1)$ subset AR model with lags $(i_1, \cdots, i_{m-1})$ because $\langle\Theta(i_m), e^*_{m-1}\rangle$ can be regarded as a projection of e^*_{m-1} on $\Theta(i_m)$. The absolute value of the projection represents how much of the information in e^*_{m-1} can be represented by $\Theta(i_m)$. A smaller absolute value of the projection indicates that there is less information in e^*_{m-1} associated with $\Theta(i_m)$. Therefore, to reduce the subset size by 1 for a specified subset AR model, the choice of which lag should be removed can be

determined by the absolute values of the projection for different lags. Now, we define a new statistical measurement, the *projection modulus* of lag i_m on lags $(i_1, \cdots, i_{m-1})$ as

$$\rho^2_{(i_1,\cdots,i_{m-1})}(i_m) = \langle \Theta(i_m), e^*_{m-1} \rangle^2 \tag{2.48}$$

From the expression for the coefficient of lag i_m in equation (2.41), the projection modulus of lag i_m on lags $(i_1, \cdots, i_{m-1})$ can also be expressed as

$$\rho^2_{(i_1,\cdots,i_{m-1})}(i_m) = a^2_{(i_1,\cdots,i_m)}(i_m) \langle \Theta(i_m), e_{m-1} \rangle^2 \tag{2.49}$$

Removing the lag with the smallest projection modulus produces an "optimal" subset AR model reduced in size by 1 for the specified subset AR model, where the "optimum" is based on the smallest projection modulus. On the other hand, the increase in residual variance due to the removal of a lag is a commonly used statistic for measuring the significance of that lag in a specified subset AR model. Removing the lag with the smallest increase in residual variance produces a different "optimal" reduced subset AR model where "optimal" is now based on the smallest increase in residual variance. Therefore, the projection modulus and the increase in residual variance are two different statistical criteria to measure the significance of a deleted lag in a specified subset AR model. A question which arises is what is the relation between the two criteria.

Bearing in mind that (see equation (2.47)) we have,

$$e^*_{m-1} = a_{(i_1,\cdots,i_m)}(i_m) e_{m-1} + c\epsilon_m$$

and $e_{m-1} \perp (\Theta(i_1), \cdots, \Theta(i_{m-1}))$, $\epsilon_m \perp (\Theta(i_1), \cdots, \Theta(i_m))$, we know that the residual variance of the size $(m-1)$ subset AR model with lags $(i_1, \cdots, i_{m-1})$ is given by

$$\langle e^*_{m-1}, e^*_{m-1} \rangle = \sigma^2_{(i_1,\cdots,i_m)}(i_m) \langle e_{m-1}, e_{m-1} \rangle + c^2 \langle \epsilon_m, \epsilon_m \rangle$$

where $c^2 \langle \epsilon_m, \epsilon_m \rangle$ is the residual variance of the size m subset AR model with lags $(i_1, \cdots, i_m)$; and where we also have the expression,

$$\sigma^2_{(i_1,\cdots,i_{m-1})}(i_m) = a^2_{(i_1,\cdots,i_m)}(i_m) \langle e_{m-1}, e_{m-1} \rangle \tag{2.50}$$

which is the increase in the residual variance due to removing lag i_m.

Comparing equations (2.49) and (2.50), we obtain the following relation between the two criteria

$$\rho^2_{(i_1,\cdots,i_{m-1})}(i_m) = \sigma^2_{(i_1,\cdots,i_m)}(i_m)\langle e_{m-1}, e_{m-1}\rangle \tag{2.51}$$

The above relation indicates that the projection modulus depends more on how well lag i_m is linearly represented by lags $(i_1,\cdots,i_{m-1})$ than it does on the increase in residual variance in judging the optimality of the size $(m-1)$ AR model with lags $(i_1,\cdots,i_{m-1})$.

To illustrate this fact, we can express the projection modulus and the increase in residual variance as follows:

Define

$$\langle \vec{\alpha}, \vec{\beta}\rangle = \vec{\alpha}\cdot\vec{\beta} = \cos(\widehat{\vec{\alpha},\vec{\beta}})\|\vec{\alpha}\|\ \|\vec{\beta}\| \tag{2.52}$$

where $(\widehat{\vec{\alpha},\vec{\beta}})$ represents the angle between $\vec{\alpha}$ and $\vec{\beta}$; then we have

$$\begin{aligned}
\rho^2_{(i_1,\cdots,i_{m-1})} &= \langle e_{m-1}, e^*_{m-1}\rangle^2 \\
&= \cos^2(\widehat{e_{m-1}, e^*_{m-1}})\langle e^*_{m-1}, e^*_{m-1}\rangle\langle e_{m-1}, e_{m-1}\rangle \\
&= \cos^2(\widehat{\Theta(i_m), e^*_{m-1}})\langle e^*_{m-1}, e^*_{m-1}\rangle\langle e_{m-1}, e_{m-1}\rangle
\end{aligned} \tag{2.53}$$

and we also have

$$\begin{aligned}
\sigma^2_{(i_1,\cdots,i_{m-1})} &= \frac{\langle e_{m-1}, e^*_{m-1}\rangle^2}{\langle e_{m-1}, e_{m-1}\rangle} \\
&= \cos^2(\widehat{e_{m-1}, e^*_{m-1}})\langle e^*_{m-1}, e^*_{m-1}\rangle \\
&= \cos^2(\widehat{\Theta(i_m), e^*_{m-1}})\langle e^*_{m-1}, e^*_{m-1}\rangle
\end{aligned} \tag{2.54}$$

The term $\cos^2(\widehat{\Theta(i_m), e^*_{m-1}})$ represents the "angle" between $\Theta(i_m)$ and e^*_{m-1}, and $\cos^2(\widehat{\Theta(i_m), e^*_{m-1}}) = 0$ when $e^*_{m-1} \perp \Theta(i_m)$. In other words, when $\Theta(i_m)$ has no extra information on the error from the regression of $\Theta(0)$ on $(\Theta(i_1),\cdots,\Theta(i_{m-1}))$, then, there is no increase in residual variance due to the removal of lag i_m ; $\cos^2(\widehat{\Theta(i_m), e^*_{m-1}}) = 1$ implies that the error term e^*_{m-1} can be totally linearly represented by $\Theta(i_m)$. The terms $\langle e^*_{m-1}, e^*_{m-1}\rangle$ and $\langle e_{m-1}, e_{m-1}\rangle$ are measures of the errors of $\Theta(0)$, and $\Theta(i_m)$

Measures		Definition	exponent on $\langle e_{m-1}, e_{m-1}\rangle$
$a^2_{(i_1,\cdots,i_m)}(i_m)$	=	$\langle \Theta(i_m), e^*_{m-1}\rangle^2 \frac{1}{\langle e_{m-1}, e_{m-1}\rangle^2}$	-2
$\sigma^2_{(i_1,\cdots,i_{m-1})}$	=	$\langle \Theta(i_m), e^*_{m-1}\rangle^2 \frac{1}{\langle e_{m-1}, e_{m-1}\rangle}$	-1
$\rho^2_{(i_1,\cdots,i_{m-1})}$	=	$\langle \Theta(i_m), e^*_{m-1}\rangle^2$	0
$\gamma^2_{(i_1,\cdots,i_{m-1})}$	=	$\langle \Theta(i_m), e^*_{m-1}\rangle^2 \langle e_{m-1}, e_{m-1}\rangle$	1
$\eta^2_{(i_1,\cdots,i_{m-1})}$	=	$\langle \Theta(i_m), e^*_{m-1}\rangle^2 \langle e_{m-1}, e_{m-1}\rangle^2$	2

Table 2.1: Different Measurements of Errors

regressed on $(\Theta(i_1),\cdots,\Theta(i_{m-1}))$, respectively. The increase in residual variance only depends on the "angle" between $\Theta(i_m)$ and e^*_{m-1} and the error from $\Theta(0)$ regressed on $(\Theta(i_1),\cdots,\Theta(i_{m-1}))$; it does not take into account the error from $\Theta(i_m)$ regressed on $(\Theta(i_1),\cdots,\Theta(i_{m-1}))$. Therefore, it is not necessarily true that lag i_m can be best represented by other lags even when the increase in residual variance due to the removal of lag i_m is at its smallest value. The projection modulus does directly take account of the effect of $\Theta(i_m)$ regressed on $(\Theta(i_1),\cdots,\Theta(i_{m-1}))$ by multiplying the increase in residual variance by $\langle e_{m-1}, e_{m-1}\rangle$. Similarly, achieving the smallest projection modulus does not mean that the corresponding increase in residual variance is at its minimum. However, the only difference between them is that projection modulus depends not only on the increase in residual variance but also on how well the removed lag is linearly related to the other included lags.

When selecting the best subset AR model of size one less than the subset model already selected, this property of the projection modulus is very useful since the best subset AR model should be the model which does not include any lag which can be well represented by the other included lags.

Discussion

It is interesting to note that if we set out the possible measures in Table 2.1. The emphasis given to, $\langle e_{m-1}, e_{m-1}\rangle$, the length of the vector, e_{m-1}, which is the failure of the removed lag i_m to be linearly represented by lags $(i_1,\cdots,i_{m-1})$, $\langle e_{m-1}, e_{m-1}\rangle$ is shown by the power transformation applied to this length. $\gamma^2_{(i_1,\cdots,i_{m-1})}$ and $\eta^2_{(i_1,\cdots,i_{m-1})}$

are derived from the power transformation and may not have some obvious statistical meaning attached to them.

Mann and Wald (1943) proved that $\{\sqrt{N}(\hat{a}_{(i_1,\cdots,i_m)} - a_{(i_1,\cdots,i_m)})\}$ has an asymptotic multivariate normal distribution with zero mean and variance-covariance matrix $c_m^2\langle\epsilon_m,\epsilon_m\rangle R^{-1}_{(i_1,\cdots,i_m)}$ where $\hat{a}_{(i_1,\cdots,i_m)}$ is the least squares estimate of $a_{(i_1,\cdots,i_m)}$. Therefore the covariance of the estimated $a_{(i_1,\cdots,i_m)}$ is asymptotically $c_m^2\langle\epsilon_m,\epsilon_m\rangle R^{-1}_{(i_1,\cdots,i_m)}(i_m,i_m)/N$.

By lemma 2.2 and equation (2.30), we have

$$\begin{aligned} R^{-1}_{(i_1,\cdots,i_m)}(i_m,i_m) &= (1-[r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})}]'\, R^{-1}_{(i_1,i_2,\cdots,i_{m-1})} r^{(i_m)}_{(i_1,i_2,\cdots,i_{m-1})})^{-1} \\ &= \langle e_{m-1},e_{m-1}\rangle^{-1} \end{aligned} \tag{2.55}$$

The $t-$squared statistics of the coefficient $a_{(i_1,\cdots,i_m)}(i_m)$ will become

$$\begin{aligned} t^2(i_m) &= \frac{a^2_{(i_1,\cdots,i_m)}(i_m)}{c_m^2\langle\epsilon_m,\epsilon_m\rangle R^{-1}_{(i_1,\cdots,i_m)}(i_m,i_m)/N} \\ &= \frac{N}{c_m^2\langle\epsilon_m,\epsilon_m\rangle} a^2_{(i_1,\cdots,i_m)}(i_m)\langle e_{m-1},e_{m-1}\rangle \\ &= \frac{N}{c_m^2\langle\epsilon_m,\epsilon_m\rangle}\sigma^2_{(i_1,\cdots,i_{m-1})}(i_m) \end{aligned} \tag{2.56}$$

In general, the above equation shows that, in a subset AR model with lags $(i_1,\cdots,i_m)$, the $t-$squared statistics of the coefficient for lag i_j is equivalent to the increase in residual variance due to the absence of the lag $i_j \in (i_1,\cdots,i_m)$.

Therefore, besides the increase in residual variance or the $t-$squared statistics for coefficient, $a^2_{(i_1,\cdots,i_m)}(i_m)$, the projection modulus, $\rho^2_{(i_1,\cdots,i_{m-1})}$, provides a new statistic to measure the significance of lag i_m. This new statistic depends more on the linear representation of lag i_m by the included lags in a specified subset AR model. Comparing the relation between the increase in residual variance and the projection modulus, we can derive further statistics, such as $\gamma^2_{(i_1,\cdots,i_{m-1})}$, $\eta^2_{(i_1,\cdots,i_{m-1})}$, with a different degree of importance given to the linear representation of lag i_m by the included lags, to measure the significance of lag i_m.

2.3 A New Subset AR Model Selection Algorithm

Hocking and Leslie (1967) have described an algorithm which always avoids checking every $m-$subset of a $k-$variate regression model before finding the subset with minimum residual variance. McClave (1975) adapted the algorithm and simplified it for the subset AR situation. The McClave algorithm first solves the full lag Yule-Walker equation (2.2) and then orders each lag in decreasing order of influence on the residual variance i.e θ_i, $(i = 1, 2, \cdots k)$ measures the increase in residual variance due to the absence of lag i in the full lag AR model

$$\theta_{i_1} \geq \theta_{i_2} \geq \cdots \geq \theta_{i_k}$$

where $\theta_i = \hat{\sigma}_k^2 a^2(i)/\hat{\sigma}_{a(i)}^2$. $\hat{\sigma}_k^2$ is the residual variance of the full lag AR model with maximum lag k; $a(i)$ is the AR coefficient of lag i in the full lag model and $\hat{\sigma}_{a(i)}^2$ is the estimated variance of the estimate of $a(i)$.

Suppose we wish to determine that $m-$ lag model with minimum residual variance whose maximum lag does not exceed k. Let $q = k - m$ and consider any subset of lags $(j_1, j_2, \cdots, j_q)$ which are ordered according to the ranking established for θ_i, $i = 1, 2, \cdots, k$. McClave established the basic idea, which is adapted from Hocking and Leslie, that *if the increase in residual variance arising from the removal of the q-subset of lags* $(j_1, j_2, \cdots, j_q)$ *is not larger than* $\theta_{i_{(j_1-1)}}$, *then no q-subset with lag exceeding* j_1 *removed can produce a smaller increase in residual variance.*

By using this idea, McClave developed his algorithm to search for the best subset AR model for a specified size m. The search scheme is described as follows:

Step 1 Remove the q lags corresponding to the q smallest θ_i values in the order system $(i_1, i_2, \cdots, i_k)$. i.e. lags $(i_{m+1}, \cdots, i_k)$ are removed.

- If the increase in residual variance due to the removal does not exceed θ_{i_m}, the best $m-$subset AR consists of lags $(i_1, \cdots, i_m)$ in $(1, 2, \cdots, k)$. The algorithm is terminated.

- If the increase exceeds θ_{i_m}, the algorithm chooses lag subsets with size q in the lag set $(i_m, i_{m+1}, \cdots, i_k)$ to be removed to form q candidates of subset AR models with size m. The model with the minimum increase in residual variance is selected among the q candidate models and then we go to the next step.

Step 2

- If the minimum increase does not exceed $\theta_{i_{m-1}}$, the corresponding complementary lag set is the $m-$subset having minimum residual variance among all $\begin{pmatrix} k \\ m \end{pmatrix}$ subsets. The optimum $m-$subset AR model is found. The algorithm is terminated.
- If the minimum increase exceeds $\theta_{i_{m-1}}$, the $q-$subset of the lag set $(i_{m-1}, \cdots, i_k)$ to be removed to form the $\begin{pmatrix} q+2 \\ q \end{pmatrix}$ candidate subset AR models with size m. The model with the minimum increase in residual variance is selected from among the $\begin{pmatrix} q+2 \\ q \end{pmatrix}$ candidate models and then we go to the next step.

$\vdots$

Step i

- If the minimum increase does not exceed $\theta_{i_{m-i+1}}$, the corresponding complementary lag set is the $m-$subset having minimum residual variance among all $\begin{pmatrix} k \\ m \end{pmatrix}$ subsets. The optimum $m-$subset AR model is found. The algorithm is terminated.
- If the minimum increase exceeds $\theta_{i_{m-i+1}}$, the $q-$subset of the lag set $(i_{m-i+1}, \cdots, i_k)$ to be removed to form the $\begin{pmatrix} q+i \\ q \end{pmatrix}$ candidate subset AR models with size m. The model with the minimum increase in residual variance is selected among the $\begin{pmatrix} q+i \\ q \end{pmatrix}$ candidate models and then we go to the next step.

⋮

Step m

- If the minimum increase does not exceed θ_{i_1}, the corresponding complementary lag set is the $m-$subset having minimum residual variance among all $\begin{pmatrix} k \\ m \end{pmatrix}$ subsets. The optimum $m-$subset AR model is found. The algorithm is terminated.
- If the minimum increase exceeds θ_{i_1}, the $q-$subset of the lag set $(i_1, \cdots, i_k)$ to be removed to form the $\begin{pmatrix} k \\ q \end{pmatrix}$ candidates of subset AR models with size m. The model with the minimum increase in residual variance is selected among the $\begin{pmatrix} k \\ q \end{pmatrix}$ and is the optimum $m-$subset AR model. The algorithm is terminated.

The above scheme provides a search mechanism for a specified size m subset AR model. In normal circumstances, McClave's algorithm is terminated before **Step m** . Therefore, the algorithm avoids checking all possible subset AR models. However, our task is to find an optimum subset AR for a data set without a specified size. The McClave's search scheme can be employed by sequentially reducing (or increasing) the size of the subset AR model. For each given size of subset AR model, we search for the minimum variance subset AR model using the above algorithm and using model selection criteria for the appropriate size in selecting the "optimum" subset AR model, i.e. we use,

Criterion 1 Based on Akaike (1970) information criterion:

$$AIC = N \log(\sigma^2_{(i_1,\cdots,i_m)}) + 2m \tag{2.57}$$

Criterion 2 Based on the method of Hannan and Quinn (1979):

$$HC = \log(\sigma^2_{(i_1,\cdots,i_m)}) + \frac{2m \log(\log(N))}{N} \tag{2.58}$$

Criterion 3 Based on the Schwarz (1978) criterion:

$$SC = N\log(\sigma^2_{(i_1,\cdots,i_m)}) + m\log(N) \tag{2.59}$$

Criterion 4 Based on the Hocking and Leslie (1967) criterion:

$$C = N\sigma^2_{(i_1,\cdots,i_m)}/\sigma^2 + 2m - N \tag{2.60}$$

where m is the size of the subset AR , $\sigma^2_{(i_1,\cdots,i_m)}$ denotes the estimated residual variance after fitting the size m subset AR model and σ^2 is the least squares estimate of the residual variance of the full lag model.

The optimum subset AR model can be determined by comparing with these criteria among the minimum variance subset AR models for each different size.

It can be seen that we have to solve at least $\begin{pmatrix} q+i \\ q \end{pmatrix}$ or more subset Yule-Walker equations in order to compare the increase in residual variances of those candidate models once the increases in residual variance of the subset AR with size m, whose complementary lags are collected in the $q-$subset of the lag set $(i_{m-i+1},\cdots,i_k)$, exceeds $\theta_{i_{m-i+1}}$, $i = 1,2,\cdots,m$. Although McClave used the Cholesky decomposition approach (see Pagano (1972)) to avoid solving the subset Yule-Walker equations for the increase in residual variance, the search algorithm consumes a lot of computer time when it deals with a large maximum lag k. Consequently, the computing speed of McClave's algorithm cannot cope with situations in practice where the maximum lag k is large.

Now, we examine McClave's algorithm, sequentially reducing the size of subset AR models from the full lag AR model. Without loss of generality, we suppose the size $(m-1)$ $(m \leq k)$ subset AR model with minimum variance is the subset AR model with the lag set $(i_1,\cdots,i_{m-1})$. It is noted from equation (2.21) that the other AR coefficients are modified due to removing lag i_m. The significance of the lags are represented by the lag number ordering, i.e. $(i_1,\cdots,i_m)$ on the basis of $\theta_{i_1} \geq \theta_{i_2} \geq \cdots \geq \theta_{i_m}$. When a lag is deleted, for example i_m, the remaining $(m-1)$

lags are not necessarily ordered $(i_1, \cdots, i_{m-1})$ and, of course, if a further deletion of lag i_{m-1} was carried out, there is no reason that the lags $(i_1, \cdots, i_{m-2})$ are still correctly ordered and therefore giving the best $(m-2)$-subset AR model. i.e. the increase in the residual variance is likely to exceed $\theta_{i_{m-3}}$. The search procedure has to continue checking $(m-2)$-subsets in the lag set $(i_1, \cdots, i_{m-1})$ according to McClave's algorithm. If all increases in residual variance of all the $(m-1)$ candidate AR subset models of size $(m-2)$ exceed $\theta_{i_{m-2}}$, the search procedure starts to search for the $(m-2)$-subset AR model in the increased lag set $(i_1, \cdots, i_m)$. The minimum variance $(m-2)$-subset AR is found if the increase in residual variance does not exceed $\theta_{i_{m-3}}$, and so on.

From the above analysis, it can be seen that the ordering based on $(i_1, \cdots, i_m)$ may be invalid for the reduced size model and so causes the search procedure to search more candidate subset AR models. If we can re-rank the order system after finding the minimum variance subset AR model with size m, the minimum variance subset AR model with size $(m-1)$ may be more easily found according to the new order system.

Suppose having the minimum variance subset AR model with size m and lag set $(i_1, \cdots, i_m)$, we calculate the τ^2 values of newly estimated AR coefficients $\hat{a}(i_j)$, $j = 1, 2, \cdots, m-1$.

$$\tau_{i_j}^2 = \hat{a}_{i_j}^2 / \sigma_{\hat{a}_{i_j}}^2 \tag{2.61}$$

and sort them into decreasing order so that we have,

$$\tau_{l_1}^2 \geq \tau_{l_2}^2 \geq \cdots \geq \tau_{l_{m-1}}^2 \tag{2.62}$$

where $l_j \in (i_1, \cdots, i_{m-1}), j = 1, 2, \cdots, m-1$.

We re-rank the first $(m-1)$ elements of the order system obtained after finding the minimum variance subset AR model with size m according to $(l_1, l_2, \cdots, l_{m-1})$ to form a new order system. The lag corresponding to the $(m-1)$th element of the new system has the minimum τ^2 among other lags. We should bear in mind that the increase of residual variance due to removing lag i_j in a subset AR model consists of two parts

$A_{(i_1,\cdots,i_m)}(i_j)$ and $E_{(i_1,\cdots,i_m)}(i_j)$ (see equation (2.26)). The first part is the square of the AR coefficient of the removed lag. The second part is the error due to the part of the characteristic behaviour of the removed lag not being represented by the remaining included lags. Therefore, the square of each AR coefficient is only the "amplitude" of the removed lag. Since the two parts have no "direct linkage" between them, in the sense that if $A_{(i_1,\cdots,i_m)}(i_j)$ is large (or small) there is no reason why $E_{(i_1,\cdots,i_m)}(i_j)$ should be small (or large), and since both are positive in value, the large value of the squared AR coefficient will not automatically be compensated for by the second term which indicates the failure of the removed lag to be represented by other included lags and vice versa. The increase in residual variance has to be used to search for the best subset model for a fixed size under the new order system which serves as a quicker path to find the best subset model. Under the new order system, the lag l_{m-1} and the set $(i_m,\cdots,i_k)$ are removed to calculate $\sigma^2(l_{m-1})$ first. Sequentially proceeding with the calculations of $\sigma^2(l_j)$ due to removing in turn the lag l_j for $j = m-2$ to $j = 1$ together with lags $(i_m,\cdots,i_k)$. The "best" $(m-2)$-subset AR model is found once we have established that $\sigma^2(l_j) \leq \sigma^2(l_{m-1})$, since it implies that lag l_j can be represented better by other included lags than lag l_{m-1} although the l_{m-1}th lag's amplitude is smaller than is the l_jth lag's, i.e. $\tau^2(l_{m-1}) \leq \tau^2(l_j),\ j = 1,\cdots,m-2$.

It is noticed that the "best" $m-2$-subset AR model produced by the above procedure is the best in the lag set $(i_1,\cdots,i_m)$, and we cannot be sure it is the best subset AR model among all $\begin{pmatrix} k \\ m-2 \end{pmatrix}$ candidates drawn from the lags $(1,2,\cdots,k)$. In other words, the "best" of all subset models is the local best instead of the global best, in the sense of minimal increase in residual variance, since the candidate lag set, $(i_1,\cdots,i_m)$, is only a subset of the full lag set. Therefore, it will not always be the case that the local minimum residual variance model is the same as the global one. However, we have seen that the increase in residual variance due to the removing of lag i_j from an AR model with lag set $(i_1,i_2,\cdots,i_m)$ can be represented as follows, $\sigma^2_{(i_1,i_2,\cdots,i_m)}(i_j) = A_{(i_1,i_2,\cdots,i_m)}(i_j)\ E_{(i_1,i_2,\cdots,i_m)}(i_j)$. There are two contributing parts $A_{(i_1,i_2,\cdots,i_m)}(i_j)$ and $E_{(i_1,i_2,\cdots,i_m)}(i_j)$ to $\sigma^2_{(i_1,i_2,\cdots,i_m)}(i_j)$. We construct another

statistic, $\delta^{(\beta)}_{(i_1,i_2,\cdots,i_m)}(i_j)$, as proposed below

$$\delta^{(\beta)}_{(i_1,i_2,\cdots,i_m)}(i_j) = 2\beta \log A_{(i_1,i_2,\cdots,i_m)}(i_j) + 2(1-\beta)\log E_{(i_1,i_2,\cdots,i_m)}(i_j) \tag{2.63}$$

where β ranges from 0 to 1, and is a trade-off parameter.

Different choice of β will result in different emphases being given to removing a lag. For instance, suppose that when lag i_j is removed

- If β is chosen as 1, $\delta^{(1)}_{(i_1,\cdots,i_m)}(i_j)$ is solely dependent on the squared coefficient, $a^2_{(i_1,\cdots,i_m)}(i_j)$, for lag i_j (see Table 2.1) since $\delta^{(1)}_{(i_1,\cdots,i_m)}(i_j) = 2\log a^2_{(i_1,\cdots,i_m)}(i_j)$;
- If β is chosen as 1/2, $\delta^{(1/2)}_{(i_1,\cdots,i_m)}(i_j)$ is solely dependent on the increase in residual variance, $\sigma^2_{(i_1,\cdots,i_m)}(i_j)$ (see Table 2.1) since $\delta^{(1/2)}_{(i_1,\cdots,i_m)}(i_j) = \log \sigma^2_{(i_1,\cdots,i_m)}(i_j)$;
- if β is chosen as 1/3, $\delta^{(1/3)}_{(i_1,\cdots,i_m)}(i_j)$ is solely dependent on the projection modulus, $\rho^2_{(i_1,\cdots,i_m)}(i_j)$ (see Table 2.1) since $\delta^{(1/3)}_{(i_1,\cdots,i_m)}(i_j) = \frac{2}{3}\log \rho^2_{(i_1,\cdots,i_m)}(i_j)$;
- if β is chosen as 1/4, $\delta^{(1/4)}_{(i_1,\cdots,i_m)}(i_j)$ is solely dependent on the $\gamma^2_{(i_1,\cdots,i_m)}(i_j)$ (see Table (2.1) since $\delta^{(1/4)}_{(i_1,\cdots,i_m)}(i_j) = \frac{1}{2}\log \gamma^2_{(i_1,\cdots,i_m)}(i_j)$;
- if β is chosen as 1/5, $\delta^{(1/5)}_{(i_1,\cdots,i_m)}(i_j)$ is solely dependent on the $\eta^2_{(i_1,\cdots,i_m)}(i_j)$ (see Table 2.1) since $\delta^{(1/5)}_{(i_1,\cdots,i_m)}(i_j) = \frac{2}{5}\log \eta^2_{(i_1,\cdots,i_m)}(i_j)$.

In general, $\delta^{(\beta)}_{(i_1,\cdots,i_m)}(i_j)$ gives more weight to $E_{(i_1,\cdots,i_m)}(i_j)$ and makes $\delta^{(\beta)}_{(i_1,\cdots,i_m)}(i_j)$ more sensitive to $E_{(i_1,\cdots,i_m)}(i_j)$ if β is chosen in the range (0, 0.5); and gives more weight to $A_{(i_1,\cdots,i_m)}(i_j)$ and make $\delta^{(\beta)}_{(i_1,\cdots,i_m)}(i_j)$ more sensitive to $A_{(i_1,\cdots,i_m)}(i_j)$ if β is chosen in the range (0.5, 1). So we see β is a *trade-off parameter* between $A_{(i_1,\cdots,i_m)}(i_j)$ and $E_{(i_1,\cdots,i_m)}(i_j)$. Our decision will be to choose an appropriate β to best retain the true lags in the local best models, though we remember that the local best subset models may not be the global best. In other words, the appropriate choice of β should set up an appropriate sensitivity level to $A_{(i_1,i_2,\cdots,i_m)}(i_j)$ and $E_{(i_1,i_2,\cdots,i_m)}(i_j)$ to ensure, whenever possible, that true lags are not removed and also that the true subset model is one of the best local models found by reducing the size of the subset AR models.

Adapting McClave's algorithm, we develop a new algorithm to search for the optimum subset AR model for any data set as follows:

Stage 1: solve the full lag Yule-Walker equation (2.2) and sort the increase in residual variance θ_i, $(i = 1, 2, \cdots k)$ due to the absence of each lag i in the full lag AR model in decreasing order

$$\theta_{i_1} \geq \theta_{i_2} \geq \cdots \geq \theta_{i_k} \tag{2.64}$$

where $\theta_i = \hat{\sigma}_k^2 a^2(i)/\hat{\sigma}_{a(i)}^2$, and where $\hat{\sigma}_k^2$ is the residual variance of the full lag AR model with maximum lag k. $a(i)$ is the AR coefficient of lag i in the full lag model. $\hat{\sigma}_{a(i)}^2$ is the estimated variance of $a(i)$.

Stage 2:

This stage begins with size m set to k and reduces one by one to size 1.

Step 1 Removing lag i_m corresponding to the smallest θ_{i_m} value in the order system $(i_1, \cdots, i_m)$.

Step 2 If the increase in residual variance, $\sigma^2_{(i_1,\cdots,i_m)}(i_m)$, due to the removal does not exceed the increase in residual variance with θ_{i_m-1}, the local best $(m-1)$-subset AR model with lags $(i_1, i_2, \cdots, i_{m-1})$ is found. Calculate the model selection criterion, and go to to **Step 4.**

Step 3 If the increase exceeds $\theta_{i_{m-1}}$, using theorem 2.1, calculate $\sigma^2_{(i_1,\cdots,i_m)}(i_j)$ and so $\delta^{(\beta)}_{(i_1,\cdots,i_m)}(i_j)$ for j from $(m-1)$ to 1. If $\delta^{(\beta)}_{(i_1,\cdots,i_m)}(i_j)$ is less than $\delta^{(\beta)}_{(i_1,\cdots,i_m)}(i_m)$, the local best $(m-1)$-subset AR model with lags $(i_1, \cdots, i_{j-1}, i_{j+1}, \cdots, i_m)$ is found. Calculate the model selection criterion.

Step 4 Calculate the AR coefficients of the local best model with size $(m-1)$, and re-order the θ_i system in (2.64) by the t^2 of the AR coefficients of the local best $(m-1)$-subset AR model from **Step 3.** Back to **Stage 2.**

Stage 3: Determine the smallest model selection criterion of all the local best subset models for every size, $m = 1, \cdots, k$. The subset model with the smallest criterion is the optimum subset model.

The re-ranked order system in **Step 4** of **Stage 3** gives a search path, which can avoid looking at every possible model, for the local best model with the size progressively reduced by 1. The optimum model can be found when at most $k(k+1)/2$ candidate models have been checked. It is obvious that more weight should be put on $E_{(i_1,i_2,\cdots,i_m)}(i_j)$ since it represents the error arising from projecting lag i_j on the rest of lags in the set $(i_1, i_2, \cdots, i_m)$. Computer simulation shows that β ranging from 0.1 to 0.5 is appropriate to ensure that the proposed algorithm obtains and retains the true lags in the local best subset models of different size. Consequently, the true optimum subset AR models can be found with high probability. However, we have no evidence to show that there is an optimum value of β which retains the true model lags with greater probability than any other values when the proposed algorithm is applied to those data sets simulated from different subset AR models. Nevertheless, it has been our experience so far that the true model has a greater chance of being selected as the optimum model by the proposed algorithm if β is chosen slightly less than 0.5. In the next section, two numerical examples are presented to show the efficiency of the proposed algorithm.

2.4 Numerical Examples

The subset selection algorithm described in the above section was applied to 100 independent simulations of the stationary model given by

$$X(t) + 0.8X(t-1) - 0.4X(t-3) + 0.2X(t-12) = \epsilon(t) \tag{2.65}$$

where $\epsilon(t)$ $i.i.d.(0,1)$.

Each sample contains 240 observations and the maximum lag $k = 20$ was employed to test the algorithm. Table 2.2 summarizes how the true model has been found for different choices of β where the true optimum means that the true model is chosen by the proposed algorithm as the optimum model. The choice of the optimum model is based on the model selection criterion SC specified in section 2.3.

β	0	0.03	0.05	0.07	0.1	0.13	0.15	1/5	1/4
True Optimum	7	74	85	84	86	82	81	75	74
β	0.3	1/3	0.4	1/2	0.6	0.7	0.8	0.9	1
True Optimum	69	68	65	57	56	58	60	61	61

Table 2.2: The Performance of the Optimum AR Subset Search Algorithm for Different β

Table 2.2 shows that the proposed algorithm using model selection criterion SC with β in the range of 0.05 to 0.4 performs better than β in the range of 0.5 to 1 for the simulated data; with the best choice of β around 0.1. From the 100 sample data sets, for $\beta = 0.5$, 57 optimum models are the true optimum model whereas with β set at to 0.1, 86 optimum model out of those 100 chosen optimum model are the true true optimum model. We also found in these 100 sample data sets that those sample data sets where the true model is chosen as the optimum model when $\beta = 0.5$ were always again chosen when $\beta = 0.1$. This observation suggests that the algorithm with $\beta = 0.1$ is superior to the algorithm with $\beta = 0.5$. In other words, it seems highly likely that the true model will be chosen as optimum model with a greater probability when $\beta = 0.1$. Among the model selection criteria, AIC and C seem to be the most conservative; and they choose almost the same optimum model, i.e. same size and same lags. HC seems to be less conservative in most cases than AIC and C; whereas, SC is, as expected, the most parsimonious, i.e. the least conservative. It seems that the optimum models chosen by SC have a greater chance of being the true model.

To illustrate the advantage of the proposed algorithm and selection criterion $\delta^{(\beta)}$, we choose one sample of the simulated data set for which with $\beta = 0.5$ (the proposed criterion, $\delta^{(0.5)}$ is therefore the logarithm of the increase in the residual variance) this method fails to select the lags of the true model. However, the choice of $\beta = 0.1$ selects the true model. The global best subset models for each size and corresponding model selection criteria are shown in Table 2.3 on page 48. The local best subset models chosen by $\delta^{(0.5)}$ and $\delta^{(0.1)}$ for each size and corresponding model selection criteria are presented in Table 2.4 and Table 2.5 on pages 49 and 50 respectively. The optimum

model chosen by different model selection criteria are indicated by underline.

Comparing the chosen best local models for $\beta = 0.5$ and $\beta = 0.1$ in Table 2.4 and 2.5 on page 49 and 50 respectively, the local best models for lags size 10 to 20 are the same. Lag 12 is removed when the proposed algorithm with $\beta = 0.5$ searches for the local best model of size 9, while the lag 12 is retained if the proposed algorithm uses $\beta = 0.1$. The lags of the local best model with size 9 chosen by $\beta = 0.5$ is (1,3,4,10,11,14,16,17,19) with residual variance $\sigma^2(9) = 1.0875$ while the lags of the local best model with size 9 chosen by $\beta = 0.1$ is (1,3,4,8,10,12,16,17,19) with residual variance $\sigma^2(9) = 1.0959$. Although the proposed algorithm using $\beta = 0.5$ achieves a smaller residual variance than using $\beta = 0.1$ does, the true lag 12 is removed. Therefore, the proposed algorithm with $\beta = 0.5$ removes the true lag 12 but produces a smaller increase in residual variance; and the proposed algorithm with $\beta = 0.1$ retains the true lag 12 at the cost of a greater increase in residual variance. Although it achieves the minimum increases in residual variance for size 6 to 9, the proposed algorithm with $\beta = 0.5$ misses the true model lags in the chosen models with smaller size since the removed true lag 12 cannot be in the candidate models using the proposed search scheme. The proposed algorithm with $\beta = 0.1$ retains the true lags in searching for the local best models with size 6 to 9 although the local best models do not achieve the minimum increase in residual variance. However, using $\beta = 0.1$ does ensure that the true model lags are included in the candidate models with smaller size. The proposed algorithm with $\beta = 0.1$ finds the true model lags as the local best model with size 3. Comparing the local best model in Table 2.4 and 2.5 with the global best model in Table 2.3, we find that the local best models, and those which are chosen by the search with $\beta = 0.5$, are global best models for size 7 to 20; the local best models, which are chosen by the proposed algorithm with $\beta = 0.1$, are global best models with size 1 to 5 and 10 to 20. Neither of the two local best models with size 6 are global best. The proposed algorithm with $\beta = 0.1$ will find the true model as its optimum model since the true lag size is 3.

Global Best Subset Models Based on Residual Variances

Size(m)	Lags Chosen by Increase in Residual Variances of All Subsets				
1	1				
2	1,3				
3	1,3,12				
4	1,3,12,19				
5	1,3,10,12,19				
6	1,3,4,10,12,19				
7	1,3,4,10,11,14,19				
8	1,3,4,10,11,14,16,19				
9	1,3,4,10,11,14,16,17,19				
10	1,3,4,10,11,12,14,16,17,19				
11	1,3,4,8,10,11,12,14,16,17,19				
12	1,3,4,8,10,11,12,13,14,16,17,19				
13	1,3,4,8,10,11,12,13,14,15,16,17,19				
14	1,2,3,4,8,10,11,12,13,14,15,16,17,19				
15	1,2,3,4,6,8,10,11,12,13,14,15,16,17,19				
16	1,2,3,4,6,8,10,11,12,13,14,15,16,17,19,20				
17	1,2,3,4,6,8,10,11,12,13,14,15,16,17,18,19,20				
18	1,2,3,4,5,6,8,10,11,12,13,14,15,16,17,18,19,20				
19	1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18,19,20				
20	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20				
Size(m)	HC(m)	AIC(m)	S C(m)	C(m)	$\sigma^2(m)$
1	0.71258	169.62	173.10	210.16	2.0105
2	0.21207	48.091	55.052	31.860	1.2017
3	<u>0.17736</u>	38.358	<u>48.800</u>	21.081	1.1443
4	0.18081	37.784	51.707	20.359	1.1321
5	0.18503	37.396	54.799	19.861	1.1209
6	0.18339	<u>35.600</u>	56.484	<u>17.933</u>	1.1033
7	0.19315	36.540	60.904	18.856	1.0985
8	0.20200	37.261	65.106	19.554	1.0926
9	0.21150	38.137	69.463	20.417	1.0875
10	0.22283	39.454	74.261	21.728	1.0844
11	0.23287	40.463	78.750	22.731	1.0800
12	0.24597	42.204	83.972	24.472	1.0788
13	0.25919	43.973	89.221	26.241	1.0778
14	0.27284	45.847	94.576	28.115	1.0772
15	0.28672	47.777	99.987	30.045	1.0769
16	0.30079	49.751	105.44	32.019	1.0768
17	0.31494	51.743	110.91	34.011	1.0767
18	0.32909	53.737	116.39	36.005	1.0767
19	0.34325	55.733	121.86	38.000	1.0767
20	0.35742	57.732	127.35	40.000	1.0767
215 Models are Checked Before The Optimum Model is Found					

Table 2.3: Globally Selected AR Subset Models for the Simulated Data

Local Best Subset Models Based on $\delta^{(0.5)}$

Size(m)	Lags Chosen by $\delta^{(0.5)}$				
1	1				
2	1,3				
3	1,3,11				
4	1,3,10,11				
5	1,3,10,11,19				
6	1,3,4,10,11,19				
7	1,3,4,10,11,14,19				
8	1,3,4,10,11,14,16,19				
9	1,3,4,10,11,14,16,17,19				
10	1,3,4,10,11,12,14,16,17,19				
11	1,3,4,8,10,11,12,14,16,17,19				
12	1,3,4,8,10,11,12,13,14,16,17,19				
13	1,3,4,8,10,11,12,13,14,15,16,17,19				
14	1,2,3,4,8,10,11,12,13,14,15,16,17,19				
15	1,2,3,4,6,8,10,11,12,13,14,15,16,17,19				
16	1,2,3,4,6,8,10,11,12,13,14,15,16,17,19,20				
17	1,2,3,4,6,8,10,11,12,13,14,15,16,17,18,19,20				
18	1,2,3,4,5,6,8,10,11,12,13,14,15,16,17,18,19,20				
19	1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18,19,20				
20	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20				
Size(m)	HC(m)	AIC(m)	S C(m)	C(m)	$\sigma^2(m)$
1	0.71258	169.62	173.10	210.16	2.0105
2	0.21207	48.091	55.052	31.860	1.2017
3	0.19481	42.548	52.990	25.573	1.1645
4	0.18461	38.696	<u>52.618</u>	21.320	1.1364
5	0.18663	37.780	55.183	20.261	1.1227
6	<u>0.18576</u>	<u>36.168</u>	57.052	<u>18.523</u>	1.1059
7	0.19315	36.540	60.904	18.856	1.0985
8	0.20200	37.261	65.106	19.554	1.0926
9	0.21150	38.137	69.463	20.417	1.0875
10	0.22283	39.454	74.261	21.728	1.0844
11	0.23287	40.463	78.750	22.731	1.0800
12	0.24597	42.204	83.972	24.472	1.0788
13	0.25919	43.973	89.221	26.241	1.0778
14	0.27284	45.847	94.576	28.115	1.0772
15	0.28672	47.777	99.987	30.045	1.0769
16	0.30079	49.751	105.44	32.019	1.0768
17	0.31494	51.743	110.91	34.011	1.0767
18	0.32909	53.737	116.39	36.005	1.0767
19	0.34325	55.733	121.86	38.000	1.0767
20	0.35742	57.732	127.35	40.000	1.0767
110 Models are Checked Before The Optimum Model is Found					

Table 2.4: Locally Selected AR Subset Models by $\delta^{(0.5)}$ for the Simulated Data

Local Best Subset Models Based on $\delta^{(0.1)}$

Size(m)	Lags Chosen by $\delta^{(0.1)}$
1	1
2	1,3
3	1,3,12
4	1,3,12,19
5	1,3,10,12,19
6	1,3,10,12,16,19
7	1,3,4,10,12,16,19
8	1,3,4,8,10,12,16,19
9	1,3,4,8,10,12,16,17,19
10	1,3,4,8,10,12,14,16,17,19
11	1,3,4,6,8,10,12,14,16,17,19
12	1,3,4,6,8,10,11,12,14,16,17,19
13	1,3,4,6,8,10,11,12,13,14,16,17,19
14	1,3,4,6,8,10,11,12,13,14,15,16,17,19
15	1,3,4,6,8,10,11,12,13,14,15,16,17,19,20
16	1,2,3,4,6,8,10,11,12,13,14,15,16,17,19,20
17	1,2,3,4,6,8,10,11,12,13,14,15,16,17,18,19,20
18	1,2,3,4,5,6,8,10,11,12,13,14,15,16,17,18,19,20
19	1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18,19,20
20	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20

Size(m)	HC(m)	AIC(m)	S C(m)	C(m)	$\sigma^2(m)$
1	0.71258	169.62	173.10	210.16	2.0105
2	0.21207	48.091	55.052	31.860	1.2017
3	<u>0.17736</u>	38.358	<u>48.800</u>	21.081	1.1443
4	0.18081	37.784	51.707	20.359	1.1321
5	0.18503	<u>37.396</u>	54.799	<u>19.861</u>	1.1209
6	0.19600	38.626	59.509	21.060	1.1173
7	0.19553	37.110	61.474	19.438	1.1011
8	0.20754	38.590	66.435	20.907	1.0987
9	0.21913	39.970	71.296	22.275	1.0959
10	0.22912	40.965	75.771	23.254	1.0844
11	0.24322	42.946	81.233	25.236	1.0800
12	0.24683	42.410	84.178	24.679	1.0788
13	0.25996	44.159	89.407	26.426	1.0778
14	0.27321	45.936	94.665	28.204	1.0772
15	0.28727	47.908	100.12	30.176	1.0769
16	0.30079	49.751	105.44	32.019	1.0768
17	0.31494	51.743	110.91	34.011	1.0767
18	0.32909	53.737	116.39	36.005	1.0767
19	0.34325	55.733	121.86	38.000	1.0767
20	0.35742	57.732	127.35	40.000	1.0767
57 Models are Checked Before The Optimum Model is Found					

Table 2.5: Locally Selected AR Subset Models by $\delta^{(0.1)}$ for the Simulated Data

A typical real data set which has been analysed by many authors in the subset AR literature is the log_{10} transformation of annual trappings of lynx in a Canadian region between the year 1812 and 1934. Using McClave's algorithm, Tong (1977), fitted a subset AR model containing lags 1,2,4,10,11 and $\hat{\sigma}^2 = 0.04405$ with maximum lag 11. Using the inverse tree proposed by Furnival and Wilson (1974), Penm and Terrell (1982) fitted the same subset AR model with maximum lag 16. Using Furnival (1971) for subset AR model fitting, Haggan and Oyetunji (1984) developed an efficient method to evaluate the residual variance of all possible subset models, and fitted the Canadian lynx data by the same subset AR model as the one found by Tong. If the maximum lag is over estimated, this will make the subset AR selection more difficult. So, we test the proposed algorithm with maximum lag 16 to fit the Canadian lynx data by a subset AR model. Table 2.6 on page 52 shows the global best subset models obtained by Penm and Terrell (1982) with inverse tree evaluating every subset of a given size and the corresponding criteria. Table 2.7 and 2.8 on page 53 and 54 show the local best subset models chosen by $\delta^{(1/2)}$ and $\delta^{(1/3)}$ and the corresponding criteria respectively.

It can be seen that both searches with $\beta = 1/2$ and $\beta = 1/3$ select the same model as chosen by Tong (1977), Penm and Terrell (1982) and Haggan and Oyetunji (1984) as their optimum model. The local best model for size 7 to 16 and for size 1, 2, 4, 5 are the same in both searches. However, they differ for size 3 and 6. The local best models chosen by the proposed algorithm with $\beta = 1/2$ are the global best for size 1, 2, 4, 5, 10, 15, 16. The local best models chosen by the proposed algorithm with $\beta = 1/3$ are the global best for size 1, 2, 4, 5, 6, 10, 15, 16. The local best model chosen when $\beta = 1/3$ has the greater chance to be the global best from Table 2.6 since the size 6 local best chosen by $\beta = 1/3$ is the global best while the size 6 local best chosen by $\beta = 1/2$ is not the global best because it does not coincide with the size 6 model in Table 2.6.

McClave's algorithm checks more subset models because some small size best models should include some lags which are removed for a large size best model. For

Global Best Subset Models Based on Residual Variances

Size(m)	Lags Chosen by Increase in Residual Variance of All Subset				
1	1				
2	1,2				
3	1,2,9				
4	1,2,10,11				
5	1,2,4,10,11				
6	1,2,3,4,10,11				
7	1,2,4,9,12,13,16				
8	1,2,3,4,9,12,13,16				
9	1,2,3,4,9,12,13,15,16				
10	1,2,3,4,9,10,11,12,13,16				
11	1,2,3,4,9,10,11,12,13,15,16				
12	1,2,3,4,9,10,11,12,13,14,15,16				
13	1,2,3,4,5,6,9,10,11,12,13,15,16				
14	1,2,3,4,5,6,7,9,10,11,12,13,15,16				
15	1,2,3,4,5,6,7,8,9,10,11,12,13,15,16				
16	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16				
Size(m)	HC(m)	AIC(m)	S C(m)	C(m)	$\sigma^2(m)$
1	-1.1741	-133.85	-133.85	740.88	0.30909
2	-2.8085	-322.39	-316.92	47.907	0.57092E-01
3	-2.8820	-331.88	-323.68	34.765	0.51617E-01
4	-2.9690	-342.91	-331.97	21.349	0.46044E-01
5	-2.9860	-345.96	-332.28	17.830	0.44048E-01
6	-2.9722	-345.49	-329.07	18.209	0.43462E-01
7	-2.9567	-344.84	-325.69	18.790	0.42949E-01
8	-2.9413	-344.19	-322.30	19.396	0.42445E-01
9	-2.9195	-342.82	-318.19	20.752	0.42212E-01
10	-2.9015	-341.88	-314.52	21.665	0.41819E-01
11	-2.8794	-340.47	-310.37	23.073	0.41605E-01
12	-2.8535	-338.63	-305.79	24.913	0.41547E-01
13	-2.8275	-336.77	-301.20	26.763	0.41493E-01
14	-2.8022	-334.99	-296.69	28.542	0.41413E-01
15	-2.7776	-333.31	-292.27	30.226	0.41299E-01
16	-2.7523	-331.53	-287.76	32.000	0.41217E-01
149 Models are Checked Before The Optimum Model is Found					

Table 2.6: Globally Selected AR Subset Models for the Lynx Data

Local Best Subset Models Based on $\delta^{(1/2)}$

Size(m)	Lags Chosen by $\delta^{(1/2)}$				
1	1				
2	1,2				
3	1,2,10				
4	1,2,10,11				
5	1,2,4,10,11				
6	1,2,4,10,11,16				
7	1,2,3,4,10,11,16				
8	1,2,3,4,10,11,12,16				
9	1,2,3,4,10,11,12,13,16				
10	1,2,3,4,9,10,11,12,13,16				
11	1,2,3,4,5,9,10,11,12,13,16				
12	1,2,3,4,5,6,9,10,11,12,13,16				
13	1,2,3,4,5,6,7,9,10,11,12,13,16				
14	1,2,3,4,5,6,7,8,9,10,11,12,13,16				
15	1,2,3,4,5,6,7,8,9,10,11,12,13,15,16				
16	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16				
Size(m)	HC(m)	AIC(m)	S C(m)	C(m)	$\sigma^2(m)$
1	-1.1741	-133.85	-133.85	740.88	0.30909
2	-2.8085	-322.39	-316.92	47.907	0.57092E-01
3	-2.8025	-322.82	-314.61	46.577	0.55888E-01
4	-2.9690	-342.91	-331.97	21.349	0.46044E-01
5	-2.9860	-345.96	-332.28	17.830	0.44048E-01
6	-2.9673	-344.93	-328.51	18.800	0.43676E-01
7	-2.9509	-344.18	-325.03	19.482	0.43199E-01
8	-2.9313	-343.05	-321.16	20.577	0.42872E-01
9	-2.9120	-341.96	-317.33	21.632	0.42530E-01
10	-2.9015	-341.88	-314.52	21.665	0.41819E-01
11	-2.8754	-340.01	-309.91	23.532	0.41771E-01
12	-2.8513	-338.37	-305.53	25.172	0.41641E-01
13	-2.8254	-336.53	-300.96	27.007	0.41581E-01
14	-2.8009	-334.85	-296.54	28.688	0.41466E-01
15	-2.7776	-333.31	-292.27	30.226	0.41299E-01
16	-2.7523	-331.53	-287.76	32.000	0.41217E-01
35 Models are Checked Before The Optimum Model is Found					

Table 2.7: Locally Selected AR Subset Models by $\delta^{(1/2)}$ for the Lynx Data

Local Best Subset Models Based on $\delta^{(1/3)}$

Size(m)	Lags Chosen by $\delta^{(1/3)}$				
1	1				
2	1,2				
3	1,2,10				
4	1,2,10,11				
5	1,2,4,10,11				
6	1,2,3,4,10,11				
7	1,2,3,4,9,10,11				
8	1,2,3,4,9,10,11,12				
9	1,2,3,4,9,10,11,12,13				
10	1,2,3,4,9,10,11,12,13,16				
11	1,2,3,4,5,9,10,11,12,13,16				
12	1,2,3,4,5,6,9,10,11,12,13,16				
13	1,2,3,4,5,6,7,9,10,11,12,13,16				
14	1,2,3,4,5,6,7,8,9,10,11,12,13,16				
15	1,2,3,4,5,6,7,8,9,10,11,12,13,15,16				
16	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16				
Size(m)	HC(m)	AIC(m)	S C(m)	C(m)	$\sigma^2(m)$
1	-1.1741	-133.85	-133.85	740.88	0.30909
2	-2.8085	-322.39	-316.92	47.907	0.57092E-01
3	-2.8025	-322.82	-314.61	46.577	0.55888E-01
4	-2.9690	-342.91	-331.97	21.349	0.46044E-01
5	-2.9860	-345.96	-332.28	17.830	0.44048E-01
6	-2.9722	-345.49	-329.07	18.208	0.43462E-01
7	-2.9559	-344.75	-325.60	18.885	0.42984E-01
8	-2.9357	-343.55	-321.66	20.054	0.42683E-01
9	-2.9170	-342.53	-317.91	21.040	0.42317E-01
10	-2.9015	-341.88	-314.52	21.665	0.41819E-01
11	-2.8754	-340.01	-309.91	23.532	0.41771E-01
12	-2.8513	-338.37	-305.53	25.172	0.41641E-01
13	-2.8254	-336.53	-300.96	27.007	0.41581E-01
14	-2.8009	-334.85	-296.54	28.688	0.41466E-01
15	-2.7776	-333.31	-292.27	30.226	0.41299E-01
16	-2.7523	-331.53	-287.76	32.000	0.41217E-01
31 Models are Checked Before The Optimum Model is Found					

Table 2.8: Locally Selected AR Subset Models by $\delta^{(1/3)}$ for the Lynx Data

instance, the "true" lag 10 is ordered as the second smallest in the original ordering system $(1, 2, 4, 12, 3, 9, 16, 13, 11, 6, 15, 7, 5, 8, 10, 14)$ because the squared t- statistics of lag 10's coefficient is the second smallest in the full AR model. Using McClave's algorithm described on page 37, lag 10 is highly likely to be removed for a subset AR model with size less than 14. After the best model is selected for a particular size, the proposed algorithm takes advantage of a re-ordered system to gain efficiency as it begins the choice of the best model with size now reduced by one. The proposed algorithm selects the "true" model as the optimum model after evaluating only about 1/3 of the subset models investigated in McClave's algorithm.

2.5 Summary

The numerical examples presented in the previous section allow us to conclude that

- In selecting the optimum size subset AR in model fitting, using both real and simulated data, the SC criterion performs best.

- The proposed algorithm is much more efficient than McClave's algorithm; it reduces substantially the number of possible subset AR models to be evaluated. The proposed algorithm is much more efficient when fitting subset AR models with a large maximum lag.

- The local best model selection criterion $\delta^{(\beta)}$(see equation (2.63))(when it is applied to a specified size subset AR model) represents a more general criterion than one based on only the increase in residual variance. Choosing $\beta = 1/3$ introduces an option involving the projection modulus rather than choosing $\beta = 1/2$ which uses simply the increase in the residual variance. The proposed subset AR selection algorithm shows out well in preserving the true lags in the local best models when employed for both computer simulated and real data sets. Consequently, it appears that the true model will be found with greater probability using the projection modulus rather than using the increase in the

residual variance as a criterion.

As we mentioned before, the optimum trade-off parameter β is not known, and it seems that the best value of β varies from one unknown true model to another. However, we can still use the proposed algorithm to find the true optimum subset AR model for a data set. The strategy is to set β to different values, such as 1/5, 1/4, 1/3, 1/2, and 1 which introduce different criteria and so different influences on removing lags. The chosen lags of the optimum models produced by different β can then be divided into two sets of lags:

1. A common lag set which contains the lags which belong to all the chosen optimum models;

2. Another lag set which consists of those lags which are not common to all optimum models.

The lags in the common lag set are highly likely to be the lags in the true model since these lags have been preserved for the different values of β, and associated criteria. The common lag set, however, may not include all lags of the true model. The remaining true lags are most likely to be in the uncommon lag set. In most cases for our simulated samples, the uncommon lag set is very small; and only includes 1 or 2 lags. For instance, if there are p lags in the uncommon lag set, we can proceed as follows to extend the range of models considered. Taking the lags which are included in the common set as always being included and then adding extra lags from the uncommon set to produce an expanded set of candidate models. Therefore, the expanded candidate model set contains 2^p models. For each possible model in the expanded set the chosen criterion (SC) is calculated and compared, and thus we can determine which one is likely to be the true model.

For example, the proposed algorithm with different β was applied to a sample data set of length 240 generated from model (2.65), and produced the optimum models presented in the first part of Table 2.9 on page 57. The common lag set is $(1, 3)$; and the uncommon lag set is $(11, 12)$. The common lags 1 and 3 are most likely to

β/lags	common	uncommon	SC
1/5	1, 3	12	-8.5477
1/4	1, 3	12	-8.5477
1/3	1, 3	12	-8.5477
1/2	1, 3	11	-5.1670
1	1, 3	11	-5.1670
Possible True Optimum Model			SC
Model (1,3)			-4.2159
Model (1,3,11)			-5.1670
Model (1,3,12)			-8.5477
Model (1,3,11,12)			-3.2444

Table 2.9: Search for the True Optimum Model from Chosen Models for the Simulated Data

be lags of the true model. The second part of the table shows all possible optimum models and the corresponding values of the model selection criteria SC. Comparing the model selection criterion SC, we can conclude that the subset AR model with lags $(1, 3, 12)$ is highly likely to be the true model.

By using this strategy, the optimum true model has been found in 96 out of the 100 sample data sets. This indicates that this extra strategy option can find the true model with higher probability than occurs with the use of a single optimum trade-off parameter(86 out 100 sample data sets produce the true optimum model when $\beta = 0.1$).

The strategy is applied to the Canadian lynx data and yields the results in Table 2.10 on page 58. The common lag set is $(1, 2, 4, 10, 11)$; and the uncommon lag set is (16). There are only two possible true optimum models, and their model selection criteria SC listed in the second part of Table 2.10. It is obvious that the subset AR model with lags $(1, 2, 4, 10, 11)$ is most likely to be the optimum model, since it has the smaller SC value. Assuming the selected optimum model is the "true" model and finding the size 5 subset AR model from the model with lags $(1, 2, 4, 10, 11, 16)$ which is the size 6 local best chosen by $\beta \leq 1/5$, we list some statistics for lag 2 and 6 and $\delta^{(\beta)}$ in Table 2.11. Comparing the values $\delta^{(\beta)}$ of lag 2 and 6 in Table 2.11, we

β/lags	common	uncommon	SC
1/5	1,2,4,10,11	16	-328.51
1/4	1,2,4,10,11	null	-332.28
1/3	1,2,4,10,11	null	-332.28
1/2	1,2,4,10,11	null	-332.28
1	1,2,4,10,11	null	-332.28

Possible True Optimum Model	SC
Model (1,2,4,10,11)	-332.28
Model (1,2,4,10,11,16)	-328.51

Table 2.10: Search for the True Optimum Model from Chosen Models from the Lynx Data

lag i	$\hat{a}_{(1,2,4,10,11,16)}(i)$	$\langle e_i, e_i \rangle$	$t-$ ratio
2	0.37438	0.14545	3.35587
16	0.04951	0.49184	0.8167

lag	$\delta^{(\beta)}$					
i	$\beta = 0$	$\beta = 1/5$	$\beta = 1/4$	$\beta = 1/3$	$\beta = 1/2$	$\beta = 1$
2	-3.8558	-3.8706	-3.8744	-3.8805	-3.8929	-3.9299
16	-1.4192	-3.5399	-4.0700	-4.9536	-6.7209	-12.0225

Table 2.11: β Weight Effects on Criterion $\delta^{(\beta)}$

can see that the size 5 local best model with lags $(1, 4, 10, 11, 16)$ will be chosen if the trade-off parameter β is less than $1/4$; while the size 5 local best model with lags $(1, 2, 4, 10, 11)$ will be chosen if the trade-off parameter β is greater or equal to $1/4$. It is obvious that the proposed procedure will miss the "true" model when $\beta \leq 1/5$ since the trade-off parameter β over-emphasizes the impact of removing a particular lag on the rest of remaining lags. In other words, lag 2 should stay instead of lag 16 in finding the size 5 local best model and so the "true" model can be selected although lag 2 can be better represented by lags $(1, 4, 10, 11, 16)$ than can lag 16 by lags $(1, 2, 4, 10, 11)$. This illustrates an instance of how the range of the trade-off parameter β is critical in preserving the true lags.

This strategy, which searches a grid of β values, always produces the same or a higher probability of selecting the true model than can be obtained with a single

chosen value of the trade-off parameter β. However, this advantage is at the cost of lessened efficiency since this strategy needs more than five times the number of operations associated with using a single trade-off parameter.

If you choose the more efficient approach there remains a major task of choosing an appropriate β for $\delta^{(\beta)}$ to make it highly likely that the true lags will be preserved; and so select the true optimum model.

Chapter 3

On State Space Model

3.1 Introduction

The structural state space model is a more flexible statistical model which can provide a description of a time series in terms of its components of interests, such as, trend component, seasonal component, etc.. With the utilization of the Kalman filter technique, the structural state space model can handle non-stationary time series, and the state space model can be identified. In other words, the unknown parameters in the state space model can be estimated via various criteria, such as, maximum likelihood, mean square error of one step head prediction, etc.

After constructing a structural state space model for a practical data set, we always face two problems: (1) the initial conditions of the state vector; (2) the unknown system parameters, such as, the values of some elements of the system matrices, the covariance matrix of the state disturbances. When the initial state vector is normally distributed, the Kalman filter allows the likelihood function of the state space model to be calculated via what is known as the prediction error decomposition. In this way, the unknown system parameters in the model can be estimated. If the normality condition for the initial state vector is dropped, there is no longer any guarantee that the Kalman filter will give the conditional mean of the state vector and the system parameter estimation correctly. Suppose that the system parameters are all known, Caines and Meyne (1970) and, Anderson and Moore (1979) (see Theorem 3.2 and 3.3)

give sufficient conditions on system matrices and the initial state covariance matrix, and guarantee that the state covariance matrix converges exponentially fast. The properties of state space models which are not stabilizable have been examined by Chan et al. (1984) with particular reference to cases where some of the roots of the transition matrix lie on the unit circle. However, the system parameters are usually unknown in practice or only partially known and need to be estimated.

In section 3.2, we give a general form of a state space model, some pre-specified assumptions and the likehood function of the model. In section 3.3, we discuss the influence of initial conditions for the state vector on the performance of the Kalman filter, generalize the *Increasing and Decreasing Properties* of the state vector covariance matrix in Theorem 3.5 and 3.6, where the results from Chan et al. (1984) is a special case of Theorem 3.6, and conclude in Theorem 3.7 that an over-estimated initial state covariance matrix leads to a faster convergence speed in general than does an under-estimated one. In section 3.4, we investigate the fixed point smoothing and estimation of the initial state conditions, and show that the fixed point smoothing is an efficient Bayesian estimation for the initial conditions of the state vector. After the sensitivity analysis in section 3.5 for the state space model, we develop a state space model estimation procedure for the unknown system parameters in section 3.6. The application of this procedure will be presented in chapter 7.

3.2 Model and Assumptions

We consider a stochastic process $\{y(t), t \geq 1\}$ generated by the state space model

$$\begin{cases} x(t+1) = A(\theta)x(t) + \xi(t) & \text{Transition equation} \\ y(t) = C(\theta)x(t) + \epsilon(t) & \text{Observation equation} \end{cases} \tag{3.1}$$

Assumption 3.1 *(i)* $\{x(t), t \geq 1\}$ *is a sequence of* $m \times 1$ *state vectors.*

(ii) $\forall t$, $\xi(t)$ *and* $\epsilon(t)$ *are, respectively,* $m \times 1$ *and* 1×1 *Gaussian random disturbances.* $\forall t$, $s \in T^+$, $\xi(t)$, $\epsilon(t)$ *are conditionally independent of* $x(0)$ *and of given*

$y^{(t-1)}$ *where* $y^{(t-1)} = \{y(0), \cdots, y(t-1)\}$. *Both have zero mean and have variance-covariance matrices*

$$\mathbf{E}\left\{\begin{pmatrix} \xi(s) \\ \epsilon(s) \end{pmatrix} (\xi(t), \epsilon(t))\right\} = \begin{pmatrix} Q & S \\ S' & R \end{pmatrix} \delta(t-s) \tag{3.2}$$

and $R > 0$ *and* Q *can be decomposed into* $Q = BB'$ *where* $T^+ = \{1, 2, 3, \cdots, T\}$.

(iii) The matrices $A(\theta)$ *and* $C(\theta)$ *are* $n \times n$ *and* $1 \times m$, *respectively.*

Assumption 3.2 (Initial Conditions)

$$x(0) = \Psi(\psi)\zeta + \eta \tag{3.3}$$

where (i) $\Psi(\psi)$ *is an* $m \times q$ $(q \leq m)$ *matrix. (ii)* $\eta \sim \mathbf{N}(0, \Sigma_\eta(\psi))$ *and* η *is an* $m \times 1$ *vector having an unspecified distribution. Both* η *and* ζ *are independent of* $\delta(t)$ *and* $\epsilon(t)$.

ψ *is a parameter vector belonging to a subset* Ψ *of a finite dimensional Euclidean space.*

We observe $y(t\ )$, $t \in T^+$. Now let $\gamma(0) = \eta$ and for $t > 1$ define

$$F(t, \theta, \psi) = C(\theta)A^{t-1}(\theta)\Psi \tag{3.4}$$

$$\begin{cases} \gamma(t+1) = A(\theta)\gamma(t) + \xi(t) \\ \omega(t) = C(\theta)\gamma(t) + \epsilon(t) \end{cases} \tag{3.5}$$

Then

$$y(t) = F(t, \theta, \psi)\zeta + \omega(t) \quad t \geq 1 \tag{3.6}$$

When the distribution of ζ is known, the Kalman filter applied to the model yields the mean and covariance matrix of the distribution of $x(t)$ conditional on the information available at time $t-1$. Thus

$$x(t|t-1) = \mathbf{E}[x(t)|y^{(t-1)}]$$

where $y^{(t-1)} = \{y(t-1), \cdots, y(1)\}$ and $\Sigma(t) = \mathbf{E}[x(t) - x(t|t-1)][x(t) - x(t|t-1)]'$.

If $y(t)$ is stationary, then the system has a stable transition matrix with $|\lambda_i(A(\theta))| < 1$. The initial values of $x(0)$ or the distribution of ζ does not affect $y(t)$ when $t \to \infty$ because $\lim_{t\to\infty} F(t, \theta, \psi) = 0$. In principle, the starting state vector values for a Kalman filter which is employed to estimate an unobservable state vector are given by the mean and covariance matrix of the unconditional distribution of the state vector. However, structural time series models are used often to model economic time series which can be non-stationary.

Note:

The concepts of *reachability, controllability* and *stabilizability* are all concerned with the interconnection between the input and state of a system. Intuitively, these concepts describe to what extent the system state can be steered by using the system input. On the other hand, the concepts of *reconstructibility, observability* and *detectability* are all concerned with the interaction between the system output and state. Intuitively, these concepts describe what part of the system can be seen from the output. The formal definitions for the above concepts can be found in Anderson and Moore (1979), Caines (1988), etc. and would not be presented here.

$A(\theta)$ and $C(\theta)$ indicate that the system matrices of the state space model may not be fully known. For reason of simplicity, the notation A and C will be used in the following sections.

3.2.1 The Exact Likelihood Function

Now in order to construct the exact likelihood function for a given nonstationary process $\{y(t)|1 \leq t \leq T\}$, let us assume that the random variable $y(t)$ has a density that depends on the deterministic quantities θ for all t. From Baye s'rule or equivalently from the definition of a conditional density and Schweppe (1965), we obtain

$$f(y^{(T)}|\theta) = f(y^{(T)}|y^{(T-1)}, \theta) f(y^{(T-1)}|\theta)$$

$$= \prod_{t=1}^{T} f(y(t)|y^{(t-1)}, \theta) \tag{3.7}$$

We see that $f(y^{(T)}|\theta)$ in the above equation is a likelihood function for the $\{y(t)|1 \leq t \leq T\}$ and that the likelihood function is parameterized by deterministic quantities θ.

Now let

$$\bar{y}_{t|t-1}(\theta) = \int_{R(t)} y(t) f(y(t)|y^{(t-1)}, \theta) dy(t) \tag{3.8}$$

that is $\bar{y}_{t|t-1}(\theta)$ is the conditional expectation of $y(t)$ given $y^{(t-1)}$. Making the change of variable

$$y(t) \rightarrow y(t) - \bar{y}_{t|t-1}(\theta), \quad 1 \leq t \leq T \tag{3.9}$$

we obtain

$$f(y^{(T)}|\theta) = \prod_{t=1}^{T} f(y(t) - \bar{y}_{t|t-1}(\theta)|y^{(t-1)}, \theta) \tag{3.10}$$

since the determinant of the Jacobian of the change of variable is the identity matrix.

The above equation yields a likelihood function for the observations of the form

$$exp \sum_{t=1}^{T} \log f(v_t(\theta)|y^{(t-1)}, \theta) + \{\text{function of } y(0)\} \tag{3.11}$$

where $v_t(\theta) = y(t) - \bar{y}_{t|t-1}(\theta)$ is an innovation sequence.

Since $y(0) = Cx(0) + \epsilon(0)$ and $x(0) = \Psi(\psi)\zeta + \eta$, the above likelihood function can be written

$$exp \sum_{t=1}^{T} \log f(v_t(\theta)|y^{(t-1)}, \theta) + \{\text{function of } \zeta \text{ and } \psi \text{ not depending upon } \theta\} \tag{3.12}$$

From here, we can clearly see that (1) the likelihood function can be expressed as density function of innovations plus a function of an initial condition; (2) the likelihood function is affected not only by the parameter set θ but also by initial conditions in general.

If the disturbances and initial state vector in model (3.1) have proper multivariate normal distributions, the distribution of $y(t)$, conditional on $(y^{(t-1)}, \theta)$ is itself normal.

Furthermore, the mean and covariance matrix of this conditional distribution are given directly by the Kalman filter. Conditional on $(y^{(t-1)}, \theta)$, the state vector $x(t)$ is normally distributed with a mean of $x(t|t-1)$ and a covariance matrix of $\Sigma(t)$. The innovation process $v(t)$ is Gaussian with density $N(0, \Sigma_v(t))$ where $\Sigma_v(t) = C\Sigma(t)C' + R$. The following expression gives the first part of the logarithm of the likelihood function

$$-\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}[\log|\Sigma_v(t)| + v'(t)\Sigma_v^{-1}(t)v(t)] \tag{3.13}$$

The choice of the initial state vector will often significantly influence the evaluation of the likelihood function and is also related to *a priori* knowledge of the state vector. It is reasonable to take $\hat{x}(0)$ as an *a priori* estimate of $x(0)$ and let $\Sigma(0)$ reflect the confidence in this initial estimate, $\hat{x}(0)$. If $\Sigma(0)$ is small the Kalman filter gain $K(t)$ will be small for all t and the state vector estimates will therefore not change too much from $\hat{x}(0)$. On the other hand if $\Sigma(0)$ is big, the vector estimates will quickly jump away from $\hat{x}(0)$.

3.3 The Influence of the Initial State Covariance Matrix

Note that none of the assumptions in Assumption 3.1 of the state space model (3.1) imply the stability properties of the system or the stationarity of the output.

If $x(0)$ has the conditional gaussian distribution $\mathbf{N}(\bar{x}(0), \Sigma_0)$, the innovation representation is

$$\begin{cases} x(t+1|t) = Ax(t|t-1) + K(t)v(t) \\ y(t) = Cx(t|t-1) + v(t) \end{cases} \tag{3.14}$$

where $v(t) = y(t) - Cx(t|t-1) = C(x(t) - x(t|t-1)) + \epsilon(t)$ and $K(t)$ is the Kalman gain.

Define

$$\tilde{x}(t+1) \;=\; x(t+1) - x(t+1|t)$$

$$
\begin{aligned}
&= A\tilde{x}(t) + \xi(t) - K(t)[C\tilde{x}(t) + \epsilon(t)] \\
&= [A - K(t)C]\tilde{x}(t) + \xi(t) + K(t)\epsilon(t)
\end{aligned}
$$

and

$$\Sigma(t+1) = \mathbf{E}\tilde{x}(t+1)\tilde{x}(t+1)'$$

Hence

$$
\begin{aligned}
v(t) &= y(t) - Cx(t|t-1) \\
&= C\tilde{x}(t) + \epsilon(t) \qquad (3.15)
\end{aligned}
$$

Since $\tilde{x}(t+1) \perp v(t)$, $\tilde{x}(t) \perp \xi(t)$ and $\tilde{x}(t) \perp \epsilon(t)$, we have that

$$
\begin{aligned}
0 &= \mathbf{E}\tilde{x}(t+1)v(t) \\
&= \mathbf{E}[A\tilde{x}(t) + \xi(t) - K(t)(C\tilde{x}(t) + \epsilon(t))][C\tilde{x}(t) + \epsilon(t)]' \\
&= A\Sigma(t)C' + S - K(t)[C\Sigma(t)C' + R]
\end{aligned}
$$

Thus

$$K(t) = [A\Sigma(t)C' + S][C\Sigma C' + R]^{-1} \qquad (3.16)$$

$$
\begin{aligned}
\Sigma(t+1) &= \mathbf{E}\tilde{x}(t+1)\tilde{x}'(t+1) \\
&= A\Sigma(t)A' - A\Sigma(t)C'K'(t) + Q - SK'(t) \\
&\quad -K(t)C\Sigma(t)A' - K(t)S' + K(t)[C\Sigma(t)C' + R]K'(t) \\
&= A\Sigma(t)A' - K(t)[C\Sigma(t)C' + R]K'(t) + Q \qquad (3.17) \\
&= [A - K(t)C]\Sigma(t)[A - K(t)C]' \\
&\quad +[K(t) - SR^{-1}]R[K(t) - SR^{-1}]' \\
&\quad +Q - SR^{-1}S' \qquad (3.18)
\end{aligned}
$$

The above equation is called the Riccati Difference Equation (RDE). If the Σ is time invariant, the equation becomes

$$
\begin{aligned}
\Sigma &= [A - KC]\Sigma[A - KC]' + [K - SR^{-1}]R[K - SR^{-1}]' \\
&\quad +Q - SR^{-1}S' \qquad (3.19)
\end{aligned}
$$

where $K = [A\Sigma C' + S][C\Sigma C' + R]^{-1}$, is called the Algebraic Riccati Equation (ARE).

In the last section, we have shown that the initial state vector $x(0)$ and its covariance matrix $\Sigma(0)$ affect the likelihood estimation for the system parameter set $\Theta = (A, C, Q, R, S)$.

There are two questions which have arisen and need to be answered (1) How does the initial state variance matrix affect the Kalman filter performance ? (2) How do errors in the system parameter set Θ affect the Kalman filter performance ?

The answers to the above questions are important for estimates of Θ. Especially, in practice, when only a small sample data set is available to estimate Θ. For instance, a maximum likelihood estimation is an estimation in the criterion of the highest probability of occurrence. However, it is not efficient due to the complex nature of the likelihood function and of the first and second derivatives of the parameter vector Θ for a general form of the state space model. So, numerical approximations have to be used to evaluate these quantities in practice. The estimation procedure is computationally intensive when a large data set is involved. Therefore, we employ an initial part of a whole data set to estimate parameters in a specifically parameterized state space form model, and then initiate the Kalman filtering process with an on-line parameter estimation algorithm to finally adjust the parameters. However, the initial state variance matrix affects the Kalman filtering considerably during the likelihood estimation because a small data set is involved even though the specified form of the state space model satisfies the various conditions needed to ensure the state covariance matrix converges.

In this section, we are going to prove that an over-estimated initial variance matrix leads to a relative faster convergence speed of the state variance than an under-estimated initial variance matrix.

Theorem 3.1 (Minimum Property)

Let two initial state covariance matrices $\Sigma_1(0)$ *and* $\Sigma_2(0)$ *satisfy* $\Sigma_1(0) \geq \Sigma_2(0)$, *and let the sequence* $\{\Sigma_1(t),\ t \in T^+\}$ *with initial value* $\Sigma_1(0)$ *be generated by the*

recursive scheme

$$\begin{aligned}\Sigma_1(t+1) &= [A-\Delta(t)C]\Sigma_1(t)[A-\Delta(t)C]' \\ &\quad +[\Delta(t)-SR^{-1}]R[\Delta(t)-SR^{-1}]'+Q-SR^{-1}S' \qquad (3.20)\end{aligned}$$

where $\Delta(t)$ *is an arbitrary matrix for each* $t \in T^+$. *Further let the sequence* $\{\Sigma_2(t),\ t \in T^+\}$ *with initial value* $\Sigma_2(0)$ *be generated by the RDE*

$$\begin{aligned}\Sigma_2(t+1) &= [A-K(t)C]\Sigma_2(t)[A-K(t)C]' \\ &\quad +[K(t)-SR^{-1}]R[K(t)-SR^{-1}]'+Q-SR^{-1}S' \qquad (3.21)\end{aligned}$$

where $K(t)$ *is defined by* $K(t) = [A\Sigma_2(t)C' + S][C\Sigma_2(t)C' + R]^{-1}$ *and is called the Kalman gain.*

Then we have,

$$\Sigma_1(t) \geq \Sigma_2(t) \geq 0$$

proof: See Caines (1988) pp.196 □

Corollary 1 (Boundary Property)

If two initial covariance matrices, $\Sigma_1(0)$ *and* $\Sigma_2(0)$, *satisfy* $\Sigma_1(0) \geq \Sigma_2(0) \geq 0$. *The sequences* $\{\Sigma_1(t),\ t \in T^+\}$ *and* $\{\Sigma_2(t),\ t \in T^+\}$ *are generated by the RDE,*

$$\begin{aligned}\Sigma_i(t+1) &= [A-K_i(t)C]\Sigma_i(t)[A-K_i(t)C]' \\ &\quad +[K_i(t)-SR^{-1}]R[K_i(t)-SR^{-1}]'+Q-SR^{-1}S' \qquad (3.22)\end{aligned}$$

where $K_i(t)$ *is defined by* $K_i(t) = [A\Sigma_i(t)C' + S][C\Sigma_i(t)C' + R]^{-1}$ *and is the Kalman gain with the initial covariance matrix* $\Sigma_i(0)$, $i = 1, 2$. *Then we have the relationship,*

$$\Sigma_1(t) \geq \Sigma_2(t) \geq 0$$

proof: Set $\Delta(t)$ in the minimum property theorem equal to the Kalman gain with the initial state covariance matrix $\Sigma_1(0)$

□

Theorem 3.2 *Subject to the assumptions*

(1) (A, B) is stabilizable

(2) (C, A) is detectable

(3) $\Sigma_0 \geq 0$

then $\lim_{t\to\infty} \Sigma(t) = \Sigma$

where $\Sigma(t)$ *is the solution of the RDE with initial covariance matrix* Σ_0 *and where* Σ *is the unique stabilizing solution of the ARE.*

proof: See Caines and Meyne (1970) or Anderson and Moore (1979)

□

Theorem 3.3 *Subject to the assumptions*

(1) There is no uncontrollable modes of (A, B) on the unit circle.

(2) (C, A) is detectable

(3) $\Sigma_0 \geq 0$

Then $\lim_{t\to\infty} \Sigma(t) = \Sigma$

proof: See Chan et al. (1984)

□

Theorem 3.2 and 3.3 ensure that the state covariance matrix sequence $\{\Sigma(t),\ t \in T^+\}$ generated by the RDE will converge to a steady covariance matrix Σ which satisfies the ARD when the conditions are met. In general, we do not know the exact initial state covariance matrix corresponding to the initial state. There are two ways to deal with the unknown initial state covariance matrix $\Sigma(0)$. (1) set $\Sigma(0) = 0$ (an under-estimated initial state variance matrix), or (2) set $\Sigma(0) = kI$ (an over estimated initial state variance matrix) where k is "big enough" to approximate the diffuse initial condition or lack of any information a priori. Theoretically, the state variance matrix will converge to a steady state no matter which of the above initial state variance matrix assumptions are used. In practice, however, we usually have some partial information about the initial conditions. For instance, we may know the lower and upper bounds of the eigenvalues of the initial state covariance matrix. A

problem can arise in using the partial information to ensure that the state covariance matrix converges quickly to the steady state. Is it best to use an over estimated initial state covariance matrix or an under estimated one? The following theorem and corollaries give us further analytical results to assist with this decision.

We deal with the under-estimated initial state matrix first.

Theorem 3.4 (Increasing Property)

(1) If the initial state covariance matrix $\Sigma_0(0) = 0$, *then the sequence* $\{\Sigma_0(t),\ t \in T^+\}$ *generated by the RDE is monotonically increasing.*

(2) If, in addition, $(C,\ A)$ *is detectable, then the sequence* $\{\Sigma_0(t),\ t \in T^+\}$ *is bounded and converges to a matrix* Σ *that satisfies the ARE*

$$\Sigma = [A - KC]\Sigma[A - KC]' + [K - SR^{-1}]R[K - SR^{-1}]' + Q - SR^{-1}S' \quad (3.23)$$

where K *is defined by* $K = [A\Sigma C' + S][C\Sigma C' + R]^{-1}$.

proof: See Anderson and Moore (1979), pp. 81 and Caines (1988) pp. 171

□

Lemma 3.1 *Subject to the assumptions*

(1) (C,A) is detectable.

(2) There exists a solution Σ *for the ARE.*

Then, the Kalman gain sequence $\{K(t), t \in T^+\}$ *and the state error covariance matrix sequence* $\{\Sigma(t),\ t \in T^+\}$ *has the following relationships with the steady-state Kalman gain,* K, *and the steady-state error covariance matrix* Σ:

$$K(t) = K + [(A - KC)\Sigma(t)C + S - KR](C\Sigma(t)C' + R)^{-1}$$

$$D(t+1) = (A - KC)[D(t) - D(t)C'(C\Sigma(t)C' + R)^{-1}CD(t)](A - KC)^{-1}$$

where $D(t) = \Sigma - \Sigma(t)$ *when* $\Sigma(0) \leq \Sigma$, *and* $D(t) = \Sigma(t) - \Sigma$ *when* $\Sigma(0) \geq \Sigma$.

proof:

Without loss of generality, we assume $\Sigma(0) \leq \Sigma$. By the Minimum Property (Theorem 3.1), we know that $\Sigma(t) \leq \Sigma$.

Considering the model (3.1) we have

$$\begin{aligned} x(t+1) &= Ax(t) + \xi(t+1) - K(y(t) - y(t)) \\ &= Ax(t) + \xi(t+1) - K(Cx(t) + \epsilon(t) - y(t)) \\ &= \bar{A}x(t) + \xi^*(t+1) + Ky(t) \end{aligned} \tag{3.24}$$

where $\bar{A} = (A - KC)$, $\xi^*(t+1) = \xi(t+1) - K\epsilon(t)$.

The equivalent representation of model (3.1) is

$$\begin{cases} x(t+1) = \bar{A}x(t) + Ky(t) + \xi^*(t+1) \\ y(t) = Cx(t) + \epsilon(t) \end{cases} \tag{3.25}$$

The covariance matrix of the disturbance becomes

$$\begin{aligned} \mathbf{E}\left\{ \begin{pmatrix} \xi^*(s) \\ \epsilon(s) \end{pmatrix} (\xi^{*'}(t), \epsilon'(t)) \right\} &= \begin{pmatrix} I & -K \\ 0 & I \end{pmatrix} \begin{pmatrix} Q & S \\ S' & R \end{pmatrix} \begin{pmatrix} I & S \\ -K' & I \end{pmatrix} \delta(t-s) \\ &= \begin{pmatrix} \bar{Q} & \bar{S} \\ \bar{S}' & R \end{pmatrix} \delta(t-s) \end{aligned} \tag{3.26}$$

where $\bar{Q} = Q - KS' - SK' + KRK'$, $\bar{S} = S - KR$.

The innovation representation is

$$\begin{cases} x(t+1|t) = \bar{A}x(t|t-1) + Ky(t) + \bar{K}(t)v(t) \\ y(t) = Cx(t|t-1) + v(t) \end{cases} \tag{3.27}$$

where $v(t) = y(t) - Cx(t|t-1) = C(x(t) - x(t|t-1)) + \epsilon(t)$, $\bar{K}(t)$ is the Kalman gain.

Because $\tilde{x}(t+1)$ can be expressed by,

$$\begin{aligned} \tilde{x}(t+1) &= x(t+1) - x(t+1|t) \\ &= \bar{A}x(t) + \xi^*(t+1) + Ky(t) - \bar{A}x(t|t-1) - Ky(t) + \bar{K}(t)v(t) \\ &= \bar{A}x(t) + \xi^*(t+1) - \bar{A}x(t|t-1) + \bar{K}(t)v(t) \\ &= \bar{A}\tilde{x}(t) + \xi^*(t+1) + \bar{K}(t)v(t) \end{aligned} \tag{3.28}$$

we proceed in the same fashion for $K(t)$, and $\Sigma(t+1)$ in the original model (see equation (3.16), (3.18)) since $\tilde{x}(t+1) \perp v(t)$, $\tilde{x}(t) \perp \xi^*(t)$ and $\tilde{x}(t) \perp \epsilon(t)$, and so we have

$$\Sigma(t+1) = \bar{A}\Sigma(t)\bar{A}' + \bar{Q} + \bar{K}(t)G(t)\bar{K}' \tag{3.29}$$

where $G(t) = C\Sigma(t)C' + R$, and $\bar{K}(t)$ is given by

$$\bar{K}(t) = (\bar{A}\Sigma(t)C' + \bar{S})G(t)^{-1} \tag{3.30}$$

and is called the Kalman gain.

Now, we look at the steady-state Kalman filter of the modified model (3.27), with $\bar{K}$ defined in equation (3.30). So we have

$$\begin{aligned}
\bar{K} &= (\bar{A}\Sigma C' + \bar{S})G^{-1} \\
&= [(A - KC)\Sigma C' + S - KR]G^{-1} \\
&= [(A - (A\Sigma C' + S)(C\Sigma C' + R)^{-1}C)\Sigma C' \\
&\quad + S - (A\Sigma C' + S)(C\Sigma C' + R)^{-1}R]G^{-1} \\
&= [A\Sigma C' + S - (A\Sigma C' + S)(C\Sigma C' + R)^{-1}(C\Sigma C' + R)]G^{-1} \\
&= 0
\end{aligned} \tag{3.31}$$

since the steady state Kalman gain of the original model is $K = (A\Sigma C' + S)(C\Sigma C' + R)^{-1}$. We find that

$$\begin{aligned}
\bar{K}(t) &= (\bar{A}\Sigma(t)C' + \bar{S})G^{-1}(t) \\
&= [(A - KC)\Sigma(t)C' + S - KR]G^{-1}(t) \\
&= (A\Sigma(t)C' + S)G^{-1}(t) - K(C\Sigma(t)C' + R)G^{-1}(t) \\
&= K(t) - K
\end{aligned} \tag{3.32}$$

So, actually we have,

$$\begin{aligned}
K(t) &= K + \bar{K}(t) \\
&= K + [(A - KC)\Sigma(t)C + S - KR](C\Sigma(t)C' + R)^{-1}
\end{aligned} \tag{3.33}$$

By replacing $\bar{K}(t)$ in equation (3.29), we obtain

$$\begin{aligned}
\Sigma(t+1) &= \bar{A}\Sigma(t)\bar{A}' + \bar{Q} + \bar{A}\Sigma(t)C'G^{-1}(t)C\Sigma(t)\bar{A} \\
&\quad -\bar{A}\Sigma(t)C'G^{-1}(t)\bar{S}' - \bar{S}G^{-1}(t)C\Sigma(t)\bar{A}' - \bar{S}G^{-1}(t)\bar{S}' \\
&= \bar{A}[\Sigma(t) - \Sigma(t)C'G^{-1}(t)C\Sigma(t)]\bar{A}'
\end{aligned}$$

$$
\begin{aligned}
& -\bar{A}\Sigma(t)C'G^{-1}(t)\bar{S}' - \bar{S}G^{-1}(t)C\Sigma(t)\bar{A}' \\
& -\bar{S}G^{-1}(t)\bar{S}' + \bar{Q} \\
= \; & \bar{A}[\Sigma(t) - \Sigma(t)C'G^{-1}(t)C\Sigma(t)]\bar{A}' \\
& -\bar{K}(t)\bar{S}' - \bar{S}\bar{K}'(t) + \bar{S}G^{-1}(t)\bar{S}' + \bar{Q} \qquad (3.34)
\end{aligned}
$$

We define

$$\Sigma^{(1)}(t+1) = \bar{A}[\Sigma(t) - \Sigma(t)C'G^{-1}(t)C\Sigma(t)]\bar{A}' \qquad (3.35)$$

$$\Sigma^{(2)}(t+1) = -\bar{K}(t)\bar{S}' - \bar{S}\bar{K}'(t) + \bar{S}G^{-1}(t)\bar{S}' + \bar{Q} \qquad (3.36)$$

Therefore we have,

$$\Sigma(t+1) = \Sigma^{(1)}(t+1) + \Sigma^{(2)}(t+1) \qquad (3.37)$$

$$\Sigma^{(1)} = \bar{A}[\Sigma - \Sigma C'G^{-1}C\Sigma]\bar{A}' \qquad (3.38)$$

$$\Sigma^{(2)} = \bar{S}G^{-1}\bar{S}' + \bar{Q} \qquad (3.39)$$

since $\bar{K} = 0$, and

$$\Sigma = \Sigma^{(1)} + \Sigma^{(2)} \qquad (3.40)$$

Now, we examine $D(t)$ by considering

$$
\begin{aligned}
\Sigma^{(1)} - \Sigma^{(1)}(t+1) &= \bar{A}[D(t) - (\Sigma C'G^{-1}C\Sigma - \Sigma(t)C'G^{-1}C\Sigma(t))]\bar{A}' \\
&= \bar{A}[D(t) - D(t)C'G^{-1}(t)CD(t)]\bar{A}' \\
&\quad +\bar{A}[\Sigma(t)C'G^{-1}(t)C\Sigma - \Sigma C'G^{-1}(t)C\Sigma \\
&\quad +\Sigma C'G^{-1}(t)C\Sigma(t) - \Sigma C'G^{-1}C\Sigma]\bar{A}' \qquad (3.41)
\end{aligned}
$$

By noting that $\bar{K}(t)$ is given by

$$
\begin{aligned}
& \bar{K}(t) = (\bar{A}\Sigma(t)C' + \bar{S})G^{-1}(t) \\
\Rightarrow \quad & A\Sigma(t)C' = \bar{K}(t)G(t) - \bar{S} \qquad (3.42)
\end{aligned}
$$

and $$A\Sigma C' = -\bar{S} \tag{3.43}$$

Substituting equations (3.42), (3.43) into equation (3.41), we produce,

$$\begin{aligned}
\Sigma^{(1)} - \Sigma^{(1)}(t+1) &= \bar{A}[D(t) - D(t)C'G^{-1}(t)CD(t)]\bar{A}' \\
&\quad +(\bar{K}(t)G(t) - \bar{S})G^{-1}(t)(-\bar{S}') - (-\bar{S})G^{-1}(t)(-\bar{S}') \\
&\quad +(-\bar{S})G^{-1}(t)(\bar{K}(t)G(t) - \bar{S}') - (-\bar{S})G^{-1}(-\bar{S}') \\
&= \bar{A}[D(t) - D(t)C'G^{-1}(t)CD(t)]\bar{A}' \\
&\quad -\bar{K}(t)\bar{S}' - \bar{S}\bar{K}'(t) + \bar{S}G^{-1}(t)\bar{S}' - \bar{S}G^{-1}\bar{S}' \\
&= \bar{A}[D(t) - D(t)C'G^{-1}(t)CD(t)]\bar{A}' \\
&\quad +\Sigma^{(2)}(t+1) - \Sigma^{(2)}
\end{aligned} \tag{3.44}$$

Hence, we have the recursive expression for $D(t+1)$

$$\begin{aligned}
D(t+1) &= \Sigma^{(1)} + \Sigma^{(2)} - \Sigma^{(1)}(t+1) - \Sigma^{(2)}(t+1) \\
&= \bar{A}[D(t) - D(t)C'G^{-1}(t)CD(t)]\bar{A}'
\end{aligned} \tag{3.45}$$

□

Theorem 3.5 (General Increasing Property)

Subject to the assumptions,

(1) (C,A) is detectable.

(2) there exists a solution, Σ, for the ARE.

(3) the initial state covariance matrix $\Sigma(0)$ satisfies $0 \leq \Sigma(0) \leq \Sigma$.

then the sequence $\{\Sigma(t),\ t \in T^+\}$ generated by the RDE with initial state covariance matrix $\Sigma(0)$ is monotone increasing and converges to Σ.

proof: We prove the convergence of the sequence $\{\Sigma(t),\ t \in T^+\}$ first. Consider the RDE with initial state covariance matrix, $\Sigma(0)$

$$\begin{aligned}
\Sigma(t+1) &= [A - K(t)C]\Sigma(t)[A - K(t)C]' \\
&\quad +[K(t) - SR^{-1}]R[K(t) - SR^{-1}]' + Q - SR^{-1}S'
\end{aligned} \tag{3.46}$$

and the ARE is

$$\Sigma = [A - KC]\Sigma[A - KC]' + [K - SR^{-1}]R[K - SR^{-1}]' + Q - SR^{-1}S' \quad (3.47)$$

Using the RDE, the ARE and the Minimum Property, we have

$$\begin{aligned} 0 &\leq [A - K(t)C][\Sigma - \Sigma(t)][A - K(t)C]' \\ &\leq \Sigma - \Sigma(t+1) \\ &\leq [A - KC][\Sigma - \Sigma(t)][A - KC]' \end{aligned} \quad (3.48)$$

and through induction, we obtain

$$\Lambda(t)[\Sigma - \Sigma(0)]\Lambda'(t) \leq \Sigma - \Sigma(t+1) \leq \bar{A}[\Sigma - \Sigma(0)]\bar{A}' \quad (3.49)$$

where $\Lambda(t) = \bar{A}(0)\bar{A}(1)\cdots\bar{A}(t-1)$, $\bar{A}(t) = A - K(t)C$ and $\bar{A} = A - KC$.

Since the detectability of the pair (C, A) implies that $|\lambda_i(\bar{A})| < 1$ for $i = 1, 2, \cdots, m$, we have

$$\lim_{t\to\infty}[\bar{A}]^t[\Sigma - \Sigma(0)][\bar{A}']^t = 0$$

thus, $\lim_{t\to\infty}\Sigma(t) = \Sigma$.

Now, we prove that $\Sigma(t)$ increases monotonically converging to Σ.

Defining $D(t) = \Sigma - \Sigma(t)$, we have shown that

$$D(t+1) = \bar{A}[D(t) - D(t)C'(R + C\Sigma(t)C')^{-1}CD(t)]\bar{A}' \quad (3.50)$$

in Lemma 3.1.

Since $|\lambda_i(\bar{A})| < 1$, we have

$$\begin{aligned} \Sigma(t+1) - \Sigma(t) &= D(t) - D(t+1) \\ &\geq \bar{A}D(t)\bar{A}' - D(t+1) \\ &= \bar{A}D(t)C'(R + C\Sigma(t)C')^{-1}CD(t)\bar{A}' \geq 0 \end{aligned} \quad (3.51)$$

□

Therefore, the second conclusion in the Increasing Theorem is a special case of the above General Increasing Property. In practice, this increasing property and the

Minimum Property together ensure that we can employ partial information about the initial state covariance instead of a zero initial covariance and so gain a faster convergence speed.

Now, we deal with the case of an over-estimated initial covariance matrix.

Theorem 3.6 (General Decreasing Property)

Subject to the assumptions

(1) (C,A) is detectable.

(2) There exists a solution, Σ*, for the ARE.*

(3) the initial state covariance matrix $\Sigma(0)$ *satisfies* $\Sigma(0) \geq \Sigma \geq 0$.

the sequence $\{\Sigma(t),\ t \in T^+\}$ *generated by the RDE with initial state covariance matrix* $\Sigma(0)$ *is monotone decreasing and converges to* Σ.

proof: Using the same proof procedure as we used in the General Increasing Property, and replacing $D(t)$ by $-D(t)$, we can obtain the following results,

$$\lim_{t\to\infty} \Sigma(t) = \Sigma$$

$$\Sigma(t+1) \leq \Sigma(t)$$

where $t \in T^+$. □

Now, we conclude that the sequence $\{\Sigma(t),\ t \in T^+\}$ generated by the RDE will converge to the steady state covariance matrix Σ which satisfies the ARE with either an under-estimated or an over-estimated initial state covariance if certain conditions are met. However, we do not know which one will converge faster when the distances of the under-estimated and over-estimated initial covariance matrices from the true initial covariance matrix are the same.

Lemma 3.2 *Subject to the assumptions*

(1) A, B are non-singular matrices

(2) $[B^{-1} + CA^{-1}C']$ *is non-singular matrix*

then we have the result,

$$[A+CBC']^{-1} = A^{-1} - A^{-1}C'[B^{-1}+CA^{-1}C']^{-1}CA^{-1}$$

Theorem 3.7 *Subject to the assumptions*

(1) (C, A) detectable.

(2) There is a unique solution of the ARE, Σ.

(3) Two initial state covariance matrices $\Sigma_1(0)$ and $\Sigma_2(0)$ satisfy

$$\Sigma_1(0) > \Sigma, \quad \Sigma_2(0) < \Sigma$$

and

$$\|\Sigma_1(0) - \Sigma\| = \|\Sigma_2(0) - \Sigma\|$$

then we have,

$$\|\Sigma_1(t) - \Sigma\| \leq \|\Sigma_2(t) - \Sigma\|$$

proof: We define

$$D_1(t) = \Sigma_1(0) - \Sigma, \quad D_2(t) = \Sigma - \Sigma_2(0)$$

for $t \in T^+$ with initial $D_1(0) = D_2(0) > 0$.

By Lemma 3.1, we have

$$D_i(t+1) = \bar{A}[D_i(t) - D_i(t)C'(R+C\Sigma_i(t)C')^{-1}CD_i'(t)]\bar{A}' \tag{3.52}$$

where $\bar{A} = A - KC$, $K = [A\Sigma C' + S][C\Sigma C' + R]^{-1}$. The sequence $\Sigma_i(t)$ is generated by the RDE with initial value, $\Sigma_i(0)$, $i = 1, 2$.

By noting Lemma 3.2 and the results,

$$\Sigma_1(t) = \Sigma + D_1(t) \tag{3.53}$$

$$\Sigma_2(t) = \Sigma - D_2(t) \tag{3.54}$$

we have the recursive expressions,

$$D_1(t+1) = \bar{A}[D_1^{-1}(t) + C'(R+C\Sigma C')^{-1}C]^{-1}\bar{A}' \tag{3.55}$$

$$D_2(t+1) = \bar{A}[D_2^{-1}(t) - C'(R+C\Sigma C')^{-1}C]^{-1}\bar{A}' \tag{3.56}$$

for all $t \in T^+$.

We can also conclude that

$$D_1(1) \leq D_2(1)$$

since

$$\begin{array}{rccc} 0 < & D_1(0) & = & D_2(0) \\ \Rightarrow & D_1^{-1}(0) & = & D_2^{-1}(0) \\ \Rightarrow & D_1^{-1}(0) + C'(R + C'\Sigma C')^{-1}C & \geq & D_2^{-1}(0) - C'(R + C\Sigma C')^{-1}C \\ \Rightarrow & D_1^{-1}(1) & \geq & D_2^{-1}(1) \end{array}$$

Suppose $D_1(t) \leq D_2(t)$, we then prove that $D_1(t+1) \leq D_2(t+1)$ and define

$$D_i^\epsilon(t) = D_i(t) + \epsilon I, \quad i = 1, 2$$

If $D_1(t)$ is singular, there exists an $\epsilon > 0$ to make both $D_1^\epsilon(t)$ and $D_2^\epsilon(t)$ positive definite.

By noting that we have

$$\begin{array}{rccc} 0 < & D_1^\epsilon(t) & \leq & D_2^\epsilon(t) \\ \Rightarrow & [D_1^\epsilon(t)]^{-1} & \geq & [D_2^\epsilon(t)]^{-1} \\ \Rightarrow & [D_1^\epsilon(t)]^{-1} + C'(R + C\Sigma C')^{-1}C & \geq & [D_2^\epsilon(t)]^{-1} - C'(R + C\Sigma C')^{-1}C \\ \Rightarrow & [D_1^\epsilon(t+1)]^{-1} & \geq & [D_2^\epsilon(t+1)]^{-1} \\ \Rightarrow & D_1^\epsilon(t+1) & \leq & D_2^\epsilon(t+1) \end{array}$$

and letting $\epsilon \to 0$ yields as a consequence the relation,

$$D_1(t+1) \leq D_2(t+1) \tag{3.57}$$

If $D_1(t)$ is non-singular, the above inequality still holds if we set $\epsilon = 0$.

Therefore, for any $t \in T^+$ we know that

$$\begin{array}{rcc} & D_1(t) & \leq & D_2(t) \\ \Leftrightarrow & \|\Sigma_1(t) - \Sigma\| & \leq & \|\Sigma_2(t) - \Sigma\| \end{array}$$

□

An example is presented in Appendix A. Although the initial state can be forgotten quickly by a fast state convergence speed, the initial conditions still affect the performance of Kalman process in a small data set. In the following section, we show that the fixed point smoothing algorithm provides an efficient way to estimate the initial conditions and to analyse the properties of the algorithm in detail.

3.4 Estimation of Initial Conditions

The unconditional distribution of the state vector is not defined when the transition equation is non-stationary, unless genuine prior information is available. In general, the initial condition covariance $\Sigma(0)$ is the parametric quantity required to specify the joint distributions of $(x(0), y^{(T)})$, and to initiate the recursion for the sequence $\{\Sigma(t)\}$ in the Kalman filter. However, this quantity cannot be consistently estimated if it cannot be concentrated out of likelihood function i.e. it is not completely "free" from the parameter set Θ. There are three different approaches to estimate the initial state conditions in the existing literature, namely

1. The Diffuse Approach
2. The General Least Squares Estimation Approach
3. The Fixed Initial Vector Approach

See details of the above approaches in Harvey and Phillips (1979), Harvey (1989), and Rosenberg (1973). In this section, we shall show that fixed point smoothing, which is a Bayesian estimation, can be used to estimate the initial conditions of the state vector.

3.4.1 Fixed Point Smoothing and the Estimation of Initial State Condition

For a given initial state vector $x(0)$ and its covariance matrix $\Sigma(0)$, each state $x(t)$ is a random vector that is jointly distributed with all future system states and outputs,

as in probability spaces $\mathbf{P}(\cdots, x(t-2), x(t-1))$ and $\mathbf{P}(\cdots, y(t-1), y(t))$. The result of the filtering calculation is that the conditional density of $x(t)$ given the output observation $y^{(t-1)}$ and system parameters Θ, is Gaussian with mean $x(t|t-1)$ and covariance matrix $\Sigma(t)$, where these are generated by the Kalman filtering equations and the RDE.

The Kalman filtering problem is most closely related to Bayesian parameter estimation in the case of a fixed-point smoothing problem. This problem constitutes an important example of Gaussian conjugate densities in Bayesian estimation.

In the smoothing problem we wish to estimate a random state at a fixed time using observations on the output process $y(t)$ which are observed on and after the fixed time. We can, therefore, view estimation of the initial state problem as a smoothing problem.

By (ii) of Assumption 3.1, we have

$$\mathbf{E}(x(t)|y^{(t-1)}) = \mathbf{E}(Ax(t-1) + \xi(t)|y^{(t-1)}) = x(t|t-1) \tag{3.58}$$

To obtain a recursion for $\{x(t+1|t)$ and $x(0|t)$, $t \in T^+\}$, we draw upon the following relations,

$$\begin{pmatrix} x(t+1|t) \\ x(0|t) \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} x(t|t-1) \\ x(0|t-1) \end{pmatrix} + \begin{pmatrix} K(t) \\ K^*(t) \end{pmatrix} [y(t) - Cx(t|t-1)] \tag{3.59}$$

$$\begin{pmatrix} K(t) \\ K^*(t) \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) \\ \Sigma_{21}(t) & \Sigma_{22}(t) \end{pmatrix} \begin{pmatrix} C' \\ 0 \end{pmatrix} [C\Sigma_{11}C' + R]^{-1} \tag{3.60}$$

$$\begin{aligned} \begin{pmatrix} \Sigma_{11}(t+1) & \Sigma_{12}(t+1) \\ \Sigma_{21}(t+1) & \Sigma_{22}(t+1) \end{pmatrix} &= \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) \\ \Sigma_{21}(t) & \Sigma_{22}(t) \end{pmatrix} \begin{pmatrix} A' & 0 \\ 0 & I \end{pmatrix} \\ &- \begin{pmatrix} K(t) \\ K^*(t) \end{pmatrix} [C\Sigma_{11}C' + R] \begin{pmatrix} K(t) \\ K^*(t) \end{pmatrix} \\ &+ \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \end{aligned} \tag{3.61}$$

In view of the initial condition, it follows that the starting values for equation (3.61) are

$$\begin{pmatrix} \Sigma_{11}(t_0) & \Sigma_{12}(t_0) \\ \Sigma_{21}(t_0) & \Sigma_{22}(t_0) \end{pmatrix} = \begin{pmatrix} \Sigma_{11}(t_0) & \Sigma_{11}(t_0) \\ \Sigma_{11}(t_0) & \Sigma_{11}(t_0) \end{pmatrix}$$

Under (i) and (ii) of Assumption 3.1, a joint conditional Gaussian distribution exists for $(x(1), x(0), y(1))'$ when the conditioning is carried out with respect to $y^{(0)}$. To be specific

$$\begin{aligned} \begin{pmatrix} x(1) \\ x(0) \\ y(1) \end{pmatrix} &\sim \mathbf{P}\left(\begin{pmatrix} x(1) \\ x(0) \\ y(1) \end{pmatrix} \middle| y^{(0)} \right) \\ &= \mathbf{N}_{y^{(0)}}\left(\begin{pmatrix} x(1|0) \\ \bar{x}(0) \\ Cx(1|0) \end{pmatrix}, \begin{pmatrix} \Sigma_{11}(1) & \Sigma_{12}(1) & \Sigma_{11}(1)C' \\ \Sigma_{21}(1) & \Sigma_{22}(0) & \Sigma_{12}(1)C' \\ C\Sigma_{11}(1) & C\Sigma_{21}(1) & C\Sigma_{11}(1)C' + R \end{pmatrix} \right) \end{aligned}$$

where

$$\begin{aligned} \Sigma_{11}(1) &= \mathbf{E}(x(1) - x(1|0))(x(1) - x(1|0))' \\ &= A\Sigma_{11}(0)A' + Q - K(0)[C\Sigma_{11}(0)C' + R]^{-1}K'(0) \end{aligned}$$

$$\begin{aligned} \Sigma_{21}(1) &= \mathbf{E}(x(0) - \bar{x}(0))(x(1) - x(1|0))' \\ &= \Sigma_{21}(0)[A - K^*(0)C]' \\ \Sigma_{12}(1) &= \Sigma_{21}(1) \end{aligned}$$

$$\begin{aligned} K(0) &= A\Sigma_{11}(0)C'[C\Sigma_{11}(0)C' + R]^{-1} \\ K^*(0) &= \Sigma_{21}(0)C'[C\Sigma_{11}(0)C' + R]^{-1} \end{aligned}$$

Through an induction argument, we obtain for a general instant $t \in T^+$

$$\begin{pmatrix} x(t+1) \\ x(0) \\ y(t+1) \end{pmatrix} \sim \mathbf{P}\left(\begin{pmatrix} x(t+1) \\ x(0) \\ y(t+1) \end{pmatrix} \middle| y^{(t)} \right)$$

$$= \mathbf{N}_{y(t)}\left(\begin{pmatrix} x(t+1|t) \\ x(0|t) \\ Cx(t+1|t) \end{pmatrix}, \Sigma(t+1)\right) \tag{3.62}$$

where

$$\Sigma(t+1) = \begin{pmatrix} \Sigma_{11}(t+1) & \Sigma_{12}(t+1) & \Sigma_{11}(t+1)C' \\ \Sigma_{21}(t+1) & \Sigma_{22}(t) & \Sigma_{12}(t+1)C' \\ C\Sigma_{11}(t+1) & C\Sigma_{21}(t+1) & C\Sigma_{11}(t+1)C'+R \end{pmatrix} \tag{3.63}$$

with

$$x(t+1|t) = Ax(t|t-1) + K(t)(y(t) - Cx(t|t-1)) \tag{3.64}$$

$$x(0|t) = x(0|t-1) + K^*(t)(y(t) - Cx(t|t-1)) \tag{3.65}$$

$$\begin{aligned} \Sigma_{11}(t+1) &= \mathbf{E}(x(t+1) - x(t+1|t))(x(t+1) - x(t+1|t))' \\ &= A\Sigma_{11}(t)A' + Q - K(t)[C\Sigma_{11}(t)C' + R]^{-1}K'(t) \end{aligned} \tag{3.66}$$

$$\begin{aligned} \Sigma_{21}(t+1) &= \mathbf{E}(x(0) - x(0|t))(x(t+1) - x(t+1|t))' \\ &= \Sigma_{21}(t)[A - K^*(t)C]' \end{aligned} \tag{3.67}$$

$$\Sigma_{12}(t+1) = \Sigma_{21}(t+1) \tag{3.68}$$

$$\begin{aligned} \Sigma_{22}(t) &= \mathbf{E}(x(0) - x(0|t))(x(0) - x(0|t))' \\ &= \Sigma_{22}(t-1) - \Sigma_{21}(t-1)C'K^*(t) \\ &= \Sigma_{21}(t-1)C'[C\Sigma_{11}(t)C' + R]^{-1}C\Sigma_{21}'(t-1) \end{aligned} \tag{3.69}$$

$$K(t) = A\Sigma_{11}(t)C'[C\Sigma_{11}(t)C' + R]^{-1} \tag{3.70}$$

$$K^*(t) = \Sigma_{21}(t)C'[C\Sigma_{11}(t)C' + R]^{-1} \tag{3.71}$$

To process the argumented Kalman filter requires computer space and computer time raised to the power 2 in comparison with the original Kalman filter. By noting, however, the update equations from (3.64) to (3.71), the need to run fully the argumented filter can be avoided by recursion.

On the other hand, we can view the fixed-point smoothing procedure for estimation of the initial state vector as its linear representation on $\mathbf{H}_t^y \ominus \mathbf{H}_{-1}^y$ where $\mathbf{H}_t^y = Sp\{\cdots, y(-1), y(0), \cdots, y(t)\}$ is a Hilbert space. The original Kalman filter model can be view as an innovation generator,

$$\begin{cases} x(t+1|t) = Ax(t|t-1) + K(t)v(t) \\ y(t) = Cx(t|t-1) + v(t) \end{cases} \tag{3.72}$$

where $v(t) = y(t) - Cx(t|t-1)$.

The sequence $\{v(t), t \in T^+\}$ generated by the Kalman filter is an orthogonal process. Hence,

$$\begin{aligned} \mathbf{H}_t^y \ominus \mathbf{H}_{-1}^y &= Sp\{v(0), v(1), \cdots, v(t)\} \\ &= \sum_{i=0}^{t} \oplus \mathbf{H}_i^v \end{aligned}$$

where $\mathbf{H}_i^v = Sp\{v(i)\}$.

The fixed point smoothing for estimation of the initial vector $x(0)$ conditional on $y^{(t)}$ is given by

$$\begin{aligned} x(0|t) &= x(0)|y^{(t)} \\ &= x(0)|\mathbf{H}_t^y \ominus \mathbf{H}_{-1}^y \\ &= x(0|-1) + \sum_{i=1}^{t} (x(0)|\mathbf{H}_i^v) \\ &= x(0|-1) + \sum_{i=1}^{t} \mathbf{E}(x(0)v'(i))(R_i^v)^{-1}v(i) \\ &= x(0|-1) + \sum_{i=1}^{t-1} \mathbf{E}(x(0)v'(i))(R_i^v)^{-1}v(i) + \mathbf{E}(x(0)v'(t))(R_i^v)^{-1}v(t) \\ &= x(0|t-1) + \mathbf{E}[(x(0) - x(0|-1)) + x(0|-1)] \\ &\quad \times [\epsilon'(t) + (x(t) - x(t|t-1))'C'][R_t^v]^{-1}v(t) \\ &= x(0|t-1) \\ &\quad + \mathbf{E}(x(0) - x(0|-1))(x(t) - x(t|t-1))'C'][R_t^v]^{-1}v(t) \end{aligned} \tag{3.73}$$

since $x(0) \perp \epsilon(t)$ and $x(0|-1) \perp \epsilon(t)$.

We define two new random vectors, $\tilde{x}_0(t)$ and $\tilde{x}(t)$, as follows,

$$\begin{aligned}\tilde{x}_0(t) &= x(0) - x(0|t)\\ \tilde{x}(t) &= x(t) - x(t|t-1)\end{aligned}$$

From equation (3.73), we get

$$\tilde{x}_0(t) = \tilde{x}_0(t-1) - [\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(t)]C'[R_t^v]^{-1}[\epsilon'(t) + C\tilde{x}(t)]' \tag{3.74}$$

and

$$\begin{aligned}\mathbf{E}\tilde{x}_0(t)\tilde{x}_0'(t) &= \mathbf{E}\tilde{x}_0(t-1)\tilde{x}_0'(t-1)\\ &\quad -2[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(t)]C'[R_t^v]^{-1}[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(t)]'\\ &\quad [\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(t)]C'[R_t^v]^{-1}[C\Sigma(t)C' + R][R_t^v]^{-1}[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(t)]'\\ &= \mathbf{E}\tilde{x}_0(t-1)\tilde{x}_0'(t-1)\\ &\quad -[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(t)]C'[R_t^v]^{-1}[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(t)]' \qquad (3.75)\end{aligned}$$

since $\tilde{x}_0(t-1) \perp \epsilon(t)$ and $\tilde{x}(t) \perp \epsilon(t)$.

Hence we have the expression,

$$\mathbf{E}\tilde{x}_0(t)\tilde{x}_0'(t) = \mathbf{E}\tilde{x}_0(-1)\tilde{x}_0'(-1) - \sum_{i=1}^{t}[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(i)][R_i^v]^{-1}[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(i)]' \tag{3.76}$$

The term $[\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(i)][R_i^v]^{-1}\ [\mathbf{E}(\tilde{x}_0(-1)\tilde{x}'(i)]'$ is the estimation improvement for the initial state when $y(t)$ is available, and is non-negative definite. We have, therefore, the following limitation theorem.

Theorem 3.8 *For any state space model which satisfies (i) (ii) of Assumptions 3.1, the smoothing estimation for the initial vector is a Baysian estimation, and there exists a non-negative matrix,* $\Sigma_0(x(0|-1))$ *for* $x(0|-1)$

proof:

Since the sequence $\{\mathbf{E}\tilde{x}_0(t)\tilde{x}_0'(t)\}$ is bounded (≥ 0) and a monotonically decreasing sequence, there must exist a non-negative matrix $\Sigma_0(x(0|-1))$ which satisfies the expression,

$$\lim_{t\to\infty} \mathbf{E}\tilde{x}_0(t)\tilde{x}_0'(t) = \Sigma_0(x(0|-1))$$

□

This theorem does not imply that the smoothed estimation is consistent. However, the smoothed estimation does improve the initial state vector although the estimation relies on the performance of the original Kalman filtering from the recursive procedure set out in equations (3.64) to (3.71) which depend on the system matrices A, C and the disturbance process covariance matrices Q, R. An obvious weakness of the smoothed initial estimation algorithm is that Q and R are required at each step. This is an unreasonable amount of information to be required a priori. Indeed, one of the desirable products of many system identification algorithms is precisely the system and observation disturbance variances. In fact, all optimal linear or non-linear filtering solutions to parameter estimation problems require a priori information on the noise covariance matrix data, at least in their initial formulation.

3.5 The Influence of Error Disturbance Covariance

In applying the state space model to a specific system, The system matrices A, C, noise covariance matrices Q, R and a priori initial condition $(x(0), \Sigma(0))$ must be specified. Since the state space model is usually an approximation to the system to be modeled, the system matrices and noise characteristics are seldom known exactly although the model can be parameterized by a parameter set $\Theta = (A, C, Q, R)$ which can be estimated by a number of algorithms, such as Maximum Likelihood, Minimum Variance, Minimum Prediction Variance, etc.

We have already remarked in the last section that the initial conditions are seldom precisely known a priori. Under the condition that the system matrices and disturbance characteristics are known exactly, i.e. Θ is known, any arbitrarily specified initial conditions which starts the filtering process are eventually forgotten when sufficient observations have been processed. The state vector converges to a steady state when certain conditions for the state space model are met (see Theorems 3.3

and 3.4). However, the convergence speed can be accelerated when partial information on the initial conditions are used to initiate the filtering process (see Theorems 3.6 and 3.7). Now, we should ask what is the effect on the filtering performance if the parameter set is not known exactly or an approximation is made. In this section, we concentrate on the error sensitivity to the disturbance statistics. We, therefore, suppose A, C are known exactly, and Q, R are different from the true values. Then we have as the model

$$\begin{cases} x_e(t+1) = Ax_e(t) + \xi_e(t) \\ y(t) = Cx_e(t) + \epsilon_e(t) \end{cases} \tag{3.77}$$

where $\mathbf{E}\xi_e(t) = 0$, $\mathbf{E}\xi_e(s)\xi_e'(t) = Q_e\delta(t-s)$ and $\mathbf{E}\epsilon_e(t) = 0$, $\mathbf{E}\epsilon_e(s)\epsilon_e(t) = R_e\delta(t-s)$.

The innovation representation of the above model is

$$\begin{cases} x_e(t+1|t) = Ax_e(t|t-1) + K_e(t)v_e(t) \\ y(t) = Cx_e(t|t-1) + v_e(t) \end{cases} \tag{3.78}$$

where $v_e(t) = y(t) - Cx_e(t|t-1)$ are the innovations generated by the model (3.78). and $K_e(t)$, defined by, $K_e(t) = [A\Sigma_e(t)C' + S_e][C\Sigma_eC' + R]^{-1}$ is the Kalman gain, and with $\Sigma_e(t+1)$ derived from,

$$\begin{aligned} \Sigma_e(t+1) &= [A - K_e(t)C]\Sigma_e(t)[A - K_e(t)C]' \\ &\quad + [K_e(t) - S_eR_e^{-1}]R_e[K_e(t) - S_eR_e^{-1}]' + Q_e - S_eR_e^{-1}S_e' \end{aligned} \tag{3.79}$$

with the initial value, $\Sigma_e(0)$.

However, the computed $\Sigma_e(t)$ is $\mathbf{E}(x_e(t) - x_e(t|t-1))(x_e(t) - x_e(t|t-1))'$ instead of the estimation error covariance $\Sigma^{(e)}(t) = \mathbf{E}(x(t) - x_e(t|t-1))(x(t) - x_e(t|t-1))'$ since the model (3.77) is different from the real one. Now, we examine the relation between $\Sigma_e(t)$ and $\Sigma^{(e)}(t)$.

Defining

$$\tilde{x}_e(t+1) = x(t+1) - x_e(t+1|t)$$

then we have

$$\tilde{x}_e(t+1) = A\tilde{x}_e(t) + \xi(t) - K_e(t)(C(\tilde{x}_e(t) + \epsilon(t)) \tag{3.80}$$

and

$$\begin{aligned}\Sigma^{(e)}(t+1) &= \mathbf{E}\tilde{x}_e(t+1)\tilde{x}_e'(t+1) \\ &= [A-K_e(t)C]\Sigma^{(e)}(t)[A-K_e(t)C]' \\ &\quad +[K_e(t)-SR^{-1}]R[K_e(t)-SR^{-1}]'+Q-SR^{-1}S' \qquad (3.81)\end{aligned}$$

by noting that in the model (3.77) using the real $y(t)$, $\tilde{x}_e(t) \perp y(t)$, $\tilde{x}_e(t) \perp x_e(t|t-1)$, $\tilde{x}_e(t) \perp \xi(t)$ and $\tilde{x}_e(t) \perp \epsilon(t)$. Then we have the expression,

$$\begin{aligned}\Sigma_e(t+1)-\Sigma^{(e)}(t+1) &= [A-K_e(t)C][\Sigma_e(t)-\Sigma^{(e)}(t)][A-K_e(t)C]' \\ &\quad -[(S_e-S)K_e'(t)+K_e(t)(S_e-S)'] \\ &\quad +K_e(t)[R_e-R]K_e'(t)+[Q_e-Q] \qquad (3.82)\end{aligned}$$

Suppose the state and measurement disturbances $\xi(t)$ and $\epsilon(t)$ are uncorrelated, i.e. $S = 0$ and $S_e = 0$, and $Q_e \geq Q$, $R_e \geq R$, then $\Sigma_e(t+1) \geq \Sigma^{(e)}(t+1)$ if $\Sigma_e(t) \geq \Sigma^{(e)}(t)$. Therefore, by induction, we obtain

Theorem 3.9 *If the disturbances of a state space system are uncorrelated, and* $\Sigma_e(0) \geq \Sigma^{(e)}(0)$, $Q_e \geq Q$, $R_e \geq R$, *then* $\Sigma_e(t) \geq \Sigma^{(e)}(t)$ *for any* $t \in T^+$.

By comparing the RDE of the model (3.77) and the RDE of the real model (3.1) (see equation (3.81) and (3.18)), it can be see that the difference between them is the Kalman gain. Using the Minimum Property (Theorem 3.1), we can conclude that $\Sigma^{(e)}(t) \geq \Sigma(t)$ if $\Sigma^{(e)}(0) \geq \Sigma(0)$. Combining with Theorem 3.9, we, finally, have

Corollary 2 *If the disturbance of a state space model is uncorrelated and*

$$\Sigma_e(0) \geq \Sigma(0),\ Q_e \geq Q,\ R_e \geq R,$$

then we have

$$\Sigma_e(t) \geq \Sigma^{(e)}(t) \geq \Sigma(t)$$

The theorem shows that conservative estimates of $\Sigma(0)$, Q, R can often be made for a state space model in which the state and measurement disturbances are unrelated, although the actual values of those quantities are not generally available. In general, however, conservative estimates of $\Sigma(0)$, Q, R may not yield a conservative state covariance matrix unless the correlations between the disturbances is known exactly. This requirement is clearly unreasonable since Q, R are usually unknown initially.

3.6 A Model Estimation Procedure

If the model is detectable and stabilizable, the state vector will converge to a positive semi-definite matrix no matter what the distribution of the initial vector. In other words, the initial state covariance matrix would not affect the estimate of the state vector asymptotically. Only the values of the initial vector affect the estimate of the state vector.

Therefore, it is simplest to regard the initial state vector as part of the model parameters. It can be included in the unknown parameter set Θ as unknown parameters of the model and estimated by a maximum likelihood procedure. However, treating the initial state vector $x(0)$ in this way will increase the complexity of the numerical optimization considerably when the dimension of the state vector is "large" and many unknown model parameter are also to be estimated. A more practical way is to make the initial state vector estimate conditional on the model parameters. In other words, the initial state vector $x(0)$ is not estimated in the maximum likelihood procedure which estimates the model parameters only. In the last section, we have shown that the initial state vector and the corresponding covariance matrix can be estimated by fixed point smoothing with Kalman filtering. Therefore, a recursive model estimation procedure can be constructed as follows:

Step 1 With initial model parameter set Θ, and estimated initial state vector $\hat{x}(0)$ and the over-estimated variance matrix $\hat{\Sigma}(0)$, some partial information on the

initial state vector can be provided. Otherwise, $\hat{x}(0) = 0$ with a diffuse initial distribution for $\hat{x}(0)$ can be used.

Step 2 The first part of the exact likelihood function of the state space model (3.1) is (see equation (3.13))

$$-\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}[\log|\Sigma_v(t)| + v'(t)\Sigma_v^{-1}(t)v(t)] \tag{3.83}$$

where $v(t)$ is generated by the innovation representation (see equation (3.14)) of the state space model 3.1. Thus, the Kalman filter is employed to estimate the unknown model parameter set Θ. The initial state vector $x(0)$ and its covariance matrix $\Sigma(0)$ are estimated by the fixed point smoothing procedure described in §3.4.1.

Step 3 After the model parameter set $\hat{\Theta}$ is estimated, the old initial state vector $\hat{x}(0)$ and its variance $\hat{\Sigma}(0)$ are updated by smoothed estimation. If the estimated model parameter set $\hat{\Theta}$ differs considerably from the last estimation, then $\hat{\Theta}$ serves as the updated estimate of Θ and return to **Step 2**; if the estimated model parameter set $\hat{\Theta}$ is not much different from the last estimation, the recursive procedure ceases.

The estimated $\hat{\Theta}$ is the maximum likelihood estimation of Θ under the smoothed initial state conditions.

The application of this recursive model estimation procedure will be presented in chapter 7 where the recursive model estimation procedure is shown to be superior to the simple maximum likelihood estimation because the more precise initial vector estimate accelerates the convergence of the parameter estimates and produces more accuracy in the estimates.

Chapter 4

Profiles of Electricity Load

4.1 Introduction

This chapter is a part of the study of the New Zealand half-hourly electricity consumption data from April 1st, 1983 to March 31st, 1984. The level shift and other changing characteristics studied in this chapter arose when we studied a forecasting procedure for half-hourly electricity consumption developed by Moutter et al. (1986a) and Moutter et al. (1986b) and Bodger et al. (1987). Their procedures using frequency domain models are an important attempt to deal with periodical time series data. Unfortunately, the procedures may not provide good forecasts because of some assumptions used to simplify their model, which lead to a restriction on the Papoulis algorithm (Papoulis (1975)). Their procedures were used to forecast weekday data, under the assumption that weekend days' load is similar in daily profile to weekdays' but with a lower level. Our study shows that weekdays' and weekend days' load produce very different profiles as well as different levels.

We assume that the load comprises trend, periodic, and innovation components in an additive form. For the integrity of the modelling procedure, the models for the unobserved trend component and the periodic component are developed in this chapter. The stationary innovation component will be modelled by a subset AR modelling procedure developed in chapter 2. The trend component is described by an "error correction" cointegrating regression model; the periodic component described

by a periodically stationary model and the weekly periodic cycle is handled by the evolution between weekdays and weekend days which is captured by a *transition* function. Applications of the models developed in this chapter, and comparison with other models will be presented in Chapter 5.

4.1.1 Data

The New Zealand total electricity demand data recorded at 30-minute intervals from 1st April, 1983 to 31st March, 1984 (17,520 readings) is made available by Bodger et al. (1987) to enable the methods proposed to be tested on this data. The pattern of the data shows strong daily and weekly sinusoidical characteristics. A segment of two weeks of the data from Monday 11th June 1983 to Monday 24th June has been selected randomly and shows the pattern given in Figure 4.1. With very slowly changing working arrangements and human behaviour, an assumption can be made that the load pattern is similar week to week except for the effects of weather, holidays, and special events which have great influence upon electricity demands. From Figure 4.1, it can also be seen that the load curve for weekdays (Monday - Friday) is similar with almost identical characteristics. The pattern for weekend days is similar for Saturday and Sunday but is different from the weekday profile.

4.1.2 Moutter's Model

A simple model was developed by Moutter et al. (1986b) for the short-term New Zealand electricity consumption data. Because the load pattern for each new day follows an "almost" identical pattern to those preceding it, Moutter et al. (1986b) make an assumption that all spectral components must be harmonically related to the fundamental frequency component which corresponds to the daily cycle and phases of the spectral components are "identical" at the beginning of each day. Based on the above assumption, Moutter's model is established using harmonic frequency analysis. The fast Fourier transform (FFT) technique is employed in this procedure to determine the fundamental frequency component and its significant harmonic frequencies.

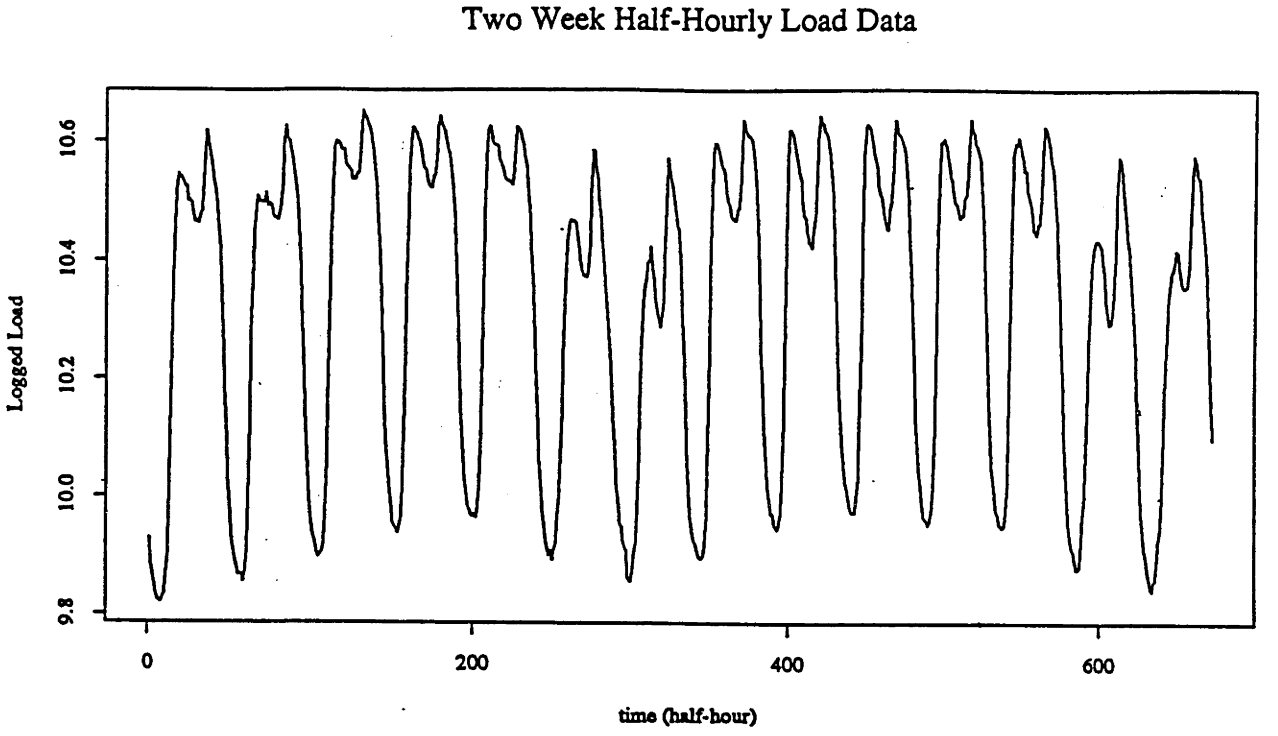


Figure 4.1: New Zealand Half-hourly Electricity Demands from 11/07/83 to 24/07/83

An empirical preparatory stage referred to as Moutter's *Super-resolution Algorithm* is used to find the most relevant frequencies and to increase the convergence speed of the Papoulis algorithm (Papoulis and Chamzas (1979)). A leakage re-introduction technique was used empirically to reduce errors which occur in spectral identification. Finally, the original Papoulis algorithm is applied to the selected spectral components to allow extrapolation from the sample data. Bodger et al. (1987) use the same idea with a mixed radix fast Fourier transform to avoid the FFT array size restriction[1] and then reduce spectral leakage. For the different load pattern characteristics of weekdays and weekend days, Moutter et al. (1986b) and Bodger et al. (1987) believe that the main difference between them is only a level shift, and neglected any other differences. For this reason, the forecasting may be much less satisfactory when sample data ends with the last of the weekdays, Friday or the data ends with Sunday.

[1] A FFT calculates the frequency spectrum of a discrete data set at frequency $\pi t/2^m$ where $2^m \geq N$, m is an integer and N is the sample size.

4.2 Model Building

From the following rough analysis[2] we find that the data shown in Figure 4.1 has three main characteristics:

1. The data is strongly oscillatory with a daily period (48 points) and a weekly period (336 points), which reflect the daily electricity consumption behavior, and working and resting day patterns, respectively. In addition, the daily electricity consumption characteristics of weekdays and weekend days are different, i.e. the profiles of the load in weekdays and in weekend day's are different, though the weekdays' profiles appear "almost" identical one to another. If the profiles of the weekday's load and the weekend days' load are similar in shape, then it is expected that the magnitude ratios of each harmonic component to the fundamental frequency component, and the phases of the corresponding frequencies would be close in values for weekdays and weekend days. Table 4.1 shows the magnitude ratios of the five harmonic components to the daily fundamental frequency component and the phases of the harmonic components for weekdays' and weekend days' from the data obtained by removing their corresponding daily sample means. The obvious differences indicate that the load profiles of weekdays' and weekend days' are not similar in shape.

2. The daily mean level electricity consumption of weekdays is higher than that of the following weekend days. To illustrate this, Table 4.2 shows the sample daily means from Monday 11th June 1983 to Sunday 24th June 1983 and the t-test statistics. The t statistics show that there are no significant difference between weekdays' level and weekend days level. However, the levels of weekdays' and weekend days' are significantly different. The above characteristics of weekdays' and weekend days' data suggest that harmonic analysis in the frequency domain

[2]The main differences between the weekdays and weekend days lie on the daily profiles and daily levels of the load. The comprehensive tests are very complicated. The differences shown here are based on intuition because our interest is of modelling the load instead of testing the explicit differences between weekdays' and weekends' load.

Order		0	1st	2ed	3rd	4th	5th
Frequency		0.1309	0.2618	0.3927	0.5236	0.6545	0.7854
Period(Hour)		24	12	8	6	4.8	4
R(i,j)[a]	Weekdays	1.0	0.6115	0.1295	0.1134	0.1333	-
	Weekends	1.0	0.5656	0.2432	-	0.1018	0.0388
Phase(Deg)	Weekdays	-139.48	157.02	52.86	147.45	-131.63	-
	Weekends	-125.54	-87.52	149.95	-	-156.13	113.93

Table 4.1: Comparison of the Five Largest Daily Harmonic Spectral Components of Two Weeks Data

[a] $R(i,j) = C(i,j)/C(i,0)$ $i = 1,2$, $j = 0,1,2,...5$ where $C(1,j)$ = the magnitude of the jth harmonic component for the weekdays data. $C(2,j)$ = the magnitude of the jth harmonic component for weekend days.

$A(i,j)$[a]	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Mon.	10.16	(-3.41)					
Tue.		10.23	(0.67)				
Wed.			10.27	(0.22)			
Thu.				10.27	(0.38)		
Fri.					10.24	(3.13)	
Sat.						10.14	(1.48)
Sun.	(2.51)						10.07

Table 4.2: Comparison of the Daily Sample Means Over Two Weeks

[a] $A(i,i)$ are sample means of ith day of a week. $A(i,i+1)$ are t statistics between ith and $i+1$th days. Under hypotheses: $\mathbf{H_0}$: Daily Mean of ith day = Daily Mean of $i+1$th day where $i = 1,2,\cdots,7$.

may be more effective if the weekday' and the weekend days' data are treated separately[3].

3. It can be seen that the Fridays' and the Mondays' daily sample mean consumption levels are between their adjacent daily sample mean consumption levels. From this point of view, Fridays and Mondays can be regarded as transition periods from weekdays' characteristics to weekend days' or in reverse.

[3] In general, to treat weekdays and weekend days separately would lose some information on the relation between weekdays' load and weekend days'. However, it can overcome the difficulty that there are too many frequency components, which are harmonics of the weekly frequency to deal with. The separate treatment will now focus on the daily frequency as the fundamental frequency because neither weekend days nor the weekdays produce strong two-day and five-day periodical patterns.

To capture all the specific characteristics, we propose that a one week period can be divided into 4 intervals:

$$\begin{aligned} T_m &= \{t|\text{Monday}\} \\ T_{wd} &= \{t|\text{Tuesday} \sim \text{Thursday}\} \\ T_f &= \{t|\text{Friday}\} \\ T_{we} &= \{t|\text{Saturday} \sim \text{Sunday}\} \end{aligned}$$

and define the following variables,

- $X_1(t)$ is weekdays' electricity consumption, where $t \in T_{wd}$.
- $X_2(t)$ is weekend days' electricity consumption, where $t \in T_{we}$.

A model can then be established as follows:

$$X(t) = \begin{cases} (1 - f_1(t))X_1(t_{wd}) + f_1(t)X_2(t_{we}) + \epsilon_m(t) & \text{if } t \in T_m \\ X_1(t) & \text{if } t \in T_{wd} \\ f_2(t)X_1(t_{wd}) + (1 - f_2(t))X_2(t_{we}) + \epsilon_f(t) & \text{if } t \in T_f \\ X_2(t) & \text{if } t \in T_{we} \end{cases} \tag{4.1}$$

where $\text{mod}(t, 48) = \text{mod}(t_{wd}, 48) = \text{mod}(t_{we}, 48)$, $f_1(t)$ and $f_2(t)$ are positive function ranging from 0 to 1. They allow us to describe the transition characteristics of the electricity demands for Mondays and Fridays, respectively. The $\epsilon_m(t)$ and $\epsilon_f(t)$ are disturbance terms when $t \in T_m$ and T_f, respectively.

According to the above analysis, the weight functions $f_1(t)$ and $f_2(t)$ play a crucial role in the model (4.1). The estimation of the weight functions and the overall forecasting of electricity demands will be discussed in later sections.

Firstly, we specify that the half-hourly load, $X(t)$, consists of three components, namely, trend (X_{trend}), periodic component (X_{period}), and an innovation term ($X_{innovation}$) in an additive form, i.e.

$$X(t) = X_{trend}(t) + X_{periodic}(t) + X_{innovation}(t) \tag{4.2}$$

Based on the linking model (4.1) and the additive model (4.2) for the weekdays and weekend days load, a modelling and forecasting scheme is established using the following steps;

1. Model and Estimate the Trend and then Detrend
2. Model and Estimate the Periodic Pattern for Weekdays and Weekend Days Separately
3. Model and Estimate the Transition between Weekdays and Weekend Days
4. Model and Estimate the Innovation Series
5. Forecast the Overall Load

4.3 Trend and Detrend

The most important feature of the New Zealand electricity "short term" data is the dominant daily and weekly periodic pattern. The annual growth trend and annual seasonal pattern are not easily visible if only a few weeks data are observed. However, the annual growth trend and annual seasonal pattern can have a significant effect on forecasting a few days or a few weeks ahead. Figure 4.2 shows the New Zealand weekly electricity demand from 1972 to 1983. It is obvious that the weekly load increases substantially from summer to winter, decreases similarly from winter to summer, which forms an annual seasonal pattern. It is not necessary to employ sample data arising from several years of short term data recorded half-hourly, in which trend and annual seasonal patterns may be visible and can be estimated, when our major purpose at this point is forecasting only a few days ahead and illustrating our ability to capture the changing daily profile. Hence, we have established an "error correction" model for the trend component which develops the short-run trend for the half-hourly data from weekly data.

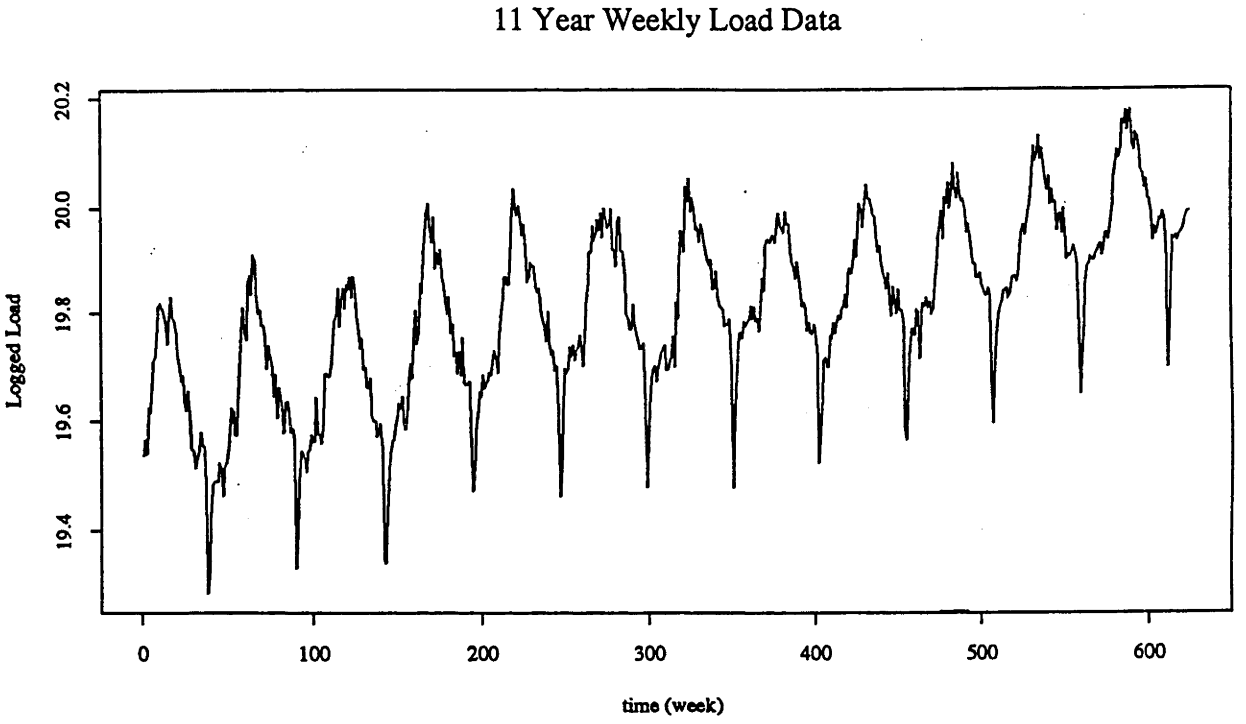


Figure 4.2: The New Zealand Weekly Electricity Demand from 1972 to 1983

In this section, an artificial trend for the half-hourly data is modelled with assistance from the weekly data and associated model characteristics. Therefore, this trend can be predicted and extracted from the raw data to form detrended data.

It is obvious that weekly data will have a marked annual seasonal pattern, i.e. the weekly data are periodical time series Pagano (1978) with period 52 (a year is approximately 52 weeks) after the natural growth trend is extracted. If we suppose the annual growth trend is exponential, it becomes linear after logarithmic transformation and it is expected that the weekly load, $W(\tau)$, is integrated at both 0 and annual seasonal frequencies, where τ is a time index on a weekly scale.

The idea of cointegration, its implications, test procedures and applications can be found in Granger (1986), Engle and Granger (1987), Engle et al. (1989). They show the one-to-one relationship between *error-correction mechanism* (ECM) and cointegration and present an easy method for estimating such relationships and testing for cointegration, now known as the *Granger and Engle Two-Step Method*. By using

their definition of integration and cointegration and their notation, the weekly load data can be expressed as $W(\tau) \sim \mathrm{SI}(d_0^{(W)}, d_{52}^{(W)})$.

On the other hand, for the half-hourly sample data set $\{X(t)\}$, we suppose that the periodic component $\{X_{periodic}(t)\}$ (see equation (4.2)) includes only those periodic components whose frequencies are equal to or higher than the frequency corresponding to a weekly period and the trend component $\{X_{trend}(t)\}$ is expected to be a smooth and slowly changing process. In addition, because the annual seasonal frequency is $2\pi/(52\times7\times48) \approx 0$ on the half-hourly scale, the short-term $\{X_{trend}(t)\}$, therefore, includes the annual variation (annual growth trend and 52 weeks annual seasonal) and has an unknown integration order $d_0^{(X)}$ at the origin. In other words, the variation from annual scale embeds in the "trend" of half-hourly scale.

Since $W(\tau)$ is formed by aggregation of $X(t)$ over a one week period, the trend of the half-hourly data can be regarded approximately as the variation from $W(\tau)$ although $W(\tau)$ is observed weekly. If we define $w(t) = W(\tau)$ when $336 \times (\tau - 1) \leq t < 336 \times \tau$, (i.e. the series $w(t)$ is created by setting it to the value of $W(\tau)$ for every time t in each week indexed by τ) and then possibly take some moving average smoothing. it is obvious then that $W(\tau)$ is a long term variable for the half-hourly load $\{X(t)\}$, and $X_{trend}(t)$ and $w(t)$ are cointegrated even though we do not know the exact integration orders, $d_0^{(W)}$, $d_{52}^{(W)}$ for $W(\tau)$ and $d_0^{(X)}$ for the trend of half hourly data. In other words, there exists a linear combination

$$z(t) = X_{trend}(t) - \alpha w(t) \sim \mathrm{I}(0) \tag{4.3}$$

We use the notation $W(\tau)$ to apply to weekly data, $W(t)$ refer to $w(t)$ and assume available sample data from 1 to T in the following sections.

4.3.1 A Modified Cointegration "Error Correction" Model

If the data process $Y(t) \sim \mathrm{I}(d), d \geq 1$ is cointegrated with another data series $W(t)$ and is assumed to be generated[4] by the following long-run model

$$Y(t) = c_0 + c_1 W(t) + \eta(t) \tag{4.4}$$

and the short-run model

$$\Delta Y(t) = b_0 + b_1' V(t) + \zeta(t) \tag{4.5}$$

where $W(t)$ are long term explanatory variables; $V(t)$ are explanatory variables for $\Delta Y(t)$; $\eta(t)$ and $\zeta(t)$ are I(0), i.e. the disturbance terms of the long-run model (4.4), $\eta(t)$ and the short-run model (4.5), $\zeta(t)$, are stationary series possibly with complex structures. Engle et al. (1989) contended that the data generation process of $Y(t)$ can be represented by a complete "error-correction" model of the form

$$\Delta Y(t) = \delta + r\, z(t-1) + \beta' V(t) + \epsilon(t) \tag{4.6}$$

where $z(t) = Y(t) - c_0 - c_1 W(t) \sim \mathrm{I}(0)$ from the long-run model (4.4) is an ECM term and $\epsilon(t)$ is a white noise disturbance with zero mean and constant variance.

The main difference between the short-run model and the complete "error-correction" model is that the latter model includes an error correction term $z(t-1)$. This term corrects the coefficients of the short-run variables $V(t)$ and the constant term.

The complete cointegrating model (4.6), however, is usually not available in practice for prediction purposes because the long-run variable $W(t)$ is unknown beyond the sample. There are two approaches allowing us to approximate the complete model.

We call the first approach the *naive* approximation which replaces $W(T+i)$ outside the sample by its forecast $\hat{W}(T+i)$ in the long-run model (4.4) if the exact model for $W(t)$ is known. A problem associated with the naive approximation is that the exact model for $W(t)$ is, in practice, most unlikely to be known or at least there are several contending models for $W(t)$. If the model for $W(t)$ is not known completely

[4]In order to avoid too cumbersome notation, we did not introduce a separate notation for the theoretical coefficients and their estimates.

and an assumed model, $\mathcal{M}$, is regarded as the "true" model, the multi-step ahead prediction $\tilde{f}^Y_{T,h}$ of $Y(T+h)$ from that model will be

$$\tilde{f}^Y_{T,h} = f^Y_{T,h} + rc_1 \sum_{i=1}^{h-1} e(T+i) \sum_{j=1}^{h-i} r^j \tag{4.7}$$

where $f^Y_{T,h}$ is the prediction made by the true model, and $e(T+i)$ is the error made by using an assumed model, $\mathcal{M}$, for $W(T+i)$.

It can be seen that the naive approximation is very sensitive to the quality of the assumed model $\mathcal{M}$. The prediction errors are accumulated in $rc_1 \sum_{i=1}^{h-1} e(T+i) \sum_{j=1}^{h-i} r^j$. These errors are caused by the inconsistency of using observed $W(t)$ within the sample and predicted $\hat{W}(T+i)$ beyond the sample.

The second approach to approximation of the full model was suggested by Engle et al. (1989) as follows

$$\Delta y(t) = \delta + r\hat{z}(t-1) + \beta' V(t) + \epsilon(t) \tag{4.8}$$

where $\hat{z}(t) = y(t) - \hat{f}^Y_{t-1,1}$ is the error from the estimated long-run model and $\hat{f}^Y_{t-1,1}$ is the forecast of $y(t)$ made at $(t-1)$ from the long-run model, i.e.

$$\hat{f}^Y_{t-1,1} = c_0 + c_1 \hat{f}^W_{t-1,1} \tag{4.9}$$

where $\hat{f}^W_{t-1,1}$ is the one step ahead prediction made at $(t-1)$ from the assumed model, $\mathcal{M}$.

Once $\hat{f}^W_{t-1,1}$ is obtained from the assumed model, $\mathcal{M}$, $\hat{f}^Y_{t-1,1}$ can be obtained from the approximate long-run model (4.9) and so we can also obtain $\hat{z}(t)$. The one step ahead post sample prediction for $y(T+1)$ is $\hat{f}^Y_{T,1} = y(T) + \delta + r(y(T) - \hat{f}^Y_{T-1,1}) + \beta' f^V_{T,1}(T,1)$. It also can be iterated out to form medium and long term forecasts which would be essentially the same as those obtained from (4.6) by replacing everything by its forecast

$$f^Y_{T,h} = f^Y_{T,h-1} + \delta + r(f^Y_{T,h-1} - \hat{f}^Y_{T,h-1}) + \beta' f^V_{T,h} \tag{4.10}$$

This approximate full model does not require the exact model for $\{W(t)\}$ and avoids the accumulated error problem because it consistently uses $\hat{z}(t)$ and $\hat{z}(T+h)$

generated from the long-term model by the estimated $\hat{W}(t)$ and $\hat{W}(T+h)$ from the assumed model, $\mathcal{M}$. Now, the approximate full model (4.8) is arisen from model (4.6) where $W(t)$ is replaced by $\hat{W}(t)$. Therefore, an additional term, $W(t) - \hat{W}(t)$ may should be introduced into (4.8) as an ECM term. To clarify the above question, we do the following analysis.

Definition:

$$\|Y - \mathbf{P}_{\mathbf{H}(X)}Y\| = \mathbf{E}|Y - \mathbf{P}_{\mathbf{H}(X)}Y|^2 = \min_{C} \mathbf{E}|Y - C'X|^2 \tag{4.11}$$

where $\mathbf{H}(X)$ is a linear space spanned by X, $\mathbf{P}_{\mathbf{H}(X)}Y$ is a projection of Y on $\mathbf{H}(X)$ with properties

- $Y = \mathbf{P}_{\mathbf{H}(X)}Y + Z,\ Z \perp \mathbf{H}(X)$ or $\mathbf{cov}(Z, \mathbf{H}(X)) = 0$.
- If $X_1 \perp X_2$, $\mathbf{P}_{\mathbf{H}(X_1 \cup X_2)}Y = \mathbf{P}_{\mathbf{H}(X_1)}Y + \mathbf{P}_{\mathbf{H}(X_2)}Y = \mathbf{P}_{\mathbf{H}(X_1, X_2)Y}$.

Therefore the long-, short-run and complete model can be re-written as

$$Y = \mathbf{P}_{\mathbf{H}(W)}Y + \eta \quad \text{long-run model} \tag{4.12}$$

$$\Delta Y = \mathbf{P}_{\mathbf{H}(V)}\Delta Y + \zeta \quad \text{short-run model} \tag{4.13}$$

$$\Delta Y = \mathbf{P}_{\mathbf{H}(V \cup Z)}\Delta Y + \epsilon \quad \text{complete model} \tag{4.14}$$

Theorem 4.1 *Suppose we have Y, W, and V with sample $t = 1, \cdots, T$; the model $\mathcal{M}$ is an approximate model for $\{W(t)\}$; $\hat{W}(t)$ is estimation of $W(t)$ made from the model $\mathcal{M}$ in the sample; the residual $e(t) = W(t) - \hat{W}(t)$ is stationary series, i.e. $e \sim \mathrm{I}(0)$, $Z = Y - \mathbf{P}_{\mathbf{H}(W)}Y \sim \mathrm{I}(0)$ $\hat{Z} = Y - \mathbf{P}_{\mathbf{H}(\hat{W})}Y \sim \mathrm{I}(0)$. Then, within sample, there are the following relations,*

$$\begin{aligned} \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z} \cup e)}\Delta Y\| &= \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z \cup e)}\Delta Y\| \\ \leq \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z)}\Delta Y\| &\leq \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z})}\Delta Y\| \end{aligned} \tag{4.15}$$

and if $e(t)$ is white noise,

$$\|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z)}\Delta Y\| = \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z})}\Delta Y\|$$

For post sample, Z is not available and is replaced by $\hat{Z}$. There are the following relations,

$$\|\Delta Y - \mathbf{P}_{1\ \mathbf{H}(V\cup\hat{Z})}\Delta Y\| \leq \|\Delta Y - \mathbf{P}_{0\ \mathbf{H}(V\cup\hat{Z})}\Delta Y\| \tag{4.16}$$

where $\mathbf{P}_{0\ \mathbf{H}(V\cup Z)} = (\delta_0, r_0, \beta_0)$ and $\mathbf{P}_{1\ \mathbf{H}(V\cup\hat{Z})} = (\delta_1, r_1, \beta_1)$ are the estimated coefficients from the complete model by using $z(t-1)$ and $\hat{z}(t-1)$, respectively.

Similarly,

$$\|\Delta Y - \mathbf{P}^e_{1\ \mathbf{H}(V\cup\hat{Z}\cup e)}\Delta Y\| \leq \|\Delta Y - \mathbf{P}^e_{0\ \mathbf{H}(V\cup\hat{Z}\cup e)}\Delta Y\| \tag{4.17}$$

where $\mathbf{P}^e_{0\ \mathbf{H}(V\cup Z\cup e)}$ and $\mathbf{P}^e_{1\ \mathbf{H}(V\cup\hat{Z}\cup e)}$ are the estimated coefficients from $\|\Delta Y - \mathbf{P}_{\mathbf{H}(V\cup Z\cup e)}\Delta Y\|$ by using $z(t-1)$ and $\hat{z}(t-1)$ in the place of Z, respectively.

Consequently,

$$\|\Delta Y - \mathbf{P}^e_{1\ \mathbf{H}(V\cup\hat{Z}\cup e)}\Delta Y\| \leq \|\Delta Y - \mathbf{P}_{1\ \mathbf{H}(V\cup\hat{Z})}\Delta Y\| \leq \|\Delta Y - \mathbf{P}_{0\ \mathbf{H}(V\cup\hat{Z})}\Delta Y\| \tag{4.18}$$

Proof:

We first prove relation (4.15). Because $Z = \hat{Z} + e^*$ and $\hat{Z} \perp e^*$ where $e^* = \mathbf{P}_{\mathbf{H}(e)}Y$, $\mathbf{H}(V \cup Z) = \mathbf{H}(V\cup\hat{Z}) \cup \mathbf{H}(e^* - V\cap e^*)$ and $\mathbf{H}(V\cup\hat{Z})\perp\mathbf{H}(e^* - V\cap e^*)$, then

$$\begin{aligned}
& \|\Delta Y - \mathbf{P}_{\mathbf{H}(V\cup Z)}\Delta Y\| \\
= \ & \|\Delta Y - \mathbf{P}_{\mathbf{H}(V\cup\hat{Z})}\Delta Y - \mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y\| \\
= \ & \|\Delta Y - \mathbf{P}_{\mathbf{H}(V\cup\hat{Z})}\Delta Y\| + \|\mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y\| \\
& -2\mathbf{E}[\Delta Y - \mathbf{P}_{\mathbf{H}(V\cup\hat{Z})}\Delta Y][\mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y]
\end{aligned}$$

and since

$$\begin{aligned}
& \mathbf{E}[\Delta Y - \mathbf{P}_{\mathbf{H}(V\cup\hat{Z})}\Delta Y][\mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y] \\
= \ & \mathbf{E}[\Delta Y][\mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y] \\
= \ & \mathbf{E}[(\Delta Y - \mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y) + \mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y][\mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y] \\
= \ & \mathbf{E}[\mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y]^2 \\
= \ & \|\mathbf{P}_{\mathbf{H}(e^*-V\cap e^*)}\Delta Y\|
\end{aligned}$$

therefore,

$$\begin{aligned} & \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z)} \Delta Y\| \\ = \; & \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z})} \Delta Y\| - \|\mathbf{P}_{\mathbf{H}(e^* - V \cap e^*)} \Delta Y\| \\ \leq \; & \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z})} \Delta Y\| \end{aligned} \tag{4.19}$$

Because $e^* = \mathbf{P}_{\mathbf{H}(e)} Y \subseteq \mathbf{H}(e)$, $\mathbf{H}(Z) = \mathbf{H}(\hat{Z} \cup e^*) \subseteq \mathbf{H}(\hat{Z} \cup e) = \mathbf{H}(Z \cup e)$ and we obtain

$$\begin{aligned} & \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z} \cup e)} \Delta Y\| \; = \; \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z \cup e)} \Delta Y\| \\ \leq \; & \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z)} \Delta Y\| \; \leq \; \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z})} \Delta Y\| \end{aligned}$$

If e is white noise, then $\mathbf{P}_{\mathbf{H}(e)} Y = 0$, i.e. $e^* = 0$. Therefore, from relation (4.19), we have

$$\|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z)} \Delta Y\| = \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z})} \Delta Y\|$$

Now, for post sample, we have

$$\begin{aligned} & \|\Delta Y - \mathbf{P}_{1 \; \mathbf{H}(V \cup \hat{Z})} \Delta Y\| = \|\Delta Y - \delta_1 - r_1 \hat{z}(t-1) - \beta' V(t)\| \\ = \; & \min_{(\delta, r, \beta)} \mathbf{E} |\Delta Y - \delta - r \hat{z}(t-1) - \beta' V(t)|^2 \leq \mathbf{E} |\Delta Y - \delta_0 - r_0 \hat{z}(t-1) - \beta_0' V(t)|^2 \\ = \; & \|\Delta Y - \mathbf{P}_{0 \; \mathbf{H}(V \cup \hat{Z})} \Delta Y\| \end{aligned} \tag{4.20}$$

With the similar argument, we can also obtain,

$$\|\Delta Y - \mathbf{P}^e_{1 \; \mathbf{H}(V \cup \hat{Z} \cup e)} \Delta Y\| \leq \|\Delta Y - \mathbf{P}^e_{0 \; \mathbf{H}(V \cup \hat{Z})} \Delta Y\| \tag{4.21}$$

Because of $\|\Delta Y - \mathbf{P}^e_{1 \; \mathbf{H}(V \cup \hat{Z} \cup e)} \Delta Y\| \leq \|\Delta Y - \mathbf{P}_{1 \; \mathbf{H}(V \cup \hat{Z})} \Delta Y\|$, we establish,

$$\|\Delta Y - \mathbf{P}^e_{1 \; \mathbf{H}(V \cup \hat{Z} \cup e)} \Delta Y\| \leq \|\Delta Y - \mathbf{P}_{1 \; \mathbf{H}(V \cup \hat{Z})} \Delta Y\| \leq \|\Delta Y - \mathbf{P}_{0 \; \mathbf{H}(V \cup \hat{Z})} \Delta Y\|$$

□

Remarks:

The sufficient condition for $e^* = 0$ is that $\{e(t)\}$ is a white noise. Under this condition, from the above theorem, we obain $\|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup Z)} \Delta Y\| = \|\Delta Y - \mathbf{P}_{\mathbf{H}(V \cup \hat{Z})} \Delta Y\|$.

This implies that there are no differences between the naive model and Engle's model and there is no reason to include the error term e in the complete model if the model $\mathcal{M}$ is an adequate model for the long term variable W. However, the adequacy of an approximate model often fails because of parsimony. Some trade-offs occur when using AIC or other information criteria to select a proper model and e is often a stationary series but may not necessarily be white noise.

The above theorem also shows that the post sample prediction from Engle's approximate full model (4.8) is better than that from the naive approximate full model if the model $\mathcal{M}$ is not well enough specified to ensure $\{e(t)\}$ is pure white noise even though the naive model has better fit within sample. Furthermore a better model (see relation (4.15)) can be obtained from,

$$\Delta y(t) = \delta_1 + r_1 z(t-1) + \beta_1' V(t) + \rho_1 e(t-1) + \eta_1(t) \tag{4.22}$$

or

$$\Delta y(t) = \delta_2 + r_2 \hat{z}(t-1) + \beta_2' V(t) + \rho_2 e(t-1) + \eta_2(t) \tag{4.23}$$

Within the sample data set, the above two modified complete models are equivalent. The prediction error, $\{e(T+h)\}$, of the assumed model for $\{W(T+h)\}$ will be unknown for post sample and assumed zero. Although the modified complete model (4.22) has the same form as the naive full model (4.6) for post sample prediction, the estimated parameters δ_1, r_1, β_1 are influenced by the inclusion of $\{e(t-1)\}$. Therefore, the pure effects of $z(t-1)$ and $V(t)$ on $\Delta y(t)$ can be best estimated by the modified model. However, from relation (4.18), we know that model (4.22) would be worse than model (4.23) in post sample prediction. The reason is that model (4.22) still suffers from the accumulated errors as does the naive full model if $z(T+h)$ is constructed by using predicted $\hat{W}(T+h-1)$ from the assumed model for post sample points while model (4.23) does not.

The modified complete model (4.23) not only has the same fitting performance within sample as the modified complete model (4.22) but also, more importantly, avoids the accumulated errors for post sample forecasts as does Engle's approximate

full model since it consistently uses $\hat{z}(t-1)$ within sample and $\hat{z}(T+h-1)$ outside the sample as an "error correction" term. Therefore, we conclude that the error term e should be included in the complete error-correction model if $\{e(t)\}$ is stationary but not necessarily white noise and the following modified long- and short-run merged model is superior to both the naive full model and the modified complete model (4.22), and will be better than Engle's approximate full model too (or at least will be no worse).

$$\begin{cases} W(t) = \hat{W}(t) + e(t) & \text{model } \mathcal{M} \\ y(t) = c_0 + c_1\hat{W}(t) + \eta(t) & \text{long-run model} \\ \Delta y(t) = b_0 + b_1' V(t) + \zeta(t) & \text{short-run model} \\ \Delta y(t) = \delta + r\,\hat{z}(t-1) + \beta' V(t) + \rho e(t-1) + \epsilon(t) & \text{complete model} \end{cases} \tag{4.24}$$

where $\hat{W}(t)$ is one step ahead prediction made at $t-1$ from an assumed model, $\mathcal{M}$, for $W(t)$; $e(t) = W(t) - \hat{W}(t) \sim \mathrm{I}(0)$ is the long term variable prediction error; $\hat{z}(t) = y(t) - c_0 - c_1\hat{W}(t) \sim \mathrm{I}(0)$.

This model is identical to Engle's approximate full model (4.8), and collapses to the naive full model (4.6) if the model, $\mathcal{M}$, for the long-run variable $W(t)$ is sufficiently well specified to ensure that the innovation series $\{e(t)\}$ is white noise. Therefore, this modified complete cointegration "error correction" model (4.23) is a more general model and (4.24) will be referred as the proposed model and will be employed here.

4.3.2 An Artificial Trend from the Half-hourly Data

The real problem being considered here is that $\{X_{trend}(t)\}$ is an unobservable trend component and the approximation of the complete "error correction" model (4.24) cannot be used directly to model the trend component. An approximation is made by assuming that the "weekly adjusted" series

$$\tilde{X}(t) = (1 + B + \cdots B^{(s-1)})X(t) \tag{4.25}$$

is an approximation to $X_{trend}(t)$ where $s = 336$ (one week for half hourly data).

The long-term explanatory variable, $W(\tau)$, in the long-run model (4.4) is the weekly load level. A reasonably simple model can be used to fit $W(\tau)$ since the inadequate error of the model can be partially corrected in the final "error correction" model (4.24). We used the "Airline model" ARIMA$(0,1,1)\times(0,1,1)_{52}$ although the model may not be strictly adequate for any particular segment of the data set. The proposed complete "error correction" model would be

$$\Delta\tilde{X}(t) = (1-B^s)X(t) = \delta + r\hat{z}(t-1) + \beta' V(t) + \rho e(t-1) + \epsilon(t) \tag{4.26}$$

where $\hat{z}(t) = \tilde{X}(t) - c_0 - c_1\hat{W}(t)$, $e(t) = W(t) - \hat{W}(t)$.

Since there are no short term explanatory variables (e.g. weather variables[5]) available, $V(t)$ is assigned to be the unit vector when the trend is believed or assumed to be locally linear. The final modified "error correction" model would be

$$(1-B^s)X(t) = \delta^* + r\hat{z}(t-1) + \rho e(t-1) + \epsilon(t) \tag{4.27}$$

where $\delta^* = \delta + \beta$, and the trend model

$$X_{trend}(t) = X_{trend}(t-1) + \delta^* + r\hat{z}(t-1) + \rho e(t-1) + \epsilon(t) \tag{4.28}$$

Consequently, the one step ahead post sample prediction for $X_{trend}(T+1)$ will be

$$X_{trend}(T+1) = X_{trend}(T) + \delta^* + r(X_{trend}(T) - f^X_{T-1,1}) + \rho e(T) \tag{4.29}$$

and the term $e(T+h)$, $(h>1)$ will be zero for multiple step ahead predictions and post sample predictions can be obtained iteratively by

$$f^X_{T,h} = f^X_{t,h-1} + \delta^* + r(f^X_{t,h-1} - \hat{f}^X_{T,h-1}) \tag{4.30}$$

where $\hat{f}^X_{T,h} = f^X_{T,h} - c_0 - c_1 f^W_{T,h}$.

Although the long-run part of the proposed model may not fit the trend as well as the long-run model for the naive approach, it retains the one step ahead prediction error from the assumed model for $\{W(t)\}$, $\mathcal{M}$, in $\{\hat{z}(t)\}$, which provides more

[5]The relationship between load and temperature is non-linear and time dependent. This topic will be discussed in chapter 6. A non-linear transformation is needed to transfer temperature into another variable which is linearly related to the load and serves as a explanatory variable.

information for the post sample prediction, This part can then be corrected in the modified complete model (4.29) by influencing the coefficients of the short-run explanatory variables.

The three "error correction" models are applied to half-hourly data for a period randomly chosen, i.e. April 24, 1983 to July 2, 1983, to illustrate the differences between the three cointegration "error correction" models. The data formed by using the weekly adjustment operation (4.25) are assumed to be the "trend" component since our task is to model the unobserved trend component.

The results from the naive model, Engle's model and our proposed model are listed in Table 4.3 and Table 4.4. For the long run model, we see that the naive model fits better than does Engle's or our proposed model since it uses the real long-term data $W(t)$ while Engle's and the proposed model use the one step predicted values $\hat{W}(t)$ from the assumed long-term model. For the complete model, by comparing the R^2 values of the three complete models in Table 4.3 and 4.4, we can see that the proposed model fits best followed by Engle's model. The proposed model employs more short-run information from its long-run model than does the other two models. The worst performance is the naive model because its complete model does not include possible short-run information contained in the error term $\{e(t)\}$. This result is consistent with the conclusion of Theorem 4.1. From the t statistic value for ρ in Table 4.4, it is also confirmed that the one step prediction error of the long-term model of $W(t)$ does play a significant role in the complete "error correction" model.

The first plot in Figure 4.3 shows the estimated trend components from the modified model (M.C.trend) and the Engle's model (E.C.trend) over the naive model (N.C.trend) for within sample data when the assumed "trend" (Trend) is formed by the weekly adjusting operation (4.25). Since the short-run model for trend includes time only and there are no other explanatory and reference variables available, the trend prediction post sample is a linear function of time for a weekly interval. The second plot of Figure 4.3 presents the post sample prediction performance of the different models after the vertical line. The dot curve (legend W.trend) is obtained from

coef.	value	std.err.	t.stat.	p.value
Naive Long Run Model				
Residual Standard Error = 0.0141, Multiple R-Square = 0.9502				
N = 3024,F-statistic = 57655.4 on 1 and 3022 df, p-value = 0				
c_0	-9.3151	0.0817	-114.0838	0
c_1	0.9798	0.0041	240.1154	0
Naive Complete Model				
Residual Standard Error = 1e-04, Multiple R-Square = 0.8527				
N = 3022, F-statistic = 17478.47 on 1 and 3020 df, p-value = 0				
δ	0.0001	0.0000	35.9350	0
r	0.9520	0.0072	132.2062	0

Table 4.3: The Naive Cointegrating 'Error Correction' Model

coef.	value	std.err.	t.stat.	p.value
Long Run Model				
Residual Standard Error = 0.0305, Multiple R-Square = 0.7674				
N = 3024, F-statistic = 9972.934 on 1 and 3022 df, p-value = 0				
c_0	-11.3460	0.2167	-52.3677	0
c_1	1.0811	0.0108	99.8646	0
Engle's Complete Model				
Residual Standard Error = 1e-04, Multiple R-Square = 0.8696				
N = 3022, F-statistic = 20140.89 on 1 and 3020 df, p-value = 0				
δ	0.0000	0.0000	27.2571	0
r	0.8912	0.0063	141.9186	0
Proposed Complete Model				
Residual Standard Error = 1e-04, Multiple R-Square = 0.8923				
N = 3022, F-statistic = 12512.63 on 2 and 3019 df, p-value = 0				
δ	0.0001	0.0000	35.1390	0
r	0.9495	0.0062	154.2384	0
ρ	-0.4282	0.0170	-25.2538	0

Table 4.4: Comparison Between Engle's & the Proposed Cointegration Models

i^a	W.trend	N.C.trend	E.C.trend	M.C.trend
1	1.717E-05	9.784E-06	1.340E-05	3.916E-06*[b]
2	3.486E-05	1.592E-05	5.845E-06	1.520E-06*
3	1.462E-04	9.224E-05	4.760E-05	2.123E-05*
4	5.098E-04	3.814E-04	2.472E-04	1.559E-04*
5	8.884E-04	6.867E-04	4.549E-04	2.954E-04*
6	8.040E-04	5.862E-04	3.348E-04	1.774E-04*
7	3.129E-04	1.738E-04	4.996E-05	2.315E-05*
8	6.476E-05	1.655E-05*	5.176E-05	1.757E-04
9	4.193E-05*	2.109E-04	4.492E-04	7.767E-04
10	1.609E-04*	5.460E-04	8.827E-04	1.342E-03
11	2.922E-04*	9.057E-04	1.295E-03	1.859E-03
12	3.998E-04*	1.245E-03	1.657E-03	2.306E-03
13	4.885E-04*	1.573E-03	1.990E-03	2.717E-03
14	3.647E-04*	1.510E-03	1.879E-03	2.606E-03

Table 4.5: The Comparison of Post Sample Predictions in the Different Models

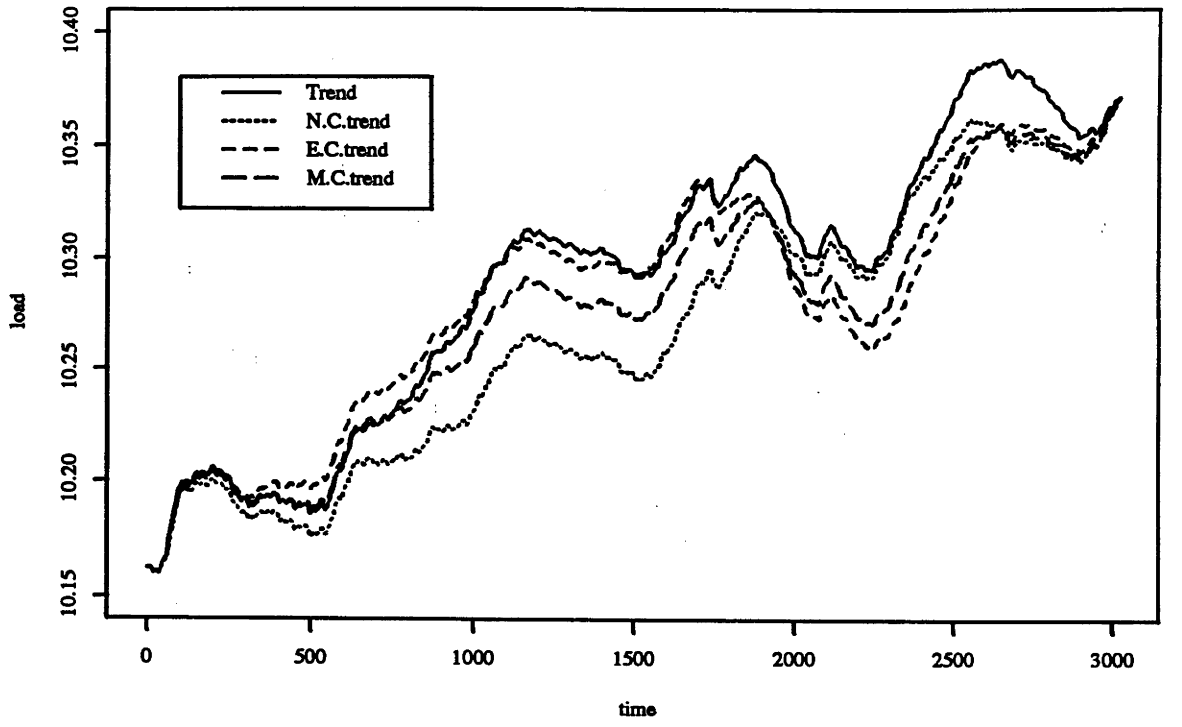
[a] i = the number of days ahead.
[b] * indicates the smallest MSE among the four differet models.

the long-run model in which long-term weekly data and its post sample prediction are included, i.e. regressing $Y(t)$ on the long-term variable $W(t)$[6] and there is no ECM included.

Engle's model improves short-term prediction over the naive model and the proposed model further improves short-term prediction over the Engle's model. Therefore, the accuracy of short-term post sample trend prediction can be compared in the proposed model, the Engle's model, the naive model and the naive long run model. On the other hand, the accuracy of the long-term post sample trend prediction moves in the opposite direction. We use MSE as measure of the post sample prediction accuracy properties of the different models. The MSE of the different models over a daily period for 14 days are presented in Table 4.5. We can see clearly that the proposed model has the best short term prediction performance. Therefore, it will be employed since our interest is in the short-term trend behaviour. In chapter 5, we will present the overall post sample prediction of the proposed model and comparisons with other

[6] $W(t)$ is estimated and predicted by using the assumed model, $\mathcal{M}$, on the weekly data, generating a set of step functions with values at every half-hourly t in the corresponding weeks and then taking a weekly moving average.

Fitting of Cointegration 'Error Correction' Models



Post Sample Prediction of Cointegration 'Error Correction' Models

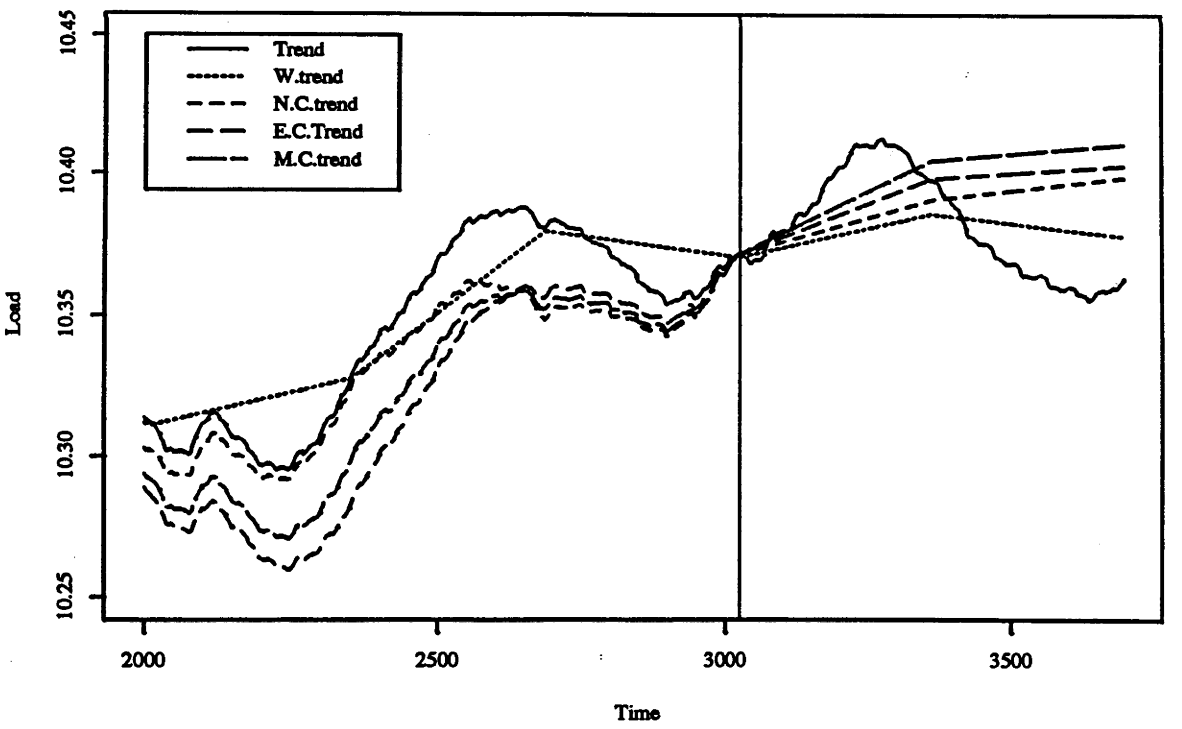


Figure 4.3: The Performance of Cointegration "Error Correction" Models

models.

The approximate detrended data, $Y(t)$, in the sample are obtained by forming

$$Y(t) = X(t) - X_{trend}(t) \tag{4.31}$$

and assuming that there is no trend presented in $\{Y(t)\}$ and the post sample forecast of $X_{trend}(T+h)$ can be made by (4.30).

4.4 A Basic Model for Periodical Time Series

Given a time series $\{Y(t); t = 0, 1, 2, 3, \cdots\}$ whose first and second order moments exist though they are not necessarily constant. We define its mean as

$$m(t) = \mathbf{E}Y(t) \tag{4.32}$$

and its autocovariance as

$$R(s,t) = \mathbf{E}(Y(s) - m(s))(Y(t) - m(t)) \tag{4.33}$$

Such a time series with the above properties is a non-stationary time series. A periodically stationary time series is defined below:

Definition:

The time series $\{Y(t)\}$ is defined as periodically stationary of period d, if for a positive integer d and for all integers, s, t, the following conditions are satisfied

$$m(t) = m(t+p), \quad R(s,t) = R(s+p, t+p) \tag{4.34}$$

The basic approach to modelling the periodical time series $Y(t)$ with period d, $t = 1, 2, 3, \cdots$ is to decompose the value $Y(t)$ into the sum of two components as follows:

$$Y(t) = Y_{(e)}(t) + Y_{(u)}(t) \tag{4.35}$$

where $Y_{(e)}(t)$ is the explained or predictable part of $Y(t)$, and $Y_{(u)}(t)$ is the error or unexplained, or unpredictable part of $Y(t)$.

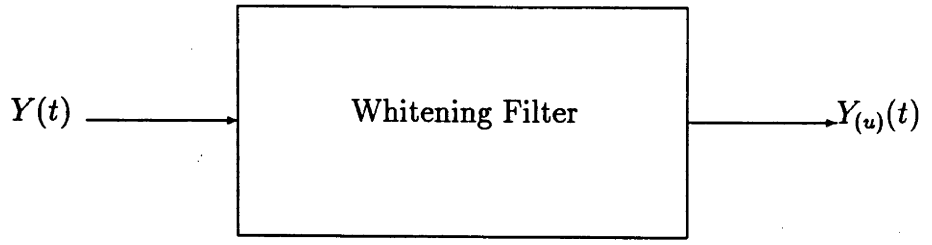


Figure 4.4: Pre-Whitening Filter

It is assumed that $Y(t)$ is a normal process. We define

$$Y_{(e)}(t) = \mathbf{E}[Y(t)|Y(t-1), Y(t-2), \cdots] \tag{4.36}$$

Hence, $Y_{(e)}(t)$ is a linear combination of $Y(t-1)$, $Y(t-2)$, $\cdots$.

$$Y_{(u)}(t) = Y(t) - \mathbf{E}[Y(t)|Y(t-1),\ Y(t-2),\ \cdots] \tag{4.37}$$

$$Y_{(u)}(t) = \sum_{i=0}^{\infty} a(i)Y(t-i), \quad a(0) = 1 \tag{4.38}$$

$Y_{(u)}(t)$ is called the *innovation* series, which is an independent zero-mean random variable and its variance is denoted:

$$\sigma^2(t) = \mathbf{E}[|Y_{(u)}(t)|^2] \tag{4.39}$$

To find $Y_{(e)}(t)$ or $Y_{(u)}(t)$ is an equivalent problem. The problem of modelling $Y(t)$ includes finding a pre-whitening filter showed in Figure 4.4 which transforms $Y(t)$ to $Y_{(u)}(t)$.

A general model for a time series $Y(t)$ is that there is a deterministic curve $m(t)$ representing mean values about which $Y(t)$ is fluctuating with variance $\sigma^2(t)$ that may be time-varying. The following model can be assumed:

$$Y(t) = m(t) + \sigma(t)Z(t) \tag{4.40}$$

where, $m(t)$ is the periodical mean value function which period d, $m(t) = \mathbf{E}[Y(t)]$, $\sigma^2(t)$ is the periodical variance function, $\sigma^2(t) = \mathbf{var}[Y(t)]$, and $Z(t)$ is the fluctuation function,

$$Z(t) = \frac{Y(t) - m(t)}{\sigma(t)} \tag{4.41}$$

Therefore, the pre-whitening filter is designed to remove the periodical mean $m(t)$ and variance $\sigma^2(t)$ from $Y(t)$ and generate the innovation series $Z(t)$.

To model the fluctuation function $Z(t)$, we assume that it is normally distributed with zero mean and unit variance, which is stationary in the sense that there exists a correlation function $\rho(\nu)$ satisfying

$$\mathbf{E}[Z(t)Z(t+\nu)] = \rho(\nu) \tag{4.42}$$

Another assumption is that the fluctuation function $Z(t)$ is *periodically stationary* with period d in the sense that there exists a correlation function $\rho_1(\nu), \rho_2(\nu), \cdots, \rho_d(\nu)$ satisfying

$$\mathbf{E}[Z(t)Z(t+\nu)] = \rho_{j(t)}(\nu) \tag{4.43}$$

where, $j(t)$ is an index j satisfying $j(t) = \mathrm{mod}(t, d)$, for $j = 1, 2, \cdots, T$

The periodically stationary model for $Z(t)$ may be more general in some circumstances. Pagano (1978) suggests one uses a d dimensional autoregression model where the d dimensional vector is comprised by $V(k) = (Z(kd+1), Z(kd+2), \cdots, Z(kd+d))$, $k = 1, 2, \cdots, [N/d](= K)$ to fit the periodically stationary $Z(t)$. However, there could too many parameters to be estimated from the available sample size N when the period, d, is large. For instance, if $d = 48$ (for daily period of half-hourly data) and the maximum order of the vector autoregression is $p = 1$, and the number of periods is $K = 10$, there are only $Kd = 480$ data points available to estimate $d^2 = 2304$ parameters in the lag 1 coefficient matrix of the vector autoregression model and $d(d-1)/2 = 1128$ parameters in the correlation matrix of the 48 dimension vector. In practice, it is impossible to estimate the model although Penm and Terrell

(1982) provided a way to reduce the number of parameters involved. Therefore, the assumption that $Z(t)$ is periodically stationary is not realistic in our case.

Suppose $Y(t)$, $t = 1, 2, 3, \cdots, t_0$ where $t_0 = nd$. We need to carry out the following steps to pre-whiten $Y(t)$:

1. Estimate and model the time series mean value function $m(j)$, $j = 1, 2, 3, \cdots, d$
2. Estimate and model the time series variance function $\sigma^2(j)$, $j = 1, 2, 3, \cdots, d$
3. If $\bar{m}(j)$ and $\bar{\sigma}(j)$ denote the fitted means and variances, $Z(t)$ is estimated by

$$\bar{Z}(t) = \frac{Y(t) - \bar{m}(j(t))}{\bar{\sigma}(j(t))} \tag{4.44}$$

The estimators of $m(j)$ and $\sigma^2(j)$ for a fixed $j = 1, 2, \cdots, d$ are given by

$$\bar{m}_n(j) = \frac{1}{n}\sum_{k=0}^{n-1} Y(j + kd) \tag{4.45}$$

$$\bar{\sigma}_n^2(j) = \frac{1}{n}\sum_{k=0}^{n-1} [Y(j + kd) - \bar{m}_n(j)]^2 \tag{4.46}$$

$\bar{m}_n(j)$ and $\bar{\sigma}_n^2(j)$, $j = 1, 2, 3, \cdots, d$ are called estimated *periodical* means and *periodical variances* , respectively.

The smooth relations among the $m(j)$ $(j = 1, 2, 3, \cdots, d)$ can be handled by fitting a harmonic representation;

$$\hat{m}(j) = m_n + \sum_{k=1}^{[d/2]} [a_n(k)\cos(j\frac{2\pi}{d}k) + b_n(k)\sin(j\frac{2\pi}{52}k)] + a_n([d/2]+1)\cos \pi j \tag{4.47}$$

The estimators of m_n, $a_n(j)$, $b_n(j)$ can be obtained from $\bar{m}_n(j)$ by the formulae;

$$\bar{m}_n = \frac{2\pi}{d}\sum_{k=1}^{d} \bar{m}_n(k) \tag{4.48}$$

$$\bar{a}_n(j) = \frac{2\pi}{d}\sum_{k=1}^{d} \bar{m}_n(k)\cos(j\frac{2\pi}{d}k) \tag{4.49}$$

$$\bar{b}_n(j) = \frac{2\pi}{d}\sum_{k=1}^{d} \bar{m}_n(k)\sin(j\frac{2\pi}{d}k) \tag{4.50}$$

The possibility that $m(j)$ is equal to a common constant m for $j = 1, 2, 3, \cdots$ as well as other smooth relations among $m(j)$ can be identified by testing whether $a_n(j)$ and $b_n(j)$ are significantly different from zero. In order to do the test, the probability distributions of $\bar{a}_n(j)$ and $\bar{b}_n(j)$, $j = 1, 2, 3, \cdots, d$ are required. However, these distributions depend on the properties of the fluctuation function $Z(t)$ whose estimation is the key issue of the investigation. Initially, under the assumption that $Z(t)$ is *white noise* , one can test $a_n(j)$ and $b_n(j)$ by

$$R_n(j) = n\frac{\bar{a}_n^2(j) + \bar{b}_n^2(j)}{\bar{\sigma}_n^2} \tag{4.51}$$

where $\bar{\sigma}_n^2 = \frac{1}{d}\sum_{j=1}^{d} \bar{\sigma}_n^2(j)$. $R_n(j)$ has a probability distribution of χ^2 with 2 degrees of freedom. The index j is regarded as significant if $R_n(j)$ is above a suitable threshold. Hence, the final periodical fitted means are defined by

$$\begin{aligned} \bar{m}_n(j) \;=\; & m_n + \sum_{\text{significant } k} [\bar{a}_n(k)\cos(j\frac{2\pi}{d}k) + \bar{b}_n(k)\sin(j\frac{2\pi}{d}k)] \\ & + \bar{a}_n([d/2]+1)\cos(\pi j) \end{aligned} \tag{4.52}$$

The estimated periodical means $\bar{m}_n(j)$ and the periodical means $m_n(j)$ are not expected to be significantly different and, the periodical variances $\sigma_n^2(j)$ and the estimated periodical variances $\bar{\sigma}_n^2(j)$ are also not expected to be significantly different either. The estimated means and variances are used in our model because it seems preferable to fit as smooth a periodic mean value function as possible to a periodically varying time series.

The estimated innovation time series $\bar{Z}(t)$ is found according to equation (4.44). $Z(t)$ is a stationary series with zero means and unit variance, and is not white noise in most circumstances. In other words, $Z(t)$ may be thought of as the sum of an explained part, $Z_{(e)}(t)$ and an unexplained part $Z_{(u)}(t)$, i.e.

$$Z(t) = Z_{(e)}(t) + Z_{(u)}(t) \tag{4.53}$$

Following the approach set out in equations (4.36), (4.37), (4.38), the whitening filter, which transforms $\{Z(t)\}$ to its innovation series $\{Z_{(u)}(t)\}$ with a time-invariant

variance is,

$$Z_{(u)}(t) = \sum_{i=0}^{\infty} \alpha(i) Z(t-i), \quad \alpha(0) = 0 \tag{4.54}$$

where $\sigma^2 = \mathbf{E}[|Z_{(u)}(t)|^2]$.

There are many methods in the literature to model the stationary innovation series, $\{Z_{(u)}(t)\}$. For the sake of parsimony, we use the subset AR model and the selection procedure developed in Chapter 2.

4.4.1 Periodic Pattern of Weekdays and Weekend Days

According to the new model described and the basic model for the periodic time series in section 4.1, we sort out a weekdays' series $\{Y_{wd}(\cdot)\}$ and a weekend days' series $\{Y_{we}(\cdot)\}$ from the detrended series $\{Y(t)\}$. Unlike $\{Y(t)\}$, the series $\{Y_{wd}(\cdot)\}$ and $\{Y_{we}(\cdot)\}$ are periodically stationary processes with period 48 (daily). For the sake of simplicity, we treat $\{Y_{wd}(\cdot)\}$ and $\{Y_{we}(\cdot)\}$ as $\{Y(\cdot)\}$ without confusion because they have same properties except differents in their profiles.

Since the daily periodic pattern dominates the profile of $\{Y(\cdot)\}$, it can be assumed that the periodic mean of $\{Y(\cdot)\}$ can be approximately expressed by the daily frequency and its harmonic frequencies plus an error term as mentioned in last section (see equation (4.47))

$$m_Y(j) = m + \sum_{k=1}^{24} (a(k)\cos(k\frac{2\pi}{48}j \quad) + b(k)\sin(k\frac{2\pi}{48}j \quad)) + Z_w(j) \tag{4.55}$$

In practice, however, the harmonic frequencies may not be adequate to fit the periodic pattern, i.e. the periodic mean may not be exactly harmonic with respect to the daily fundamental frequency. Some non-harmonic frequencies may be involved. On the other hand, some harmonic frequencies may not be significant in the harmonic model of the periodical means. In general, therefore, the formula for fitting periodical means should be

$$m(j) = m + \sum_{k=1}^{K} (A_k \cos(\omega_k j) + B_k \sin(\omega_k j)) \tag{4.56}$$

where $\{\omega_k\}$ may not include all harmonic frequencies and may not be only the harmonics of the fundamental frequency.

The harmonic model is a special case of the above model. For the estimation of the frequencies ω_k, $k = 1, 2, \cdots, K$ and the corresponding A_k, B_k, the least squares estimation algorithm of Bloomfield (1976) can be employed. to analysis the $m(j)$ function for $\{Y(\cdot)\}$ if the initial values of $\omega_k, A_k, B_k, \ k = 1, 2, 3, \cdots, m$ are provided. The problem we are facing is how to get those initial values, especially, ω_k which are essential for the Bloomfield's least square estimation algorithm to obtain the optimal multiple frequency model for the periodical means for both weekdays' and weekend days' data.

Initial Values of General Multiple Frequency Model

From Bloomfield's least square estimation algorithm, we know that the initial frequency values predominate over the other initial parameters in the algorithm. Based on the *empirical preparatory stage* in Moutter's super-resolution algorithm Moutter et al. (1986b), a procedure which selects the initial frequencies is established as follows:

Step 1 Suppose that the sampled $\{Y(t)\}$ is $\{Y_t\}$ with sample interval Δt. i.e. $Y_t = Y(t \times \Delta t)$, $t = 0, \pm 1, \pm 2, \pm 3, \cdots$.

It is known that Y_t has a spectral representation which extends only over the frequency range $(-\pi/\Delta t, \ \pi/\Delta t)$.

The basic reason being that when t is restricted to integer multiples of Δt, we can not distinguish the frequency components between $e^{i\omega t}$ and $\{e^{i(\omega + 2k\pi/\Delta t)t}\}$. The components in $Y(t)$ with frequencies $\omega \pm 2\pi/\Delta t, \omega \pm 4\pi/\Delta t, \omega \pm 6\pi/\Delta t, \cdots$ will all appear to have frequency ω. These frequencies are said to be aliases of ω and every frequency outside the range $(-\pi/\Delta t, \ \pi/\Delta t)$ has an "alias" inside this range (see Priestley (1982), pp. 504 - 508).

If Y_t is a bandlimited, say at ω_0, process, then the spectrum of Y_t within the bandlimit ω_0 is of interest and we can apply a band-pass filter to Y_t to reduce the aliasing effect. In practice, the frequency limit beyond which Y_t has negligible power has to be based on a knowledge the nature of the process generating the data set. For example, if Δt is equal to one half-hour, the frequencies of $Y(t)$ higher than a cycle per hour (f_0) cannot be presented correctly in the spectrum of Y_t according to the sampling theorem that Δt must satisfy $\Delta t \leq 1/2f_0$ where $\omega_0 = \pi f_0$. There are various low frequency band-pass filters which can be applied to Y_t to reduce (effectively) all frequencies higher than f_0. A symmetric moving average filter is a simple low frequency band-pass filter which has the advantage of retaining the same phase as Y_t after filtering. The length of the symmetric moving average filter can be determined by the frequency response of the filter needed to cut out the frequencies higher than ω_0.

Step 2 The filtered time series is loaded into a mixed radix fast Fourier transform (MXFFT) array whose length is equal to the sample data segment. The MXFFT is run to form $F_0(H)$, and the initial spectral estimate of the time series is:

$$F_0(H) = MXFFT[Y_t] \tag{4.57}$$

where, H is used to represent the harmonic order such that $\omega_H = 2\pi H/N$; N is the length of the MXFFT array.

Step 3 The spectrum is truncated to the known or desired harmonic limit by deleting high frequency components which we are not interested in (e.g. frequencies which are higher than one hour per cycle). An indicator function $B(H)$ is used.

$$B(H) = \begin{cases} 1 & \text{if } |H| \leq \mu \\ 0 & \text{otherwise} \end{cases}$$

where H is the bandlimit harmonic order.

Hence, the truncated spectrum, say $F_1(H)$, is formed by:[7]

$$F_1(H) = F_0(H) * B(H) \tag{4.58}$$

Step 4 In the spectrum $F_0(H)$, a component is most likely to be the true component of $F(H)$ if it has an amplitude greater than the components on either side of it. In practice, adjacent components are unlikely to be exactly equal in amplitude. $F_2(H)$ is formed by reducing $F_1(H)$ to groups of three harmonic components in which it is expected that the true components will be contained.

In mathematical terms:

$$\begin{aligned} F_2(H) &= F_1(H) \quad \text{if } |F_1(H-2)| < |F_1(H-1)| > |F_1(H)| \\ &\qquad\quad \text{or } |F_1(H-1)| < |F_1(H)| > |F_1(H+1)| \\ &\qquad\quad \text{or } |F_1(H)| < |F_1(H+1)| > |F_1(H+2)| \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

for $H = 0,\ 1,\ 2,\ 3,\ \cdots,\ [\mu n]$[8]

Step 5 Delete all components with an amplitude below a lower threshold E_l:

$$F_3(H) = \begin{cases} F_2(H) & \text{if } |F_2(H)| < E_l \\ 0 & \text{otherwise} \end{cases}$$

for $H = 0,\ 1,\ 2,\ 3,\ \cdots,\ [\mu n]$ where $E_l \simeq 0.05 \max |F_2(H)|$.

In this way, components of relatively low amplitude and therefore of little importance can be excluded.

Step 6 The largest harmonic components are selected from the groups of three harmonic components in $F_2(H)$. It is expected that the largest harmonic components are likely to be the "true" harmonic components of $F(H)$.

[7]The symbol, *, represents the convolution operation.
[8]The Symbol $[\mu n]$ means take the integer part of μn.

In mathematical terms:

$$F_4(H) = \begin{cases} F_3(H) & \text{if } |F_3(H)| = \max(|F_3(H-1)|,\ |F_3(H)|,\ |F_3(H+1)|) \\ 0 & \text{otherwise} \end{cases}$$

In this way, we obtain the selected harmonic component orders F_4. The corresponding harmonic frequencies satisfy

$$\omega_k = 2\pi \frac{H_k}{N}$$

where H_k satisfies $F_4(H_k) > 0$.

The ω_k obtained from the above 6 step procedure and $A_k = B_k = 0$ can be used as a set of initial values for the model (4.56). Furthermore, the significance of the ω_j can be tested by a statistic similar to $R_n(j)$ in equation (4.51) which is distributed approximately as a χ^2 with 2 degrees of freedom if there is no significant effect of ω_j.

4.5 Transition between Week and Weekend Days

In section 4.2, we have shown that weekdays' data $X_1(t_{wd})$ and weekend days data $X_2(t_{we})$ from the short-term data have different characteristics (see Table 4.1 and Table 4.2). According to the model (equation (4.1)), we assume Monday, T_{mon}, and Friday, T_{fri}, are interim periods in which the characteristics of weekend days are in transition to the weekday characteristics, or vice versa. For reasons of simplicity, we just discuss the Friday case since the proposed procedure deals similarly with Friday and Monday.

$$x(t_{fri}) = f(t_{fri})x(t_{thu}) + (1 - f(t_{fri}))x(t_{sat}) \tag{4.59}$$

where $t_{fri} \in T_{fri}$, t_{thu} is the preceding Thursday, t_{sat} is the following Saturday. They satisfy $t_{fri} = t_{thu} + 48$, and $t_{sat} = t_{fri} + 48$.

In general, the weight function $f(t_{fri})$ (see equation (4.59)) is a nonlinear function. If we suppose that most industrial companies and businesses stop working around 4

PM to 5 PM and begin a weekend when the Friday is a routine day, which excludes any special event, such as strikes, public holiday, etc. Therefore, the load characteristic of the Friday before 4 PM is still weekday's; however, the load character of the Friday after 5 PM is moving to a weekend's. The transition period is from 4 PM to 5 PM. Normally, the transition characteristics of the weight function is unknown; the transition period is also unknown and varies.

From the above example and discussion, we assume that a continuous monotone function can approximately describe the character of the weight function $f(t)$. Several numerical procedures are developed to estimate this weight function in the following sections.

4.5.1 Parametric Function Family Approach

Suppose $\{f_{(d)}(t) \mid d \in \mathcal{R}^n\}$ is a weight function family, where d is the parameter set of the function; $\mathcal{R}^n$ is a n dimensional real number space. We construct a new series for each value of the weight function $f_{(d)}(t)$

$$\gamma_{(d)}(t) = x(t_{fri}) - f_{(d)}(t)x(t_{thu}) - (1 - f_{(d)}(t))x(t_{sat}), \quad \text{for } d \in \mathcal{R}^n$$

If the weight function $f_d(t)$ describes the transition properly, the series $\gamma_{(d)}(t)$ would be normally distributed white noise with mean zero and variance $\sigma^2_{f_d}$. Our task is to choose a function in the weight family function, i.e. $\{f_{(d)}(t) \mid d \in \mathcal{R}^n\}$. For this purpose, therefore, a statistic can be constructed to choose an optimal weight function which describes the transition between weekday and weekend days in the family by

$$t_{(d)} = \frac{\bar{\gamma}_{(d)}}{S^2_{(d)}} \tag{4.60}$$

where $\bar{\gamma}_{(d)} = \frac{1}{48}\sum_{t=1}^{48}\gamma_{(d)}(t)$, and $S^2_{(d)} = \frac{1}{47}\sum_{t=1}^{48}(\gamma_{(d)}(t) - \bar{\gamma}_{(d)})^2$

If $\gamma_{(d)}(t)$ is a white noise series, the statistic t_d has a t distribution with mean zero and 47 degree of freedom. The optimal weight function in the given family of

functions $\{f_{(d)}(t)|d \in \mathcal{R}^n\}$ is chosen with the use of the following criteria.

$$f(t) = f_{(d_0)}(t)$$

where $|t_{(d_0)}| = \min(|(t_d)|,\ d \in \mathcal{R}^n)$.

There are countless choices of the family of functions which can approximately describe the transition. For parsimony, we chose those which have the least number of independent parameters and can provide the required transition characteristics.

The transition characteristic mentioned at the beginning of this section, a step function

$$S_\tau(t) = \begin{cases} 1 & t < \tau \\ 0 & t \geq \tau \end{cases} \tag{4.61}$$

can be employed to roughly describe the transition. For instance, suppose, the electrical load before 4PM on a Friday is a typical weekday. However, the load profile after 8PM on the Friday is a typical weekend. The transition could roughly be considered as taking place "suddenly" at 6PM. i.e. the step function $S_{6PM}(t)$ roughly describes the transition. We know that the transition seldom take place as suddenly as a step jump rather it is a "slowly" evolving process. The transition rate, therefore, should take this into account. Since the electricity load at the beginning of a Friday maintains the weekday load profile, but changes to the weekend day load profile by the end of the Friday, the transition rate must be a concave function with a value of zero at both the beginning and the end of the Friday. The pole point of the transition rate function is the time at which the transition takes place most rapidly. For instance, in the above example, the transition rate function satisfies

$$\frac{dS_{6PM}(t)}{dt} = \begin{cases} 0 & t \neq 6\text{PM} \\ -\infty & t = 6\text{PM} \end{cases}$$

From the above discussion, it is seen clearly that the step function will not describe the transition accurately in most of the usual circumstances. A bounded concave function may be more appropriate to describe the transition rate. However, we have to make some assumptions to simplify the concave functions since the skewness of

the concave function is not known. A simple approximation for the transition rate is based on the following assumption:

Assumption 4.1 *The transition rate function is a symmetric bounded function.*

The implication of the assumptions is that the distribution of the transition is symmetric. i.e. the transition rate function is symmetric at a certain range of times on Friday. Intuitively, the assumption is reasonable.

There is still a large range of symmetric concave functions which can be candidate functions to approximate the transition rate function. For numerical simplicity, we choose a m-lag moving average symmetric at the pole point as an approximate transition function $f_{(\tau,m)}(t)$. The first derivative of $f_{(\tau,m)}(t)$ is a concave function with its minimum value at the pole point τ. The number of lags of the moving average controls the concaveness of $\frac{df_{(\tau,m)}(t)}{dt}$ which describes the transition rate. Now, we have a transition function family $\{f_{(\tau,m)}(t)\}$ with two parameters $(\tau, m) \in \mathcal{R}^2$ where τ is the pole point and m is the length or the lags of the symmetric moving average. We, then, have a new series $\gamma_{(\tau,m)}(t)$ as follows

$$\gamma_{(\tau,m)}(t) = x(t_{fri}) - f_{(\tau,m)}(t)x(t_{thu}) - (1 - f_{(\tau,m)}(t))x(t_{sat}) \tag{4.62}$$

where $f_{(\tau,m)}(t)$ is an m-lag moving average of $f_{(\tau,1)}(t)$, $m = 3, 5, 7, \cdots$ and

$$f_{(\tau,1)}(t) = \begin{cases} 1 & t < \tau \\ 0 & t \geq \tau \end{cases}$$

For the desired transition function $f_{(\tau,m)}(t)$, $\{\gamma_{(\tau,m)}(t)\}$ should be normally distributed. Therefore, a statistic is constructed as follows

$$t_{(\tau,m)} = \frac{\bar{\gamma}_{(\tau,m)}}{S^2_{(\tau,m)}} \tag{4.63}$$

where

$$\bar{\gamma}_{(\tau,m)} = \frac{1}{48}\sum_{t=1}^{48} \gamma_{(\tau,m)}(t)$$

$$S^2_{(\tau,m)} = \frac{1}{48}\sum_{t=1}^{48} (\gamma_{(\tau,m)}(t) - \bar{\gamma}_{(\tau,m)})^2$$

$t_{(\tau,m)}$ is a t- distribution with 47 degrees of freedom. Hence, if a pair (τ_0, m_0) makes

$$|t_{(\tau_0,m_0)}| = \min_{\tau,m}\{|t_{(\tau,m)}|\} \tag{4.64}$$

the $f_{(\tau_0,m_0)}(t)$ is said to be the optimal transition function among the function family $\{f_{(\tau_0,m_0)}(t)\}$. A two step procedure is developed to locate the pair (τ_0, m_0) as follows.

In step 1, a new series is generated as

$$\gamma_{(\tau,1)}(t) = x(t_{fri}) - f_{(\tau,1)}(t)x(t_{thu}) - (1 - f_{(\tau,1)}(t))x(t_{sat}) \tag{4.65}$$

The pole point can be located by choosing τ_0 which has the minimum value among the $\{\gamma^2_{(\tau,1)}\}$, i.e.

$$\gamma^2_{(\tau_0,1)} = \min_{\tau}\{\gamma^2_{(\tau,1)}\}$$

In step 2, another series is generated as

$$\gamma_{(\tau_0,m)}(t) = x(t_{fri}) - f_{(\tau_0,m)}(t)x(t_{thu}) - (1 - f_{(\tau_0,m)}(t))x(t_{sat}) \tag{4.66}$$

and an associated statistic $t_{(\tau_0,m)}$. The lag of a moving average for $f_{(\tau_0,1)}$, m, is determined by m_0 which has the minimum absolute value among the $\{|t_{(\tau_0,m)}|\}$, i.e.

$$|t_{(\tau_0,m_0)}| = \min_{m}\{|t_{(\tau_0,m)}|\}$$

The chosen transition function $\gamma_{(\tau_0,m_0)}$ should be the most appropriate function in this family of functions to approximate the true transition. It is noted that each function has two parameters, and the chosen transition function from this family may not be smooth enough to reflect the transition process. To overcome the above two disadvantages, another family of functions is recommended.

We now consider a "filter" family of functions as follows:

$$f_{(d)}(t) = \begin{cases} g_{1\sim 0}(h_{(d)}(48-t)) & \text{for Friday} \\ g_{1\sim 0}(\max\{h_{(d)}(t)\} - h_{(d)}(t)) & \text{for Monday} \end{cases} \tag{4.67}$$

where $d \in \mathcal{I}$, $\mathcal{I}$ is an positive integer number set.

$$h_{(d)}(t) = \frac{d}{d+1} - \frac{1}{d+1}\sum_{j=1}^{d}(1 + \cos\frac{\pi j}{d+1})\cos\frac{\pi j t}{48}$$

$$g_{1\sim 0}(h(t)) = \frac{h(t) - \min(h(t))}{max(h(t)) - \min(h(t))}$$

According to the above definition, the family $\{f_{(d)}(t)|d \in \mathcal{I}\}$ has the properties below:

1. Each function in the family is monotone decreasing function ranging from 0 to 1.

2. $f_{(d)}(t) < f_{(d+1)}(t)$ for the Friday case;
   $f_{(d)}(t) > f_{(d+1)}(t)$ for the Monday case.

3. $\max(\frac{df_{(d)}(t)}{dt}) < \max(\frac{df_{(d+1)(t)}}{dt})$, and $t^*_{(d)} < t^*_{(d+1)}$ for the Friday case;
   $\max(\frac{df_{(d)}(t)}{dt}) < \max(\frac{df_{(d+1)}(t)}{dt})$, and $t^*_{(d)} > t^*_{(d+1)}$ for the Monday case;
   where $t^*_{(d)}$ is determined by $\frac{df_{(d)}(t)}{dt}|_{t=t^*} = \max(\frac{df_{(d)}(t)}{dt})$

The innovation series is generated as follows:

$$\gamma_{f_{(d)}} = x(t_{fri}) - f_{(d)}(t)x(t_{thu}) - (1 - f_{(d)}(t))x(t_{sat}) \tag{4.68}$$

Under the criterion of the minimum sum of squared errors, we can locate the most suitable function in this family of functions to approximate the transition process by searching d_0 which makes

$$\gamma^2_{f_{(d_0)}} = \min_d\{\gamma^2_{f_{(d)}}\}$$

Discussion:

If we question the normality assumption of the generated series $\{\gamma_{(\tau,m)}(t)\}$ from equation (4.63), the constructed t- statistic $t_{(\tau,m)}$ may not be a proper statistic to locate the pair of parameters (τ, m) because the distribution of $\{\gamma_{(\tau,m)}(t)\}$ is unknown in general circumstances. Any pre-specified assumption for the distribution may misspecify the estimation of the transition function. Also, it is well known that the criterion of sum of squared errors is not a robust criterion. Selection procedures without any special assumption concerning the distribution (i.e. a nonparametric selection procedure, and a more robust criterion such as absolute sum of squares, etc.) can be considered. Because only the selection criteria would be different, the nonparametric and robust selection procedure will not be discussed here.

In summary, the approaches mentioned above are designed to search for the optimal proposed transition functions from a restricted family of functions. Although

some properties of the transition function are known by analysis, there is a broad range of parametrized families which can approximate the transition process. Since there is not enough information about the detailed properties of the transition function, some assumptions have had to be made to narrow the range of the candidates for the parametrized family of functions, to reduce the size of the parameter set, and to simplify expression of the function for simplicity and numerical reasons.

However, there is no answer to the question, "which transition function family is the best ?". For this reason, another mechanism is proposed to find the transition process instead of searching for an optimal one in a parametrized family of functions.

4.5.2 Principal Components Approach

The evolution from weekdays to weekend days by way of a Friday can be in general described by a weight function $f(t_{fri})$ (see equation (4.1)). However, $f(t_{fri})$ is an unknown function. In the last section, two procedures were proposed to search for an optimal transition function in a parameterized family of functions. The optimal transition function is regarded as an approximation to the real transition function. Now, we use the linear combination of a set of step functions to approximate the real transition function.

Suppose, a linear space $\mathcal{Z}$ is expanded on the basis of a family of functions $S_t^{(d)}$, where

$$S_t^{(d)} = \begin{cases} 1 & \text{if } t < d \\ 0 & \text{if } t \geq d \end{cases}$$

From the theory of real function analysis, we know that the real function space $\mathcal{F}$ belongs in the closure of $\mathcal{Z}$, say, $\bar{\mathcal{Z}}$. i.e. $\mathcal{F} \subseteq \bar{\mathcal{Z}}$. Therefore, the weight function can be approximately expressed by a linear combination of $Sp\{S_t^{(d)} | d = 1, 2, 3, \cdots, 48\}$.

We suppose

$$\begin{aligned} X_{thu} &= (x_{thu}(1), x_{thu}(2), \cdots, x_{thu}(48))' \\ X_{fri} &= (x_{fri}(1), x_{fri}(2), \cdots, x_{fri}(48))' \\ X_{sat} &= (x_{sat}(1), x_{sat}(2), \cdots, x_{sat}(48))' \end{aligned}$$

$$X_{i,j} = S_t^{(j)} x_{thu}(t) + (1 - S_t^{(j)}) x_{sat}(t)$$

Hence,

$$X_{\cdot,j} = (X_{1,j}, X_{2,j}, \cdots, X_{48,j}) = \begin{pmatrix} I_j & 0 \\ 0 & 0 \end{pmatrix} X_{thu} + \begin{pmatrix} 0 & 0 \\ 0 & I_{48-j} \end{pmatrix} X_{sat} \tag{4.69}$$

To find the weight function, we must estimate the vector β in the following linear model

$$X_{fri} = (X_{\cdot,1}, X_{\cdot,2}, \cdots, X_{\cdot,48})\beta + \epsilon \tag{4.70}$$

where $\beta = (\beta_1, \beta_2, \beta_3, \cdots, \beta_{48})'$

Because

$$\begin{aligned} X_{fri} &= \sum_{j=1}^{48} \beta_j X_{\cdot,j} + \epsilon \\ &= \sum_{j=1}^{48} \beta_j \begin{pmatrix} I_j & 0 \\ 0 & 0 \end{pmatrix} X_{thu} + \sum_{j=1}^{48} \beta_j \begin{pmatrix} 0 & 0 \\ 0 & I_{48-j} \end{pmatrix} X_{sat} + \epsilon \end{aligned}$$

The weight function $f(t)$ at time point i is $f(j) = \sum_{j=i}^{48} \beta_i$ from the first part of the above equation; $(1 - f(i)) = \sum_{j=1}^{i-1} \beta_j$ from the second part of the above equation. According to the model equation (4.1), $f(1) = 1$. Therefore, there is a restriction on β, i.e. $\sum_{i=1}^{48} \beta_i = 1$.

Our problem becomes to solve a restricted linear regression as follows

$$\begin{cases} X_{fri} = X\beta + \epsilon \\ A\beta = a \end{cases} \tag{4.71}$$

where

$$A = \begin{pmatrix} 1 & \cdots & 1 & 1 \\ 0 & \cdots & 0 & 1 \end{pmatrix}; \quad a = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

After estimating β, the estimated weight function can be constructed by

$$\hat{f}(j) = \sum_{i=j}^{48} \hat{\beta}_i, \quad \text{for } 1 \leq i \leq 48$$

Considering the construction of X and the evolving nature of a Friday's (or a Monday's) load profile, we know that X must be at least a near multicollinear matrix.

Dhrymes (1978) has discussed several methods and their properties for handling the near multicollinear case, such as, general inversion methods, dropping multicollinear columns, and the principle components method. Nevertheless, there is a practical sense in which the principal components version is more desirable because we would have an "estimate" for the coefficients of all relevant variables. However, the other two methods generally have an estimate for only a subset of such coefficients. Hence, significant aspects of such implication may be obscured or escape our notice. Therefore, we will use the principal components method to estimate the transition weight function for Friday and Monday.

4.5.3 Application

To illustrate the evolution process between the weekdays and weekend days, we use the above three approaches on the half-hourly data from April 4, 1983 to May 1, 1983 (four weeks) and estimate the transition functions presented in Figure 4.5. The thin curves are transition functions from weekdays to weekend days (Friday); and the wide curves are transition functions from weekend days to weekdays (Monday). The estimated transition functions consistently indicate that the transition from weekday patterns to weekend's patterns are taking place around or after 3PM Friday; and the weekend days' load profile patterns have vanished after 8AM Monday. It is noted that the "moving average" and "filter" function approaches show greater lack of flexibility as compared to the principle components approach.

As we expected the sum of squared fitting errors (see equation (4.1)) from the principle components approach is the smallest among the three transition function estimation approaches but includes more computation. The "filter" function estimation approach is the most efficient in the sense of least computation. Based on accuracy, the principle component approach is used in our modelling procedure. The total sample fitting and post sample predictions are compared with other popular modelling techniques in Chapter 5.

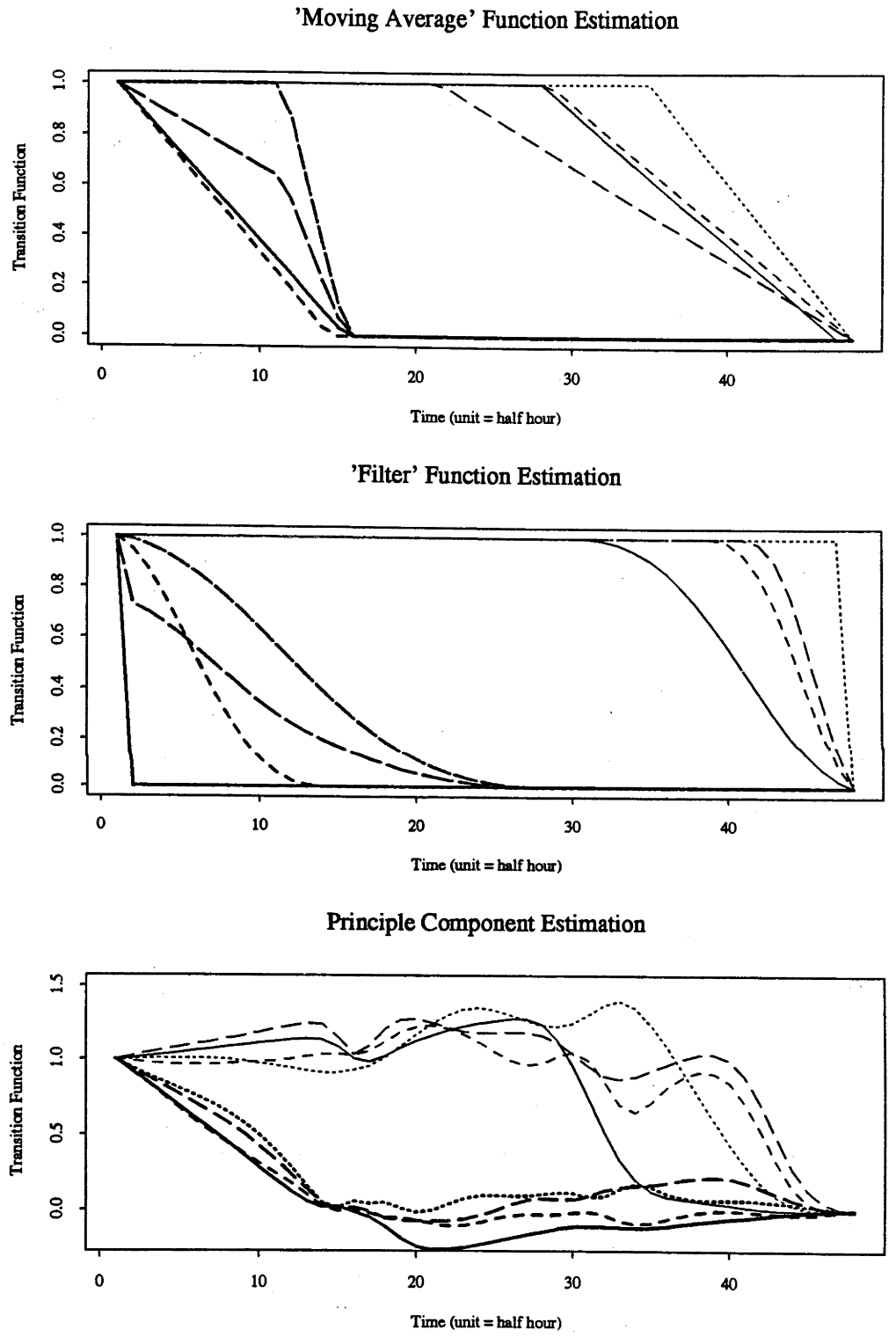


Figure 4.5: The Estimated Transition Functions

4.6 Summary

In this chapter, a new model for the half-hourly electricity load data has been established. A more general cointegration "error correction" regression model has been proposed to model and to forecast trend behaviour. The detrended data is a periodically stationary time series which can be modelled by the proposed model described in section 4.4 for data in weekdays and weekend days separately. The load evolution process between weekdays and weekends is portrayed by a weight function which can be estimated by three approaches described in section 4.5. The remaining innovation series is a stationary series and can be fitted and forecast by a subset AR model proposed in chapter 2.

After applying the proposed new model to New Zealand half-hourly electricity load data and utilizing the associated modelling procedures for different components, it shows that the overall performance is much better than Moutter's approach both in sample fitting and post sample forecasting. The proposed modelling procedure have also been successfully applied to quarter-hourly electricity load data in the Canberra region of Australia.

Our experiences show that the proposed approach is superior to Moutter et al. (1986b) and Bodger et al. (1987) approaches. To compress the size of this thesis, the results are omitted here but they can be obtained from the author on request. The comparison with other popular modelling techniques for New Zealand half-hourly electricity load data is presented in chapter 5. The results show that the modelling procedure proposed in this chapter is very promising with a potential to model short-term electricity load from any region.

Chapter 5

On Additive Deterministic & Adaptive Models

5.1 Introduction

In the literature on modelling a non-stationary periodic time series $\{x(t)\}$, where the time series $\{x(t)\}$ consists of trend, seasonal, and disturbance components in additive form[1], i.e.

$$x(t) = f(t) + S(t) + y(t) \tag{5.1}$$

and where $f(t)$, $S(t)$ and $y(t)$ are the trend, seasonal, and disturbance components, respectively, there are two basic additive models namely:

1. Conventional Additive Model

2. Adaptive Additive Model

The conventional additive model assumes that the trend and seasonal components are deterministic components. While, for the adaptive additive model, there is no such deterministic component restrictions on trend, and seasonal components, and, at least, one of them is stochastic. Therefore, the adaptive additive model has

[1]One can always transfer multiplicative form to additive form by adding a constant, C, which is large enough to make $\{x(t) + C\}$ a positive series, and then taking logarithms if $\{x(t)\}$ is in multiplicative form.

three forms. i.e. (1) Both trend and seasonal components are stochastic; (2) Trend component is stochastic while seasonal component is deterministic; (3) Trend component is deterministic and seasonal component is stochastic. In most cases, the adaptive additive model is more flexible and realistic than the conventional additive model.

In the next section of this chapter, we will discuss the relation between the conventional additive model and a special form of the adaptive additive model, and reveal that the conventional additive model in some circumstances may be more accurate in the sense of smaller residual variance, and more reliable in the sense of having a narrower forecasting confidence interval than the adaptive additive models. In section 5.3, a more complicated multiplicative model, AR(ARIMA)MA, and its special form subset ARAR, which can be "automatically" specified for a given data set by a designed model selection procedure, is presented. In section 5.4, we apply five models (three adaptive additive models and two conventional models) to the New Zealand half-hourly electric load data, and compare their performance in fitting and forecasting.

5.2 Multiplicative Model *vs* Additive Model

We know that the filter $(1-B)$ will remove an order 1 polynomial trend and level, and the filter $(1-B^s)$ will remove a periodic , or "seasonal", component with period s. Consequently, if $x(t)$ contains such trend, and seasonal components, $z(t) = (1-B)(1-B^s)x(t)$ will have a "detrended" and "deseasonalized" form and zero mean, and thus it is reasonable to assume that $\{z(t)\}$ may be represented by an ARMA model. To allow for the possibility of non-stationarity in both the $\{x(t)\}$ model and the seasonal component, Box and Jenkins (1976) generalized the seasonal model to

$$\theta_1(B)\theta_2(B^s)\Delta^d\Delta_s^D x(t) = \phi_1(B)\phi_2(B^s)\epsilon(t) \tag{5.2}$$

Where $\theta_1(z), \theta_2(z), \phi_1(z), \phi_2(z)$ are polynomials of degrees, p, P, q, Q, respectively, (having all their roots outside the unit circle), $\Delta = (1 - B)$ is a one step difference operator, i.e. $\Delta x(t) = x(t) - x(t-1)$, and $\Delta_s = (1 - B^s)$ is the $s-$ step difference operator, i.e. $\Delta_s x(t) = x(t) - x(t-s)$. Box and Jenkins refer to the model (5.2) as the *multiplicative seasonal model of order* $(p, d, q) \times (P, D, Q)$

Again, once the values of d and D have been determined, and suitable integers specified for the orders, p, P, q, Q, the further parameters of the model may be estimated by fitting the model

$$\theta_1(B)\theta_2(B^s)z(t) = \phi_1(B)\phi_2(B^s)\epsilon(t) \tag{5.3}$$

to $z(t) = \Delta^d \Delta_s^D x(t)$.

Box and Jenkins derive the model (5.2) by arguing that the seasonal component of $x(t)$ my be modelled by

$$\theta_2(B^s)\Delta_s^D x(t) = \phi_2(B^s)e(t) \tag{5.4}$$

while the "non-seasonal" component may be modelled by assuming that the "residuals" from the above model, $e(t)$, satisfy

$$\theta_1(B)\Delta^d e(t) = \phi_1(B)\epsilon(t) \tag{5.5}$$

Substituting equation (5.5) into equation (5.4) then gives the multiplicative seasonal model (5.2). Box and Jenkins maintain that model (5.2) is a fairly general model for (non-stationary) series which contain a seasonal component of period s, and it should be noted that model (5.2) implies rather more than the conventional additive model

$$x(t) = f(t) + S(t) + y(t) \tag{5.6}$$

where $f(t)$ denotes a deterministic polynomial of degree $(d-1)$ (representing the "trend"), $S(t)$ is a periodic component, (period s), and $y(t)$ is a stationary stochastic process, with $\mathbf{E}y(t) = 0$.

Underlying Additive Model for A Multiplicative Seasonal Model

A question must arise when we ask what is the underlying additive model for a multiplicative seasonal ARIMA model, and how "safe" is it to use a multiplicative seasonal ARIMA model to fit a non-stationary time series, and what is lost, if anything, in this approach. For simplicity, we take the "airline" model, ARIMA$(0,1,1)\times(0,1,1)_s$, as an example to illustrate the underlying additive model. The "airline" model has the following form

$$(1-B)(1-B^s)x(t) = (1-\phi_1 B)(1-\phi_2 B^s)\epsilon(t) \tag{5.7}$$

where the absolute values of ϕ_1, ϕ_2 are less than 1, and $\epsilon(t)$ is a white noise process with zero mean and variance σ_ϵ^2.

It is easy to construct an *adaptive additive model* which is equivalent to model (5.7) as follows

$$x(t) = f(t) + S(t) + \epsilon_1(t) \tag{5.8}$$

where $f(t)$ is a degree 1 polynomial stochastic trend function; $S(t)$ is a stochastic periodic function with period s; and $\epsilon_1(t)$ is a white noise process with zero mean and variance $\sigma_{\epsilon_1}^2$. $f(t)$ and $S(t)$ satisfy the following models respectively,

$$(1-B)^2 f(t) = \nu^{(1,\nu_1,\cdots,\nu_m)}(B)\epsilon_2(t) \tag{5.9}$$

$$(1+B+\cdots+B^{s-1})S(t) = \eta^{(1,\eta_1,\cdots,\eta_n)}(B^s)\epsilon_3(t) \tag{5.10}$$

where $\nu^{(1,\nu_1,\cdots,\nu_m)}(z) = 1+\nu_1 z+\cdots+\nu_m z^m$; $\eta^{(1,\eta_1,\cdots,\eta_n)}(z) = 1+\eta_1 z+\cdots+\eta_m z^m$; $\epsilon_i(t)$ is white noise process with zero mean and variance $\sigma_{\epsilon_i}^2$, $i=1,2,3$; and $\epsilon_1(t)$, $\epsilon_2(t)$, $\epsilon_3(t)$ are uncorrelated. The use of model (5.9) suggests that one should allow for changing trend slopes. Similarly, the seasonal model (5.10) allows for changing amplitude and phase. Box et al. (1987) show that the "airline" model yields the unobserved trend and seasonal forecasts given below

$$f_t(l) = \alpha_0^{(t)} + \alpha^{(t)} l$$

$$S_t(l) = \alpha_p^{(t)}, \quad \sum_{p=1}^{s} \alpha_p^{(t)} = 0$$

and the forecasting function is as follows

$$x_t(l) = f_t(l) + S_t(l), \quad l > 0 \tag{5.11}$$

They also show that the optimal component forecasts are independent of the admissible decompositions (5.9) - (5.10) generating these components. Within the class of decomposition (5.9) - (5.10) that are consistent with the overall adaptive additive model (5.8), the canonical decomposition minimizes the MSE of the forecasts of the trend and seasonal components, and the MSE of the irregular component forecasts is maximized in the canonical decomposition.

The changes in trend and seasonal components are achieved by updating the parameters, $\alpha_0^{(t)}$, $\alpha^{(t)}$ and $\alpha_p^{(t)}$ of the trend and seasonal components using the one step ahead forecast errors $\epsilon_t(1)$ for the overall model (5.7)

$$x_{t+1}(l) = x_t(l+1) + \psi_l \epsilon_{t+1} \tag{5.12}$$

where ϵ_{t+1} is the error made in forecasting x_{t+1} at time t. It is easy to show that the overall update parameter ψ_l satisfies (see pp. 310-311, in Box and Jenkins (1976) for details)

$$\psi_l = \psi_p = \lambda_1(1 + p\lambda_2) + \delta_s\lambda_2$$

where $p = \text{mod}(l, s)$, $\lambda_2 = 1 - \phi_1$, $\lambda_2 = 1 - \phi_2$ and $\delta_s = 1$ if $p = s$ and 0 otherwise.

The adaptive parameters of the trend and seasonal forecasting component satisfy

$$\begin{cases} \alpha^{(t+1)} = \alpha^{(t)} + \frac{\lambda_1\lambda_2}{s}\epsilon_{t+1} \\ \alpha_0^{(t+1)} = \alpha_0^{(t)} + \alpha^{(t)} + \left[\lambda_1(1 - \frac{s+1}{2s}\lambda_2) + \frac{\lambda_2}{s}\right]\epsilon_{t+1} \\ \alpha_p^{(t+1)} = \alpha_{p+1}^{(t)} + \left[\frac{s+1-2p}{2s}\lambda_1\lambda_2 - (1 - s\delta_s)\frac{\lambda}{s}\right]\epsilon_{t+1} \end{cases} \tag{5.13}$$

It is noted that the forecasted series level $\alpha_0^{(t)}$ is adjusted by the forecast increment to the trend, and incorporated into the next level $\alpha_0^{(t+1)}$ at the new origin.

It also can be seen that the overall "airline" model will be very sensitive to outliers and interventions; the adaptation of trend and seasonal parameters is heavily

dependent on the one step ahead forecast error which will be large if an outlier or intervention occurs. i.e. the one step error caused by an outlier or intervention is shared among the trend, seasonal and disturbance components, and the distribution to each component is determined by the parameters of the "airline" model. Consequently, outliers or interventions can lead to mis-adaptation of the trend and seasonal components, and then multiple step ahead forecasts could diverge far from the true value.

In general, a multiplicative seasonal model may be suitable for a forecasting a few steps ahead. However, for a large number of steps ahead forecasts made by a multiplicative seasonal model may not be reliable if the quality of a data set is poor (i.e. outliers and interventions are present).

Supposing the disturbance terms of trend and seasonal components of the adaptive additive model (5.8) are known, we examine how the model allows the parameters to adapt in the trend and seasonal components. As with the update equation (5.12), the trend and seasonal update equations have the following form:

$$f_{t+1}(l) = f_t(l+1) + \psi_l \epsilon_2(t+1) \tag{5.14}$$

$$S_{t+1}(l) = S_t(l+1) + \varphi_l \epsilon_3(t+1) \tag{5.15}$$

where ψ_l and φ_l satisfy

$$\sum_{i=0}^{+\infty} \psi_i B^i = \frac{\nu^{(1,\nu_1,\cdots,\nu_m)}(B)}{(1-B)^2} = \nu^{(1,\nu_1,\cdots,\nu_m)(B)} \sum_{j=0}^{+\infty} (j+1) B^j$$

$$\sum_{i=0}^{+\infty} \varphi_i B^i = \frac{(1-B)}{(1-B^s)} \eta^{(1,\eta_1,\cdots,\eta_n)}(B^s) = (1-B)\eta^{(1,\eta_1,\cdots,\eta_n)}(B^s) \sum_{j=0}^{+\infty} B^{sj}$$

It is easy to verify that when one step ahead new information becomes available, the update equations for the trend and seasonal components satisfy

$$\begin{cases} \alpha^{(t+1)}(l) = \alpha^{(t)}(l+1) + (\psi_{l+1} - \psi_l)\epsilon_2(t+1) \\ \alpha_0^{(t+1)}(l) = \alpha_0^{(t)}(l+1) + \alpha^{(t)}(l+1)[(l+1)\psi_l - l\psi_{l+1}]\epsilon_2(t+1) \\ \alpha_p^{(t+1)}(l) = \alpha_p^{(t)}(l+1) + \varphi_l \epsilon_3(t+1) \end{cases} \tag{5.16}$$

Consequently, the one ahead step adaptation for the parameters of trend and seasonal components is as follows

$$\begin{cases} \alpha^{(t+1)} = \alpha^{(t)} + \epsilon_2(t+1) \\ \alpha_0^{(t+1)}(l) = \alpha_0^{(t)}(l+1) + \alpha^{(t)}(l+1) + (1-\nu_1)\epsilon_2(t+1) \\ \alpha_p^{(t+1)}(l) = \alpha_p^{(t)}(l+1) + \epsilon_3(t+1) \end{cases} \tag{5.17}$$

From model (5.7) and (5.8), (5.9), (5.10) we have the relation,

$$\begin{aligned} (1-\phi_1 B)(1-\phi_2 B^s)\epsilon(t) &= \frac{(1-B^s)\nu^{(1,\nu_1,\cdots,\nu_m)}(B)}{(1-B)}\epsilon_2(t) \\ &\quad +(1-B)\eta^{(1,\eta_1,\cdots,\eta_n)}(B^s)\epsilon_3(t) \end{aligned} \tag{5.18}$$

and so the error variance of the overall "airline" model, σ_ϵ^2, is a weighted sum of the error variances of trend $\sigma_{\epsilon_2}^2$, seasonal $\sigma_{\epsilon_3}^2$, and $\sigma_{\epsilon_1}^2$

$$\sigma_\epsilon^2 = \frac{\nu^{(1,\nu_1^2,\cdots,\nu_m^2)}(1)}{(1+\phi_1^2)(1+\phi_2^2)}\sigma_{\epsilon_2}^2 + \frac{\eta^{(1,\eta_1^2,\cdots,\eta_n^2)}(1)}{(1+\phi_1^2)(1+\phi_2^2)}\sigma_{\epsilon_3}^2 + \frac{4}{(1+\phi_1^2)(1+\phi_2^2)}\sigma_{\epsilon_1}^2 \tag{5.19}$$

When we consider that the adaptation for the trend and seasonal components of the overall "airline" model (5.13) is generated by one source of disturbance, ϵ, according to (5.13). A "large" size disturbance caused by an outlier is shared by both trend and seasonal components, and may cause mis-adaptation. However, the adaptation of the trend and seasonal components of the adaptive additive model in (5.17) is determined by the disturbance terms of the trend and seasonal components, $\epsilon_2(t)$ and $\epsilon_3(t)$, whether their variances are known *a prior* or predetermined. The mis-adaptation for trend and seasonal components can be controlled if the overall one step ahead forecast error is caused by an outlier or intervention. Therefore, the adaptive additive model is a partial solution to overcome the mis-adaptation due to the presence of outliers and interventions.

One method is to restrict the variances of trend and seasonal components to a certain level to prevent the trend and seasonal components being over-adapted. This method will be discussed in a state-space form in a latter chapter. The second method is to abolish the adaptation for the trend and seasonal components if we can be convinced that the changes in the trend and seasonal components are too small

to be taken into account. Then, the trend and seasonal are regarded as deterministic components. The trend and seasonal components can be estimated by regression of $x(t)$ on a polynomial function $f(t) = \alpha_0 + \alpha t$ and a series of trigonometric functions which have harmonic frequencies related to the period s. This regression procedure is equivalent to setting $\epsilon_2(t) = 0$, and $\epsilon_3(t) = 0$ for all t in model (5.9) and (5.10), consequently, $\sigma^2_{\epsilon_2} = \sigma^2_{\epsilon_3} = 0$. Therefore, from equation (5.19), the error variance of the overall "airline model", σ^2_ϵ, satisfies

$$\sigma^2_\epsilon = \frac{4}{(1+\phi_1^2)(1+\phi_2^2)}\sigma^2_{\epsilon_1} > \sigma^2_{\epsilon_1} \tag{5.20}$$

The above equation illustrates that the overall "airline" model's performance is worse than the conventional additive regression model if the trend and seasonal are really deterministic components. In this case, The more general random assumption for the trend and seasonal components and the safety associated with use of the multiplicative seasonal ARIMA is at the cost of an increase in residual variance (see details in the section, Practical Aspects, on page 141). In general, the trend and seasonal may not be deterministic components. If the ratio values of $\sigma^2_{\epsilon_2}/\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_3}/\sigma^2_{\epsilon_1}$ are large, i.e. the trend and seasonal components are far from being deterministic, and their changes can be distinguished from the disturbance noise $\epsilon_1(t)$, the overall "airline" model may fit better than the conventional additive regression model since the "airline" model has adaptive trend and seasonal changes. On the other hand, if the ratio values of $\sigma^2_{\epsilon_2}/\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_3}/\sigma^2_{\epsilon_1}$ are small, i.e. the trend and seasonal components are nearly deterministic and their changes(if any) are submerged by the disturbance component $\epsilon_1(t)$, the conventional additive regression model may fit considerably better than "airline" model because the changes in the trend and seasonal components are insignificant compared to the disturbance noise $\epsilon_1(t)$.

5.2.1 Multiplicative Seasonal Model *vs* Conventional Additive Model

CASE 1: A Multiplicative Seasonal Model Fits A Time Series Generated by A Conventional Additive Model

If the model really is of the form (5.6) and we fit a model of the form (5.2), then this implies that $y(t)$ is a *pathological* process, in the sense that $y(t)$ itself has a "non-stationary periodic" structure. To see this, we apply the operator $\Delta^d\Delta_s^D$ to both sides of the conventional additive model (5.6), and noting that Δ^d "annihilates" $f(t)$ and Δ_s^D "annihilates" $S(t)$, we obtain,

$$\Delta^d\Delta_s^D x(t) = \Delta^d\Delta_s^D y(t)$$

Consequently, $y(t)$ also satisfies (5.2), i.e.

$$\theta_1(B)\theta_2(B^s)\Delta^d\Delta_s^D y(t) = \phi_1(B)\phi_2(B^s)\epsilon(t)$$

and the presence of the "singular" factor $(1-B^s)$, $(1-B)^d$(both of which have roots on the unit circle) mean that (a) $y(t)$ is non-stationary (if $d > 0$), and (b) it has a form of periodic structure, in the sense that its "spectral density function" is given by

$$h_y(\omega) = \frac{\sigma_\epsilon^2}{2\pi} \mid \frac{\phi_1(e^{-i\omega})\phi_2(e^{-is\omega})}{\theta_1(e^{-i\omega})\theta_2(e^{-is\omega})} \mid^2 \mid \frac{1}{(e^{-i\omega})^d(e^{-is\omega})^D} \mid^2$$

and thus has infinite "peaks" at frequencies $\omega = 2\pi k/s$, $k = 0, 1, \cdots, [s/2]$ unless, of course, $\phi_1(B)$, $\phi_2(B^s)$ contain as factors $(1-B^s)$, $(1-B)^d$. Thus, for $h_y(\omega)$ to be bounded; it is necessary for $\phi_1(B)$, $\phi_2(B^s)$ to be of the from $\phi_1(B) = (1-B)^d\tilde{\phi}_1(B)$, $\phi_2(B) = (1-B^s)^D\tilde{\phi}_2(B)$ where $\tilde{\phi}_1(L)$, $\tilde{\phi}_2(L)$ have their roots outside unit the unit circle.

In this case where $h_y(\omega)$ is bounded, we have for the solution of model (5.2),

$$x(t) = f(t) + S(t) + d(t) + y(t) \tag{5.21}$$

where $d(t)$ is the (decaying) solution of $\theta_1(B)\theta_2(B^s)x(t) = 0$, and $y(t)$ (the particular solution of (5.2)) is the stationary process,

$$y(t) = \theta_1^{-1}(B)\theta_2^{-1}(B^s)\tilde{\phi}_1(B)\tilde{\phi}_2(B^s)\epsilon(t)$$

However, if $\phi_1(B)$, $\phi_2(B^s)$ do not contain factors of the above form then $y(t)$ is non-stationary and has itself a periodic form.

Abraham and Box (1978) distinguish the above case by saying that when $\phi_1(B)$, $\phi_2(B^s)$ contain $(1-B)^d$, $(1-B^s)^D$ as factors then the seasonal component has a strictly periodic stable "deterministic" form (i.e. it has a Fourier series representation with constant amplitudes and phases), and the trend has a stable "deterministic" polynomial form, with constant coefficients. On the other hand, when $\phi_1(B)$, $\phi_2(B^s)$ do not contain these factors the seasonal component has an "adaptive" form (with the amplitudes and phases possibly changing over time) and similarly the trend has an "adaptive" polynomial form. Abraham and Box (1978) therefore argue that it is always safe to fit a model of the form (5.2) to non-stationary seasonal data; if the trend and seasonal components are both stable, this will be revealed by a near cancellation of the operators $(1-B)^d$, $(1-B^s)^D$ on both sides of the fitted model.

Another tool to help one judge whether to use deterministic or adaptive seasonal components in modelling a periodic data set is spectral analysis which may prove useful; a sharp peak at the seasonal frequency and its harmonics corresponding to a deterministic component and a broader peak to an adaptive component. However, it is not easy to judge which model is better from the spectrum.

CASE 2: Conventional Additive Model Fits A Time Series Generated by A Multiplicative Seasonal Model

However, if the model really is an "airline" type model, what will happen when we fit a model of the conventional additive form (5.6)? By noting that the solution of the homogeneous difference functions for $f(t)$ and $S(t)$ in (5.9) and (5.10) are equal to the deterministic trend and seasonal function $f(t)$ and $S(t)$ of the conventional additive form (5.6), the stochastic component $y(t)$ of model (5.6) satisfies the following model

$$(1-B)(1-B^s)y(t) = (1-\phi_1 B)(1-\phi_2 B^s)\epsilon(t) \tag{5.22}$$

The above model is exactly the same as the model for $x(t)$. It is expected, therefore, that the auto-correlation properties of $y(t)$ should be exactly the same as the auto-correlation properties of $x(t)$. Hence, although the sample variance of the innovation residual from the conventional additive model, $\widehat{\mathbf{var}}(y(t))$, is less than the corresponding sample variance of $x(t)$, say, $\widehat{\mathbf{var}}(x(t))$, the innovation residual $y(t)$ is still a non-stationary series which has the same stochastic structure as $x(t)$. We can conclude that if $x(t)$ is really a multiplicative seasonal ARIMA model, the innovation residual series, which is obtained by subtracting the "deterministic" trend and seasonal components from $x(t)$, is not stationary (actually the innovation residual series is also the same multiplicative seasonal ARIMA model as $x(t)$) unless the moving average side contains detrend and deseasonal factors which cancel with the detrend and deseasonal factors on the auto-regressive side.

Practical Aspect

From the above analysis, it can be seen that the multiplicative seasonal ARIMA model can still model a time series, which is generated by the conventional additive model (5.6), reasonably well but with an increase in residual variance from residuals which are nearly white noise. However, if a data set is generated by a multiplicative seasonal ARIMA model, the conventional additive model cannot fit the data set properly since the innovation or residual series is still of the same multiplicative seasonal ARIMA type as the original data. This is the reason why Abraham and Box (1978) claim that use of a multiplicative seasonal ARIMA model to fit seasonal non-stationary data is safer than the conventional additive model.

In practical cases it is very hard to know whether the underlying trend and seasonal component are nearly deterministic or totally adaptive. As we mentioned above, if the ratio of the error variance of trend to disturbance variance, $\sigma^2_{\epsilon_2}/\sigma^2_{\epsilon_1}$, and the ratio of error variance of seasonal component to disturbance variance, $\sigma^2_{\epsilon_3}/\sigma^2_{\epsilon_1}$, are both very small, or equivalently, if there are approximate detrend and deseasonal factors

as part of the moving average side of a fitted multiplicative seasonal ARIMA model which can nearly cancel detrend and deseasonal factors on the auto-regressive side, then the use of the conventional additive model may benefit from a smaller residual variance. However, there is no simple answer to how small those ratios should be and how close the moving average factors should be to the detrend and deseasonal factor on the auto-regressive side of any fitted multiplicative seasonal ARIMA model since this can vary from one model to another.

Therefore, the approximate detrend and deseasonal factor on the moving average average side of a fitted multiplicative seasonal model or sharp peaks in spectrum just indicate the conventional additive model may fit the data set better. On the other hand, if the innovation residuals from fitting the data set to a conventional additive model shows a similar auto-correlation pattern to the auto-correlation function of the data set, this indicates a multiplicative seasonal model may be more appropriate for the data set.

So far we have only considered cases where the trend and seasonal component are either totally deterministic or totally adaptive. However, in practical situations, one could have a model which is adaptive for certain components and not in others or both trend and seasonal are "nearly" deterministic. For these cases, the innovation residual series from a conventional additive model is nearly stationary and may have damped trend and some periodicity. Therefore, the innovation residual series may be modelled by an auto-regressive model such as

$$\gamma(B)y(t) = \epsilon(t)$$

where $\gamma(z)$ is a polynomial function of z in which all roots lie outside the unit circle and $\epsilon_1(t)$ is white noise series with variance $\sigma^2_{\epsilon_1}$.

In fitting New Zealand half-hourly load data by a multiple ARIMA$(p,d,q) \times (P,D,Q)_s$ model, we experiment with the "airline model" ARIMA$(0,1,1) \times (0,1,1)_{48}$ to fit weekly differenced data, and find that the following model fits the differenced data reasonably well.

$$(1 - B^{48})(1 - B)y(t) = (1 - 0.9B^{48})(1 - 0.4B)\epsilon(t) \tag{5.23}$$

where $y(t) = x(t) - x(t-336)$ is the weekly difference of the original data, and the estimated variance of $\epsilon(t)$ is $\hat{\sigma}_\epsilon^2 \approx$ 2.0E-4. The coefficient of B^{48} on the moving average side is 0.9 which is close to 1, i.e. the daily difference factor on the autoregressive side. This fact indicates that the use of a conventional additive model on weekly differenced data may achieve a smaller residual variance.

For data which is s periodic, the seasonal operator takes the form $(1 - B^s)$; as observed by Abraham and Box, the operator $(1 - B^s)$ has roots $e^{2\pi ik/s}$, $k = 1, \cdots, s$, which are evenly distributed over the unit circle. The homogeneous difference function corresponding to this operator therefore has the form

$$A_0 + \sum_{k=1}^{[s/2]} (A_k \cos \frac{2\pi kt}{s} + B_k \sin \frac{2\pi kt}{s}) \tag{5.24}$$

where the factors of the factorization of $(1 - B^s)$ correspond to the terms in (5.24).

It can be seen from the fitted model (5.23) that the factor $(1 - 0.9B^{48})$ on the left side of the above model is close to the term which removes the daily periodic factor, i.e. $(1 - B^{48})$ on the right side. From the analysis by Abraham and Box (1978), this fact indicates that the daily periodic component is nearly "deterministic", and use of the conventional additive model (5.6) may be a suggested.

In addition, a periodic component can be fitted by the harmonic representation (5.24) if the periodic component is strictly periodic with period s. However, in a real data set, we can only observe the most apparent periodical pattern, and there may be other hidden periodic patterns which are not properly represented by the harmonic representation (5.24). For instance, the apparently periodicity of New Zealand half-hourly load data are weekly in period (336 points) and also daily in period (48 points) for weekdays and weekend days. Other periodic patterns (if any) are not so obvious. Furthermore, there may be only a number of the harmonics in the harmonic representation which are significant. By employing the MXFFT described in chapter 4, we can search for significant seasonal harmonics and other hidden periods(frequencies) to fit the periodic component as follows

$$A_0 + \sum_{\text{significant } \omega_i} (A_i \cos 2\pi\omega_i t + B_i \sin 2\pi\omega_i t) \tag{5.25}$$

Therefore, the harmonic representation (5.24) is a special case of (5.25).

On the other hand, in fitting New Zealand half-hourly load data directly to the multiple ARIMA$(p,d,q)\times(P,D,Q)_s$ model, we experiment with the "airline model" ARIMA$(0,1,1)\times(0,1,1)_{336}$ to fit 12 data sets. Each of the data sets has three weekly samples of half-hourly electric load, starting from the first Monday of each month for April 1983, to March 1984. It can be seen from the trial results listed in Table 5.1 that the some coefficients of B^{336} on the moving average side are quite close to 1(see models for June, July, and October, 1983 in Table 5.1) and others are not so close to 1. In this case, it is very hard to determine at present which model is better the adaptive additive model or the conventional additive model. In the next section, we will show that the model ARIMA$(0,1,1)\times(0,1,1)_{336}$ is not adequate to fit the data because it ignores the daily periodic behaviour which cannot be covered by removing weekly periodic effect. Then we will discuss a more complicated model which may handle the multiplicative seasonal behaviour better than the Box-Jenkins multiplicative seasonal models.

Estimated ARIMA$(0,1,1)\times(0,1,1)_{336}$ Models for Three Weekly Sample Sets

Data set	Model	$\hat{\sigma}^2$	AIC
Apr. 83	$z(t)=(1-0.4242B^{336})(1+0.3226B)\epsilon(t)$	1.7743E-4	-8702.03
May 83	$z(t)=(1-0.3888B^{336})(1+0.3051B)\epsilon(t)$	1.4826E-4	-8883.08
Jun. 83	$z(t)=(1-0.8870B^{336})(1+0.3547B)\epsilon(t)$	1.6463E-4	-8777.51
Jul. 83	$z(t)=(1-0.9862B^{336})(1+0.2476B)\epsilon(t)$	9.8092E-5	-9299.44
Aug. 83	$z(t)=(1-0.3852B^{336})(1+0.3161B)\epsilon(t)$	1.1718E-4	-9120.21
Sep. 83	$z(t)=(1-0.4406B^{336})(1+0.2359B)\epsilon(t)$	1.0359E-4	-9244.47
Oct. 83	$z(t)=(1-0.9729B^{336})(1+0.1110B)\epsilon(t)$	8.5380E-5	-9439.35
Nov. 83	$z(t)=(1-0.6251B^{336})(1+0.0050B)\epsilon(t)$	9.7523E-5	-9305.31
Dec. 83	$z(t)=(1-0.3967B^{336})(1+0.2226B)\epsilon(t)$	1.3336E-4	-8989.84
Jan. 84	$z(t)=(1-0.5241B^{336})(1+0.1927B)\epsilon(t)$	8.5010E-5	-9443.72
Feb. 84	$z(t)=(1-0.5323B^{336})(1+0.2825B)\epsilon(t)$	1.5045E-4	-8868.30
Mar. 84	$z(t)=(1-0.2707B^{336})(1+0.0278B)\epsilon(t)$	1.1458E-4	-9142.83
$z(t)=(1-B^{336})(1-B)x(t)$			

Table 5.1: Experimental "Airline" Model

5.2.2 Long Memory and Short Memory Models

By noting that the purpose of the de-seasonal difference operation $(1 - B^s)$ in a multiplicative seasonal model or in a general ARIMA model is to transfer long memory type time series (or non-stationary periodic time series) to a short memory type series (or stationary time series), Parzen (1982) recommended ARARMA model schemes to model non-stationary periodic time series. As with an ARIMA model, an ARARMA model consists of two filters, namely, *a long memory filter* and *a short memory filter.* This model is designed to filter the non-stationary periodic time series into a short memory type series and then to a no memory type series. i.e. the long memory filter transforms the long memory type series (or non-stationary periodic time series) into a short memory type series (or stationary time series), and the short memory filter transfers this short memory type series into no memory type series (or white noise).

Unlike multiplicative seasonal and ARIMA model schemes introduced by Box and Jenkins (1976) in which the long memory filter is constrained to be composed of a pure difference operation, a linear transformation of the AR form is suggested by Parzen (1982) as the long memory filter in an ARARMA model scheme. Since the pure difference operation is a special case of an AR form linear transformation, the ARIMA schemes are a special case of the ARARMA scheme.

More generally, a suitable ARIMA form linear transformation, where the roots of the characteristic function of the autoregressive part, AR, are outside the unit circle, can also serve as the long memory filter which transforms a long memory type time series, such as non-stationary periodic series, into a short memory stationary type series which can be adequately modeled by an ARMA model. The above long memory ARIMA filter and short memory ARMA filter comprise an AR(ARIMA)MA model. If the moving average term of the long memory ARIMA filter is invertible, the long memory ARIMA filter is equivalent to an AR(∞). i.e. the roots of characteristic function of the MA term are outside the unit circle, the long memory ARIMA filter can be approximated by an AR(n) filter with a finite order n. The order n is determined by an order-determining criterion such as the AIC, Hannan, CAT, Schwarz,

etc. Therefore, the ARARMA model suggested by Parzen (1982) is a special case of AR(ARIMA)MA models. Because the long memory filter takes the form of an ARIMA, it is hard to specify unless one has *a prior* information about the long term behaviour of a time series. However, the AR(ARIMA)MA is still a very useful approach for some special circumstances.

After long memory filtering, the short memory series produced can be "whitened" by a short memory filter of the form of an ARMA(p, q) which has an equivalent AR(∞) form. With similar arguments to that for the long term memory filter, the short memory filter ARMA(p, q) can be adequately approximated by an AR(m) where $m > p$ model which has a finite order, m, although an AR(p) scheme may be sufficient in many cases.

In summary, a time series based on an AR(ARIMA)MA model can be approximated by an AR(m)AR(n) model, and furthermore, some coefficients of the AR(m)AR(n) model may not be significant. For the sake of parsimony, the subset AR selection procedure described in chapter 2 can be employed to chose an optimum subset ARAR model to approximate the AR(ARIMA)MA model.

The major disadvantage of using the Box-Jenkins multiplicative seasonal type of model or AR(ARIMA)MA is that the modelling procedure is a trial and error procedure, and it is difficult to use a model selection criterion such as AIC, CAT, Hannan, and Schwarz to select a good model unless one has a good deal of knowledge, and experience with the Box-Jenkins modelling procedure. Consequently, it is very difficult to design an automatic procedure and to obtain a good model without certain restrictions on the model. The subset ARAR modelling procedure described in section 5.2.2 overcomes the above disadvantage to a considerable extent. Once the maximum lag and suspected seasonal periods are specified, the proposed subset ARAR model procedure will automatically yield an optimum subset ARAR model. A subset ARAR model procedure is designed for the New Zealand short term data set as follows:

Step 1: Parzen (1982) has suggested a mechanism to search for lags in the long memory AR filter of his ARARMA model. However, this mechanism is not practical

for the New Zealand half-hourly electric load because of the sample data which have to be employed to obtain weekly lag auto-correlations. The weekly lag cannot be convincingly identified by this mechanism for a long memory filter if only a few weeks of sample data are employed. Practically, however, we know that a weekly periodic pattern dominates the periodicity of this data set; and a daily periodic pattern dominates weekday (Monday to Friday) load variations. From the above *a priori* information, a subset AR model with lag 48 and 336 is specified to fit the series to generate a short memory series. So the expression,

$$(1 + a_{48}B^{48} + a_{336}B^{336})x(t) = y(t) \tag{5.26}$$

represents a long memory filter to transform the long memory type series $\{x(t)\}$ into a short memory type series $\{y(t)\}$

Step 2: Determine the order m of a short memory AR(m) filter, which is an approximation to a short memory AR(∞), for the short memory type series $\{y(t)\}$ by the order determination criterion, such as the AIC. We know that the AIC is the most "conservative" criterion among the well known criteria.

Step 3: The short memory series, $\{y(t)\}$, is fitted by the subset AR scheme described in chapter 2, which, again, is an approximation for the AR(m) filter for the short memory series produced in step 1. As we know, there is an ordering in conservatism among the model selection criteria. The hierarchy governing AIC, Hannan, Schwarz, is that AIC is more conservative than Hannan and Hannan is more conservative than Schwarz. For parsimony, the first proposed short memory subset AR model is determined by Schwarz; the second proposed short memory subset AR model is determined by Hannan; and the third proposed short memory subset AR model is determined by the most conservative criterion, AIC.

Step 4: Define the ith proposed model in Step 3 ($i = 1, 2, 3$). Start with $i = 1$ and test the "whiteness " of the residual data produced by the ith proposed model.

If the residual data set passes the "whiteness" test, it indicates the subset ARAR model containing the long memory AR filter (5.26) and the ith proposed short memory subset AR model is suitable to fit the sample data set, and we proceed the next step. If the residual data set does not pass the "white noise" test, go back to the beginning of this step with $i = i + 1$. In most circumstances, the short memory subset model chosen by AIC is over-parameterized and is able to pass the "whiteness" test.

Step 5: Forecasting.

5.3 Application

For comparing the performance of the two modelling methodologies, we consider:

- Adaptive Additive Model

  Model 1: Box and Jenkins multiplicative seasonal ARIMA model

  Model 2: Multiplicative seasonal AR(ARIMA)MA model

  Model 3: Subset ARAR model

- Conventional Additive Model

  Model 4: Ordinary conventional additive model (fixed harmonic frequencies for the deterministic seasonal component)

  Model 5: Modified conventional additive model (Selected FFT frequencies for the deterministic seasonal component)

We apply them to the 4 different seasons for the New Zealand half-hourly load data set. Each sample data set for the different seasons contains three weeks' data. The one week ahead forecast performance of the five models are also compared for the different seasons. The 4 sample sets for the 4 different seasons are described as follows:

Autumn Data Set 4th April, 1983 to 24th April, 1983

Winter Data Set 4th July, 1983 to 24th July, 1983

Spring Data Set 3rd October, 1983 to 23rd October, 1983

Summer Data Set 9th January, 1984 to 29th January, 1984

5.3.1 How Well Does the Adaptive Additive Model Fit ?

Model 1: "Airline Model"

Fitting the data sets into an "airline" model frame work of the ARIMA$(0,1,1) \times (0,1,1)_{336}$ form we have the results in Table 5.2, and the goodness of fit diagnostics in figures B.1, B.2, B.3 and B.4. From the plots of the residual auto-correlations, it can be seen that the daily auto-correlation (lag 48 autocorrelation) is significant in the residual data set produced by the "airline" models for the four test sample data sets. This suggests that the "airline" model is not adequate for the four sample data sets because daily "seasonal" patterns are not properly addressed in the "airline" model and cause significant daily auto-correlation in the residual series. In other words, the "airline" models only transform the "long" memory type series (adaptive linear trend and weekly periodic pattern) into a "short" memory type series and these "short" memory series are far from white noise. Following the idea of an ARARMA model, a short memory model, such as ARMA(p, q) and a subset AR(m), etc, is needed to transform the "short" memory series into white noise.

Data set	Model	$\hat{\sigma}^2$	AIC
Autumn	$z(t) = (1 - 0.4242B^{336})(1 + 0.3226B)\epsilon_{aut}(t)$	1.77E-4	-8702.03
Winter	$z(t) = (1 - 0.9862B^{336})(1 + 0.2476B)\epsilon_{win}(t)$	1.50E-4	-8869.77
Spring	$z(t) = (1 - 0.9729B^{336})(1 + 0.1110B)\epsilon_{spr}(t)$	1.46E-4	-8895.53
Summer	$z(t) = (1 - 0.5241B^{336})(1 + 0.1927B)\epsilon_{sum}(t)$	0.95E-4	-9331.62
$z(t) = (1 - B^{336})(1 - B)y(t)$			

Table 5.2: "Airline" Model for the Data Sets in Different Seasons

Model 2: Multiplicative Seasonal AR(ARIMA)MA Model

After examining the results from the ARIMA$(0,1,0) \times (0,1,1)_{336}$ trial fitting in Table 5.1, we observe that the estimated coefficient on B^{336} from moving average side range form 0.27 to 0.98, and develop the following "long" memory filter ARIMA$(0,1,0) \times (0,1,1)_{336}$

$$(1 - B^{336})(1 - B)y(t) = (1 - 0.5B^{336})z(t) \tag{5.27}$$

as an adequate filter to convert our long memory type series (weekly periodical pattern) into a short memory type series (only daily seasonal pattern present). After applying the above "long" memory filter to the 4 sample data sets, a multiplicative model ARIMA$(1,0,1) \times (1,0,1)_{48}$ is recommended to fit those "short" memory series. These "long" and "short" memory filter models comprise our multiplicative AR(ARIMA)MA model for the four sample data set. The results from estimation of the parameters of these models are listed in Table 5.3 and the model diagnostics are given in figures B.5, B.6, B.7 and B.8.

Data set	Model
Autumn	$(1 - 0.292B^{48})(1 - 0.30B)z(t) = (1 - 0.22B^{48})(1 + 0.12B)\epsilon_{aut}(t)$
(s.e.)	(0.481) (0.211) (0.491) (0.218)
	$\hat{\sigma}^2$ = 1.37E-4, AIC = -8962.70
Winter	$(1 - 0.25B^{48})(1 + 0.18B)z(t) = (1 - 0.13B^{48})(1 + 0.32B)\epsilon_{win}(t)$
(s.e.)	(0.324) (0.258) (0.332) (0.249)
	$\hat{\sigma}^2$ = 1.42E-4, AIC = -8926.56
Spring	$(1 - 0.23B^{48})(1 + 0.53B)z(t) = (1 - 0.035B^{48})(1 + 0.61B)\epsilon_{spr}(t)$
(s.e.)	(0.148) (0.264) (0.152) (0.245)
	$\hat{\sigma}^2$ = 1.30E-4, AIC = -9015.56
Summer	$(1 - 0.38B^{48})(1 + 0.31B)z(t) = (1 - 0.16B^{48})(1 + 0.45B)\epsilon_{sum}(t)$
(s.e.)	(0.156) (0.238) (0.167) (0.224)
	$\hat{\sigma}^2$ = 0.91E-4, AIC = -9375.10
where $(1 - B^{336})(1 - B)y(t) = (1 - 0.5B^{336})z(t)$	

Table 5.3: AR(ARIMA)AR Model for the Four Sample Data Sets

Model 3: Subset ARAR Model

The proposed subset ARAR model procedure with model selection criterion AIC is employed to fit the four sample data sets and yields results listed in tables B.1, B.2,

B.3 and B.4 of Appendix B.3, and the corresponding model diagnostics are given in figures B.9, B.10, B.11 and B.12. Comparing the "long" memory AR filters estimated for the four different data sets from the tables B.1 - B.4 in Appendix B.3, we find that the "long" memory AR filters are similar, i.e. the estimated corresponding AR coefficients for the four different data sets are very close although the load profiles are quite different. This fact shows that the "long" memory type behaviour of the load data are similar. However, the "short" type memory behaviour is quite different. The differences are therefore reflected in the different nature of the "short" memory filters for the four different sample data sets in tables B.1, B.2, B.3 and B.4.

Model 4: Conventional Additive Model

The conventional additive model assumes that the deterministic periodic component consists of harmonic frequencies of the fundamental frequency corresponding to the periodicity (see equation (5.24)). By deleting those insignificant harmonic frequencies, the conventional additive model is applied to the four sample data sets and yields the estimation for those significant frequencies which are responsible for the periodic behaviour of the load. The stochastic component is produced by subtracting the estimated deterministic trend and periodic components from the data set and is fitted by the proposed subset AR model. In Appendix B.4, tables B.5, and B.6 list the estimated results for the autumn data set; tables B.7, and B.8 list the estimated results for the winter data set; tables B.9, and B.10 list the estimated results for the spring data set; tables B.11, and B.12 list the estimated results for the summer data set. The model diagnostic statistics are given in figures B.13, B.14, B.15 and B.16 for the autumn, winter, spring and summer data sets, respectively.

Model 5: Modified Conventional Additive Model

The modified conventional additive model assumes that pure discrete frequencies which are not necessarily harmonic to the fundamental frequency are responsible for apparent and hidden periodicities. The model for the periodic component is described in the last chapter and the first section of this chapter and is applied to the four sample data sets and yields the estimation of discrete frequencies for week days and weekend

days, and the subset AR model for the stochastic error terms. In Appendix B.5, the estimated periodic components and the subset AR models for the stochastic error terms are listed in tables B.13, and B.14 for the autumn data set, tables B.15, and B.16 for the winter data set, tables B.17, and B.18 for the spring data set, tables B.19, and B.20 for the summer data set. The model diagnostics are given in figures B.17, B.18, B.19 and B.20 for the autumn, winter, spring and summer data sets, respectively.

5.3.2 Model Fit Diagnostics

The idea of model fit diagnostic checking is to look for clues that may indicate the specification is deficient in some way or another. Therefore, the estimated variance of the residual is a criterion to judge the assumed model. Apart from the residual variance, it is necessary to check there is no information left in the residuals, i.e. the residuals are *white noise*. There are three major checks or tests for whiteness . The first is the *auto-correlation check* which seeks significant auto-correlations in the residuals. If some auto-correlations are significantly different from zero, this indicates the residuals are not close to white noise, and the assumed model may not adequate.

The second commonly used diagnostic checking test is the *portmanteau lack of fit test*, i.e. the Box-Pierce χ^2 test(see pp. 290 - 293, Box and Jenkins (1976)).

$$Q(k-m) = n\sum_{i=1}^{k}\hat{r}_i^2 \tag{5.28}$$

where m is the number of parameters estimated in a model; k is a constant and $k > m$; $\hat{r}_i$ is the i-th lag auto-correlation of the residuals from the model; n is the number of sample data points used to estimate the model parameters.

If the model is adequate, i.e. the residuals are white noise, $Q(k-m)$ should asymptotically satisfy a χ^2 distribution with $(k-m)$ degree of freedom.

The third is the *cumulative periodogram check* (for detail see pp. 294 - 298, Box and Jenkins (1976)). If the cumulative periodogram of the residuals from an assumed model exceed the 5% Kolmogrov-Smirnoff probability limits then this indicates that

the assumed model is not an adequate fit to the data set. However, the cumulative periodogram only supplies a very rough guide to the significance of apparent deviations. The deviations indicate the periodic effects have been inadequately accounted for by the assumed model if these deviations exceed certain probability limits. Nevertheless, it does not mean the assumed model is an adequate fit to the data even if the cumulative periodogram is within certain probability limits.

From the cumulative periodogram check for the residuals from the five models (see Figure B.1 to Figure B.20), it can be seen that the cumulative periodograms are all within the Kolmogrov-Smirnoff 5% probability limits. Therefore, the cumulative periodogram is not sensitive enough to compare the fit performance of the five models. On the other hand, the cumulative periodogram check for the five model indicates there is no strong evidence against the five models because of the failure to account for periodic effects.

Inspecting the other model diagnostic checks for Model 1 from Figure B.1 to Figure B.4, we find that the residual auto-correlations at lag 48, 96, 144, 192 are significantly different from zero. This indicates that Model 1 does not properly account for the daily pattern of the load. In addition, the plots of probabilities of the Box-Pierce statistic $Q(k-2)$ are well below the 5% level at low lags, and the cumulative periodogram are close to the 5% probability limits. These model diagnostics all indicate Model 1 is not adequate for our load data. The rest of the four models do not systematically violate zero residual auto-correlations as does Model 1.

Model 2 fits the sample data sets reasonably well but cannot pass the Box-Pierce test well at low lag auto-correlations which can be seen in the plots of the P-value for the goodness of fit statistics in figures B.5, B.6, B.7 and B.8. The only reason for that is that the "short" memory filter, ARIMA$(1,0,1) \times (1,0,1)_{48}$, may be too parsimonious to fit the data fully. A remedy can always be found through a more generously parameterized ARMA model to meet the goodness of fit test. However, a better fit within the sample data set from a more complicated model does not mean that a better model has been found, and that the more complicated model will yield

Model / Data Set	Autumn	Winter	Spring	Summer
estimated variance $\hat{\sigma}^2$				
Model 1	1.77E-4	1.50E-4	1.46E-4	0.95E-4
Model 2	1.37E-4	1.42E-4	1.30E-4	0.91E-4
Model 3	1.25E-4	1.37E-4	1.11E-4	0.82E-4*[a]
Model 4	1.26E-4	1.48E-4	1.17E-4	0.86E-4
Model 5	1.14E-4*	1.35E-4*	1.02E-4*	0.85E-4
AIC				
Model 1	-8702.03	-8869.77	-8895.53	-9331.62
Model 2	-8962.70	-8926.56	-9015.56	-9375.10
Model 3	-9042.99	-8956.62*	-9166.83	-9472.06*
Model 4	-9023.06	-8856.86	-9091.76	-9398.05
Model 5	-9111.95*	-8949.52	-9230.06*	-9421.84

Table 5.4: The Comparison of Estimated Variance and AIC for the Five Models

[a]Symbol * indicates the minimum estimated disturbance variances and minimum AIC values of the five models for the four different data sets.

good forecasting outside the sample; since more "short" model parameters are to be estimated for the "short" memory model and the forecasting confidence interval will be wider.

Box-Pierce tests at lag 336 for the residuals from Model 4 are always below the 5% level (see plots of P-value for goodness of fit statistic from Figure B.13 to B.16), and the cumulative periodogram of Model 4 for the Spring data set exceed the Kolmogrov-Smirnoff 5% boundary (see Figure B.15). These facts indicate that Model 4 is not an adequate model.

Model 2, 3 and 5 do not have any serious violations of model adequacy. The estimated variance and AIC for the five models are listed for the four sample data sets in Table 5.4. It can be seen that Model 5 achieves the minimum residual variance among the five models for three sample data sets although Model 3 achieves the minimum residual variance for the summer data set. Recognizing the penalty from the number of parameters estimated in these models, we used AIC to evaluate the five models. Model 3 achieves the minimum AIC for the winter and summer data sets, while Model 5 achieves the minimum AIC for the autumn and spring data sets.

In the adaptive model group, i.e. Models 1, 2 and 3, Model 3 performs the best

in fitting the four sample data sets, and Model 2 is better than Model 1. These results clearly indicate that daily periodicity cannot be covered by the operation that removes weekly periodicity, and both weekly and daily periodic factors must be take into account in a model. Although Model 3 performs better than Model 2, it does not mean a subset ARAR model scheme is always better than an AR(ARIMA)MA model scheme because a subset ARAR model is a special form of AR(ARIMA)MA model. The advantage of the subset ARAR model scheme is that it is easy to implement an "automatic" procedure to select an optimal subset ARAR model efficiently while it cannot be so easily and efficiently done for the AR(ARIMA)MA model.

In the conventional model group, i.e. Models 4 and 5, Model 5 performs better than model 4 in fitting the sample data sets. The obvious reason is that the harmonic frequencies for weekly periodicity are not sufficient to model the periodic behavior of the data because there are extra "hidden" periodicities whose frequencies are not harmonic with the weekly frequency. These "hidden" periodicities, then, are passed on to the stochastic component and cannot easily be modelled by a subset AR model because the maximum lag for the subset AR scheme determined by the model selection criterion, AIC, is usually less than those reflecting the "hidden" periods.

The reason that the adaptive additive and the conventional additive models are selected from different data sets is due to the effects of weather conditions, especially, temperature on the electricity load. The winter and summer sample data sets are more "irregular" than are the autumn and spring sample data sets. Therefore, the deterministic trend and periodic components of Model 5 for the winter and summer sample data sets are not adequate to describe the stochastic behaviour of the trend and periodic components. For the same reason, as mentioned in section 5.2.1, the stochastic component of Model 5 may be still non-stationary and has the same stochastic structure as the sample data. However, the autumn and spring data sets are more "regular", i.e. there is no marked weather sensitive load because of the moderate weather conditions in Autumn and Spring, and so, the trend and seasonal

components are nearly deterministic, and therefore, Model 5 may be more appropriate to model that data and so achieve a smaller residual variance than Model 3. The adaptive mechanism of model 3 is very sensitive to the size of the disturbance and therefore may cause mis-adaptation. Model 3 may not be appropriate to model the "regular" spring and autumn data set but may be more appropriate for the "irregular" summer and winter data sets. Nevertheless, Model 3 and 5 are the first or second best models in sample and which is the better varies with the season of the sample observed.

5.3.3 Post-Sample Predictive Test

Since the best fitting model for within sample data may not guarantee the best forecasts out of sample, we conduct post-sample predictive tests and model evaluation for Model 3 and 5 with the maximum forecast of one week ahead. Because our primary interest is not only one step ahead prediction but also is multi-step ahead predictions, the post-sample multiple-step ahead predictive test and model evaluation become our major interests. However, the multiple-step ahead prediction errors are not independent of each other as are the one step ahead prediction errors, and the construction of a valid post-sample predictive test would have to take account of this dependence. Nevertheless, as Box and Tiao (1976) have shown all the relevant information for multiple-step ahead predictions is effectively contained in the one-step prediction errors, so we use the Chow-test. The quantities calculated arise from the post-sample predictions and are used to evaluate the two chosen models.

$$\text{Chow}(k) = \sum_{i=1}^{k} e^{*2}_{T+i}/k\, s^* \tag{5.29}$$

where $e^*_{T+i} = (x_{T+i} - \hat{x}_{T+i|T})/{w^*_{T+i}}^{1/2}$, w^*_{T+i} = prediction error weight,

$$s^* = \frac{\sum_{i=1}^{T} e_i^2}{T-L},\ L = \text{number of data used for initial estimates}$$

Under the null hypothesis $\mathbf{H_0}$: the one step ahead post-sample prediction will be worse than one step ahead in-sample prediction, Chow(k) $\sim F(k, T-L)$.

The Chow statistics and their probabilities for a one week ahead post-sample prediction based on the four half-hourly sample data sets are listed in Table 5.5. From the Chow statistic values for Model 3, we can see that the null hypothesis $\mathbf{H_0}$ is rejected because there is no quantile of the Chow statistic exceeding 95%. This implies that the post-sample one step ahead prediction from Model 3 would not be worse. Similarly, the null hypothesis $\mathbf{H_0}$ is rejected when Model 5 is applied to the Autumn and Winter data sets. However, the null hypothesis $\mathbf{H_0}$ cannot be rejected when Model 5 is used for the Spring and Summer data sets. This indicates that Model 5 may not be appropriate. Therefore, overall the Chow-statistics for Model 3 and 5 indicate Model 3 is better than Model 5.

Model/Data Set	Autumn	Winter	Spring	Summer
Model 3	0.384584	0.535147	0.975784	0.957431
Prob. (F(p,q))	0(336,670)	0(336,626)	0.39(336,623)	0.33(336,623)
Model 5	0.429947	0.396865	2.00543	1.25816
Prob. (F(p,q))	0(336,910)	0(336,909)	1(336,911)	0.99(336,911)

Table 5.5: Post-Sample One-step Predictive Test — Chow Test

Prediction errors for more than one step ahead, therefore, provide no additional information for assessing the internal validity of a model, However, they are useful in providing a measure of predictive performance which can be used as a basis for comparison with rival models. The obvious statistic to consider here is called the *extrapolative mean sum of squares*

$$EMSS(T,k) = \sum_{i=1}^{k} e^2_{T+i|T} \tag{5.30}$$

where $e^2_{T+i|T} = x_{T+i} - \tilde{x}_{T+i|T}$, $i = 1, \cdots, k$.

The EMSS from one day ahead to one week ahead are listed in Table 5.6. From Table 5.6, it is obvious that the forecasting performance of Model 3 is better than Model 5's for the Spring and Summer data set; Model 5 is better than Model 3's for the Autumn data set; and it is hard to judge which model is better for the Winter data set.

L.T.[a]/Data Set	Autumn	Winter	Spring	Summer
Model 3 Forecasting Performance				
1	12.1E-2	3.17E-2*	7.15E-2*	2.58E-2*[b]
2	9.01E-2	4.26E-2*	6.68E-2*	2.71E-2*
3	7.48E-2	6.09E-2*	5.96E-2*	2.38E-2*
4	6.69E-2	7.54E-2	5.92E-2*	2.44E-2*
5	6.20E-2	9.94E-2	5.66E-2*	2.49E-2*
6	6.12E-2	11.5E-2	5.57E-2*	2.50E-2*
7	6.26E-2	12.6E-2	6.33E-2*	2.55E-2*
Model 5 Forecasting Performance				
1	2.27E-2*	8.09E-2	8.24E-2	7.01E-2
2	4.24E-2*	7.93E-2	8.21E-2	6.37E-2
3	3.81E-2*	7.12E-2	7.37E-2	5.89E-2
4	3.86E-2*	6.64E-2*	6.68E-2	6.21E-2
5	4.33E-2*	6.28E-2*	6.19E-2	6.24E-2
6	4.96E-2*	6.51E-2*	5.91E-2	5.88E-2
7	5.32E-2*	6.81E-2*	6.34E-2	5.50E-2

Table 5.6: Comparison of Model 3 and 5 in Forecasting Performance

[a]L.T. stands for the lead span day(s).
[b]Symbol * indicates the best predictive performance of a model compared to its competitor.

Because the above multi-step post-sample predictive performance comparison for Model 3 and 5 is not sensitive to the predictive performance on a particular day of a week, another comparison is conducted on the basis of the daily EMSS from the first day to seventh day ahead, and the results are listed in Table 5.7. From Table 5.7, we can see clearly that Model 3 performs better than Model 5 on the first day ahead in most cases except that Model 5 performs better than Model 3 in a significant way only on the first day ahead for the Autumn data set. The reason for this is that this day happens to be a public holiday Monday. Model 5 has adjusted its forecast values for this day because we know this day is a public holiday in advance. While Model 3 does not adjust its forecast values for this day. One simple adjustment can be made by replacing the forecast by the previous Sunday's values in order to avoid large forecasting errors for this day.

L.T.I.[a]/Data Set	Autumn	Winter	Spring	Summer
Model 3 Forecasting Performance				
1st	12.1E-2	3.16E-2*	7.15E-2*	2.69E-2*[b]
2nd	4.13E-2*	5.14E-2*	6.18E-2	3.23E-2*
3rd	2.38E-2*	8.66E-2	4.19E-2	1.88E-2*
4th	3.41E-2*	10.8E-2	5.85E-2	2.16E-2*
5th	3.69E-2*	16.3E-2	4.46E-2*	2.38E-2*
6th	5.74E-2*	17.2E-2	5.12E-2	2.42E-2*
7th	7.11E-2*	18.2E-2	9.74E-2	2.74E-2
Model 5 Forecasting Performance				
1st	2.47E-2*	8.09E-2	16.2E-2	7.01E-2
2nd	5.50E-2	7.79E-2	3.23E-2*	5.70E-2
3rd	2.78E-2	5.12E-2*	2.97E-2*	4.82E-2
4th	4.05E-2	4.98E-2*	2.55E-2*	7.11E-2
5th	5.88E-2	4.62E-2*	4.70E-2	6.42E-2
6th	7.34E-2	7.55E-2*	4.95E-2*	3.54E-2
7th	7.15E-2	8.42E-2*	6.28E-2*	2.13E-2*

Table 5.7: Comparison of Model 3 and 5 in Daily Forecasting Performance

[a]L.T.I. stands for the lead daily interval.
[b]Symbol * indicates the better predictive performance of a model compared to its rival model.

5.4 Summary

Although Model 3 fits the winter and summer data sets better than Model 5 does, the multi-step ahead predictions of Model 3 may not definitely be better than Model 5's. A typical example is the post-sample predictions for the winter data. From Table 5.6 and 5.7, we can see that Model 3's performance is better in first two days ahead, and Model 5's performance is better from three to seven days ahead. A similar thing happens for the spring and autumn sets. Although model 5 fits better in sample, this does not guarantee that Model 5 will produce better multi-step ahead predictions.

From examining the trend and seasonal update functions of an adaptive additive model, we know that the difference between the trend of a adaptive additive model (Model 3) and a conventional additive model is that the former trend component in the post-sample period only relies on the estimation of the adaptive trend component at the last point of a sample data set(see equation (5.14) and (5.15) while the latter trend component in the post-sample is dependent on the estimation of the "average"

trend in the sample. Therefore, because it is believed that there is no dramatic changes of trend in our data, Model 3 will be better for a "small" multi-step ahead prediction, and Model 5 will be more reliable for a "large" multi-step ahead prediction.

Chapter 6

Weather Sensitive Load

6.1 Introduction

Existing load forecasting techniques are basically categorized into two distinct classes:

- Non-Weather Related Load Models
- Weather Related Load Models

Weather conditions have a significant effect on electricity consumption whenever substantial cooling and heating loads are present. Weather Related Load Models have gained attention primarily because of their improved accuracy over the Non-Weather Related Load Models.

In most load models, the weather related load is one component of the net load demand. i.e.

$$y_t = z_t + x_t + v_t \tag{6.1}$$

where y_t is net load; z_t is the weather insensitive component or base load component; x_t is the weather sensitive component; v_t is a disturbance component.

There are many ways to model the base load component, such as ARIMA, State Space, Harmonic analysis, etc. In this chapter, we concentrate our discussion on weather sensitive load modelling.

It is known that the electricity load and the corresponding temperature measurement are nonlinearly related. One can hypothesize and also find evidence that the nonlinear relation also varies with time. Weather Load Related Models considered are further classified as Regression Models and Stochastic Models. Regression Models involve the effect or impact of more than one weather variable on the load, whereas Stochastic Models attempt to model properly the stochastic behaviour of the load.

In this chapter, we will build non-linear regression models for three-hourly electricity load for Canberra and then select and estimate parameters for the proposed models. After selecting an optimal model, we can decompose the load data into the weather sensitive and insensitive components; and the profiles of these components will be discussed. The stochastic effect of the weather sensitive component will be presented in chapter 7.

6.2 Preview

Before building a nonlinear modelling structure for the relationship of load and temperature, several nonlinear models depicting this relationship in the literature are summarized below:

Empirical Non-linear Function

In many reports, it was found that the values of correlation functions corresponding to weather variables other than Temperature and Humidity were always outside confidence bands corresponding to two standard deviations about zero. Correlations corresponding to humidity were relevant only in Summer. The above evidence reflects the reality that most cooling and heating devices are thermostatically controlled and that the cooling devices, especially air conditioners, are not only dependent on temperature but also on humidity in summer.

In general, the temperature effects on the electricity load is a nonlinear relationship. The higher the temperature the greater the electricity demand for cooling; the

lower the temperature the greater the electricity demand for heating. In the literature of modelling the electricity demand or consumption, linear time series models (or linear system models) are usually employed. In order to explain the load behaviour more precisely, the weather conditions, especially temperature, are employed as an explanatory variable for the load. Because a nonlinear relationship exists between the load and the corresponding temperature, it is unreasonable to employ temperature as an exogenous variable directly in the linear system models to explain the weather sensitive component of the load. One way to overcome this is to find the nonlinear relationship, then transforming the temperature by the nonlinear relation into a new variable which is linearly related to the load, and using this new variable as exogenous to the linear system model.

Many experiments shows that a positive (negative) temperature deviation ΔT_t from usual temperature, implying unusually hot (or cool) weather, leads to a load increase caused by extra cooling requirements in summer (or heating requirements in winter). Galiana et al. (1974) constructed the following non-linear function listed in Table 6.1 where W_t depends on the actual temperature T and the average temperature $\hat{T}$, i.e. $W_t = W_t(T, \hat{T})$

T (F)	Cooling + Heating $(T-70)-(60-\hat{T})$ $W = T + \hat{T} - 130$	Cooling $W = T - 70$	Cooling $W = T - \hat{T}$
70°	Heating $W = -(60 - \hat{T})$	No Effect $W = 0$	Cooling $W = 70 - \hat{T}$
60°	Heating $W = -(T - \hat{T})$	Heating $W = -(T - 60)$	Cooling + Heating $(70 - \hat{T}) - (T - 60)$ $W = 130 - T - \hat{T}$
0	60°	70°	$\hat{T}$ (F)

Table 6.1: Non-linear Function W Actual Temperature T to Average Temperature $\hat{T}$

This non-linear function transforms the real temperature into a new temperature index $W_t(T)$ which is roughly directly proportional to the load. Thus, it has clear physical explanations. i.e. the temperature does not lead to any extra load if it is

between $60°F$ to $70°F$. The temperature may lead to extra heating(cooling) load if it is less(greater) than $60°F(70°F)$ and the extra load is directly proportional to the new temperature index W_t. Nevertheless, this relationship is not time variant and is a crude approximation to the unknown true relationship between weather conditions and the loads.

As mentioned at beginning of this subsection, humidity is an influential meteorological factor in addition to temperature in summer. Campo and Ruiz (1987) introduced the combined effect on load of temperature and humidity. The so-called Temperature-Humidity Index (THI) is defined as follows:

$$THI = DT - 0.55(1 - RH/100)(DT - 58) \tag{6.2}$$

where DT represents Dry-bulb Temperatures in degrees Fahrenheit and RH represents Relative Humidities in percentage.

He claims that THI instead of temperature T is more reasonable in Table 6.1 if T and $\hat{T}$ are replaced by THI and $\widehat{THI}$, respectively.

Engle et al. (1986) proposed a semi-parametric estimation procedure for the relation between weather and electricity sales in which the model for the relation is presented in a regression form with temperature regressors carefully segmented into temperature variables. The optimal model or smoothing parameter is selected by the *generalized cross-validation* criterion. The problem with this approach is that it is quite expensive to search over smoothing parameters. Generally only a very rough grid search is performed and the bias in such a mixed parametric-nonparametric model is not well understood. Therefore, reliable confidence intervals are difficult to compute for the estimated curves.

Hagan and Behr (1987) suggested that if a simple model can be found for the load/temperature relationship and thus, provides significant improvement over non-weather related load models, it will indicate that a nonlinear model development is a fruitful area for future research. To use the nonlinear relationship in a load forecasting scheme, they suggest using a polynomial function to fit the kind of nonlinear relationship believed necessary after examining scatter diagrams of hourly loads

vs hourly temperatures for different data sets. The methodology for modelling the load-temperature relationship is to transform temperatures to a weather variable by a fixed polynomial function based upon fitting the past load-temperature relation, and then to relate the residuals of both load and transformed weather variables through a linear transfer function relation. In their paper, a third order polynomial was employed to fit the load/temperature relationship. Their specification demonstrates the value of using a nonlinear transformation in relating T_t and W_t and so improving the forecast performance.

Lu et al. (1989) argue that the nonlinear relation between temperature and load is time variant i.e. the sensitivity of load to temperature is different at different times of a day. Therefore, it is not appropriate to ignore the time factor in the load/temperature relation. They suggest updating the coefficients of a third order polynomial function in a recursive way to model the nonlinear relationship. In this way, the time factor can be properly taken into account.

In principle, a polynomial function with high enough order can be used to fit the relationship approximately. The order of the polynomial can be determined by various criteria. The disadvantage of polynomial fitting is the lack of physical explanation derived from this approach. Another disadvantage is that the fitted curves may not be consistent with the real world. For instance, an order three polynomial function was employed to fit the load/temperature relation at 9 and 12 o'clock based upon the load and temperature data set from the Canberra region in 1987. Three fitting approaches have been used namely least square, absolute residuals, and M-estimates(a robust estimation approach, where the bisquare function is chosen as the robust function). The latter two approaches give robust estimates to avoid outliers affecting the estimation of the polynomial coefficients. No matter which approach is used, it can be seen that the fitted curves are not realistic. The fitted curve indicates that the lower the temperature the less the load demanded when the temperature is lower than 35 *THI* at 9 o'clock (see Figure 6.1). The fitted curve shows an unrealistically high demand when the temperature is higher than 75 *THI* (see Figure 6.2). It is clearly

contrary to the expected relation. The above example suggests that a third degree polynomial function may not be a suitable way to describe the nonlinear relation between load and temperature, at least, for our data set.

Summarizing the existing methodologies in the literature, we find first the need for a transformation of weather conditions into a weather variable which is then linearly related with the load. The transformed variable is then to be used as an exogenous variable in a linear model system which describes the load. To overcome the disadvantages of Galiana's model and the polynomial model, a different approach has been developed to incorporate the nonlinear relationship between the load and weather conditions in the remainder of this chapter.

6.3 A Weather Sensitive Model

We primarily use a linear regression model to fit the daily average load by daily weather variables, such a maximum temperature (x_1), minimum temperature (x_2), evaporation (x_3) which is a function of humidity, wind speed above 3 metres (x_4), wind gust speed (x_5), sun duration (x_6), and daily type variables (x_7, x_8 are dummy indicator variables for Saturday and Sunday). The load on weekdays is assumed to be identical. In fact, the differences between the load on weekdays is insignificant if different weekday dummy variables are included in the linear model. The goodness of fit and estimated parameters are shown in Table 6.2.

There is no evidence that wind speed above 3 metres, wind gust speed or sun duration affect the load significantly. Whereas temperature(minimum and maximum), evaporation(or humidity) and weekend days are significant at the 0.01 significant level – see Table 6.2. We, therefore, adopt the Temperature Humidity Index(THI) (see equation (6.2)) as a single weather variable and investigate its relation with the load.

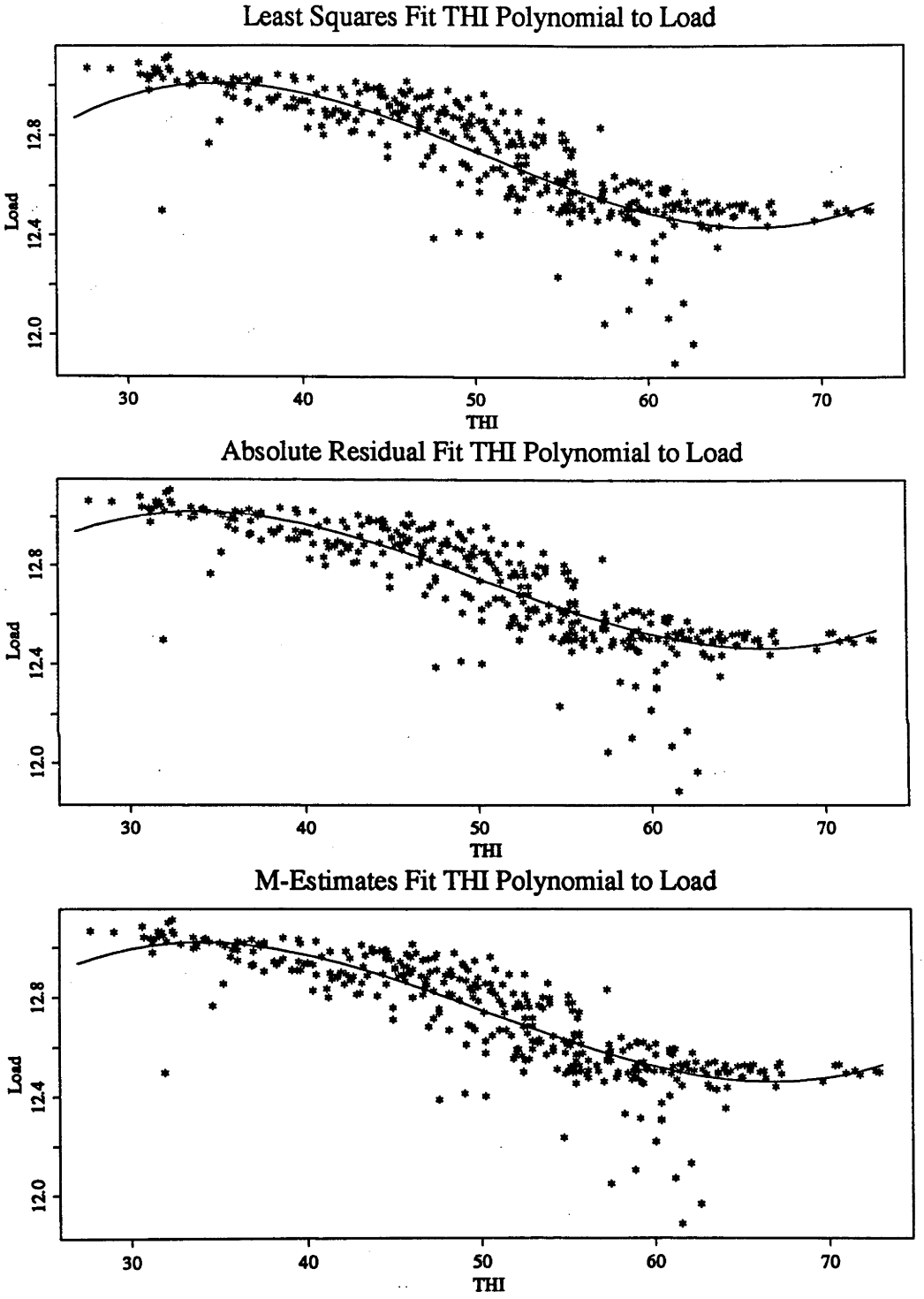


Figure 6.1: Third Order Polynomial Fit for the Load/Temperature Relation at 9 o'clock from the Load & Temperature Data Set of 1987

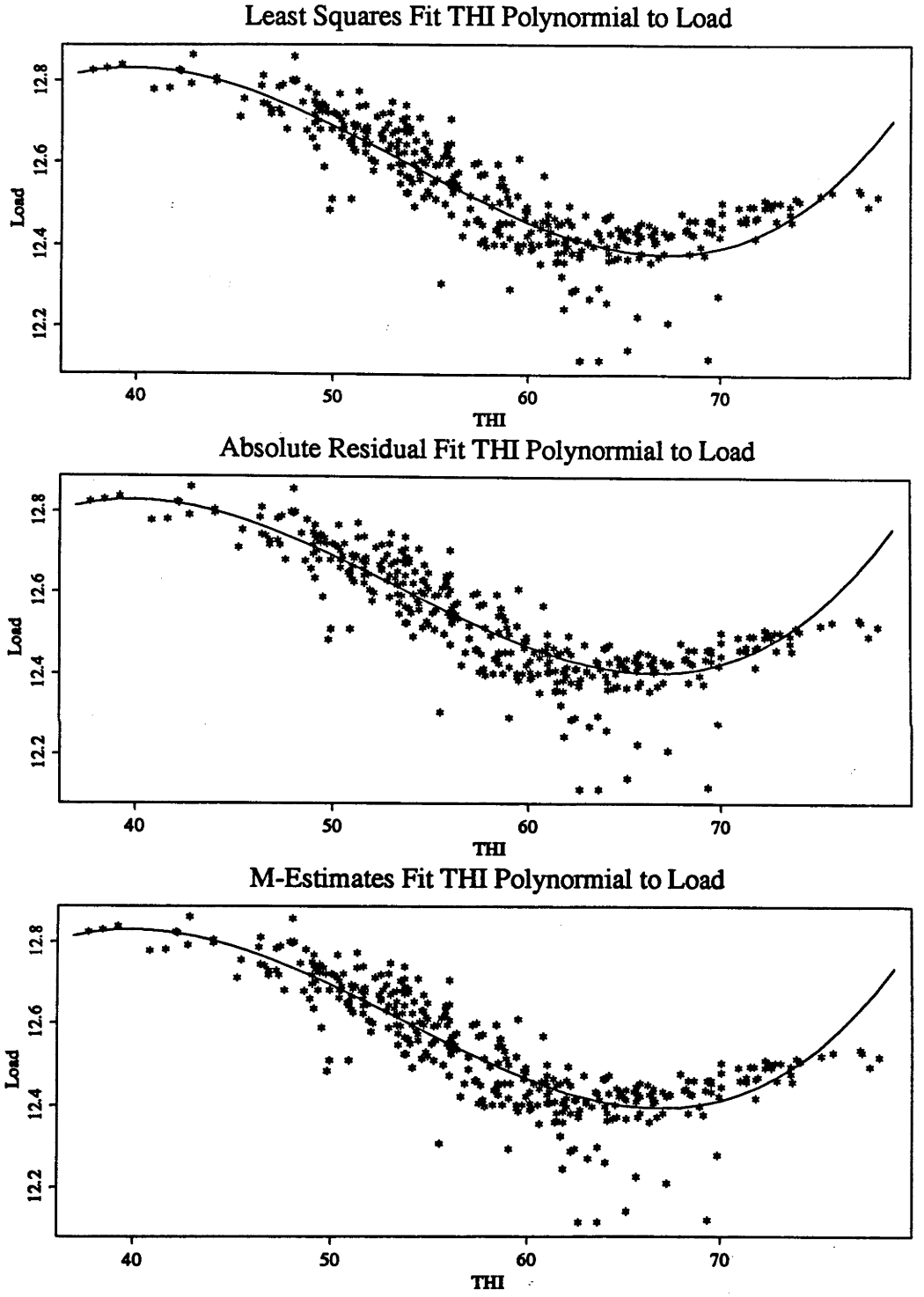


Figure 6.2: Third Order Polynomial Fit for the Load/Temperature Relation at 12 o'clock from the Load and Temperature Data Set of 1987

Residual Standard Error = 0.0659, R^2 = 0.9316 F-statistic = 563.5807 on 8 and 331 df, p-value = 0				
	coef.	std.err	t.stat	p.value
Intercept	11.6875	0.0195	599.6757	0.0
x_1	0.7133	0.0408	17.4810	0.0
x_2	0.3289	0.0377	8.7277	0.0
x_3	-0.0076	0.0019	-4.0283	0.0001
x_4	-0.0001	0.0	-1.0998	0.2722
x_5	0.0011	0.0006	1.9551	0.0514
x_6	-0.0009	0.0014	-0.6389	0.5233
x_7	-0.1066	0.0106	-10.0630	0.0
x_8	-0.1456	0.0106	-13.7774	0.0

Table 6.2: The Linear Regression of Daily Load on Weather Variables

6.3.1 Model Building

From the shape of scatter plots of the load against *THI*, we find that the shapes appear concave and that the load reaches a minimum around a value of 65 *THI* (*vertex*) at different times (three hours interval) of a day. The load increases when *THI* decreases or increases away from the *vertex*. The temperature sensitivity of the load depends on the capacity of electric heating and cooling devices and their efficient usage in the region in which the load data were collected. The efficiency of an electric heating (or cooling) device is measured by the ratio of the capacity of the device and the time required to return an existing low (or high) temperature to the desired or specified level. Roughly speaking, the bigger the capacity and the lower the efficiency, the the higher the temperature sensitive loads; and vice versa. For instance, the nonlinear relation could be symmetric about the *vertex* with a bounded saturation value if the capacities for heating and cooling are the same with the same efficiency. Furthermore, the capacities are time variant; for example, the heating and cooling facilities in an office building are fully responsible for the temperature during working hours. Those cooling and heating facilities may never be used during non-working hours.

The point is that those capacities and efficiencies are unknown. Therefore, we have to use a statistical approach to specify a model for the load/temperature relationship.

An assumption is made to specify a special case before we deal with the general case which can be approximately solved through modification of the solution of the special case.

Assumption 6.1 *The capacities for heating and cooling are same, with the same efficiency, in the region in which the load data were collected. The nonlinear function between load and temperature is a continuous function with continuous first derivatives.*

As mentioned above the nonlinear function could be symmetric about the *vertex* with a bounded saturation value arising from the above assumption. Suppose, the nonlinear function is

$$y_t = f_t(x_t) + \epsilon_x \tag{6.3}$$

where y_t, x_t, ϵ_x are the load, temperature and stochastic error at time t. According to the above analysis, the nonlinear function f_t must satisfy:

$$\frac{df}{dx} \mid_{x=D} = 0 \quad \text{(vertex point at } D\text{)}$$

$$f_t(x_t) = f_t(2D - x_t) \quad \text{(symmetric about vertex)}$$

$$\lim_{x_t \to \pm\infty} f_t(x_t) = A \quad \text{(a bounded saturation value)}$$

The other two parameters in f_t are required to describe the temperature sensitivity rate and the impact of concavity. In various families of functions, we find the following form for f_t is appropriate to meet our demands:

$$f_t(x_t) = A + C\, e^{-(x_t - D)^2/E}$$

the term $e^{-(x_t-D)^2/E}$ is a "bell" shape function, symmetric about D. The parameter $E(> 0)$ controls the "bell" shape. the larger E, the flatter the "bell" and vice versa. Therefore, the parameter E reflects the sensitivity of loads to the corresponding temperature changes. C shows the impact of the "bell".

$$\frac{df_t(x_t)}{dx_t} = -2C\frac{(x_t - D)}{E}e^{-(x_t-D)^2/E}$$

is anti-symmetric about D.

In more general cases, however, the derivative of the nonlinear function f_t may not be anti-symmetric about D due to the inequality of capacities for heating and cooling and their efficiencies. An adjustment can be made to the form of the derivative as follows

$$\frac{df_t(x_t)}{dx_t} = B(x_t) + 2C\frac{(x_t - D)}{E}e^{-(x_t-D)^2/E} \tag{6.4}$$

where $B(x_t)$ is an adjustment term to meet the asymmetry of the nonlinear relation of load and temperature. $B(x_t) = 0$, if the relation is symmetric about D.

For choosing the form of $B(x_t)$, we examine what are the properties of $B(x_t)$. Suppose, the cooling load is more sensitive than the heating load in the considered area i.e. there exists a region $\Delta_c = \{x_t : |x_t - D| < \delta_c,\ \delta_c > 0\}$. In this region, the absolute value of the first derivative of the proposed load/temperature function at any point which is less than the value of the *vertex* would be greater than the absolute value of the first derivative at the point which is symmetric to the former point about the *vertex*. In other words, the rate of change on the lower side of the *vertex* is greater than the rate of change on the other side for any pair of points which are symmetric about the *vertex* in the region Δ_c.

$\forall x_t \in \Delta_c$, we suppose $x_t^{(0)} < D$ without loss generality, the first derivative of $f_t(x_t)$ should satisfy

$$-\frac{df_t(x_t)}{dx_t}\Big|_{x_t=x_t^{(0)}} \geq \frac{df_t(x_t)}{dx_t}\Big|_{x_t=x_t^{(1)}}$$

where $x_t^{(1)} = D + (D - x_t^{(0)}) = 2D - x_t^{(0)}$

Because $x_t^{(0)}$ and $x_t^{(1)}$ are symmetric about D and

$$-2C(x_t^{(0)} - D)e^{-(x_t^{(0)}-D)^2/E} = 2C(x_t^{(1)} - D)e^{-(x_t^{(1)}-D)^2/E}$$

then,

$$-B(x_t^{(0)}) \geq B(x_t^{(1)})$$

Therefore, $\forall x_t \in \Delta_c$, $B(x_t) \leq 0$ and $B(D) = 0$.

Furthermore, because the load $y_t(x_t)$ reaches bounded saturation values when $x_t \to \pm\infty$, therefore,

$$\lim_{x_t \to \pm\infty} B(x_t) = 0$$

On the other hand, if the cooling load is less sensitive than the heating load, there exist a region $\Delta_h = \{x_t : |x_t - D| < \delta_h,\ \delta_h > 0\}$. For any $x_t \in \Delta_h$, $B(x_t) \geq 0$, $B(D) = 0$ and $\lim_{x_t \to \pm\infty} B(x_t) = 0$.

It can be concluded that the adjustment term $B(x_t)$ keeps the same sign for both $x_t < D$ and $x_t > D$ cases. i.e. $B(x_t) \leq 0$ ($B(x_t) \geq 0$) if the cooling (heating) load is more sensitive than the heating (cooling) load and is relevant to a certain region of temperature only.

Integrating the expression for the derivative (6.4) produces

$$f_t(x_t) = A + \int B(x_t)dx_t + C\, e^{-(x_t - D)^2/E}$$

and because $\lim_{x_t \to \infty} y_t(x_t) = c$ (saturated value), $\lim_{x_t \to \pm\infty} \int B(x_t)\, d(x_t) = constant$, we have

$$B(x_t) = o(x_t^{-1}) \text{ when } |x_t - D| \text{ is big enough}$$

For instance, if the cooling load is more sensitive than the heating load, $B(x_t) \leq 0$ and $B(x_t)$ has a profile as shown in Figure 6.3.

Even when the profile of $B(x_t)$ is known, the exact model for $B(x_t)$ is not easy to specify. It may have a complicated form and many unknown parameters. For the sake of parsimony, we make

Assumption 6.2

1. *The adjustment is linear, i.e. $B(x_t)$ is a constant.*

2. *The effect regions Δ_c or Δ_h must be realistic, i.e. the temperature region is based on the conditions in the area where the data was collected.*

This assumption means that $B(x_t)$ is defined in the real temperature region and then as an approximation, we replace $B(x_t)$ in the load/temperature model by its

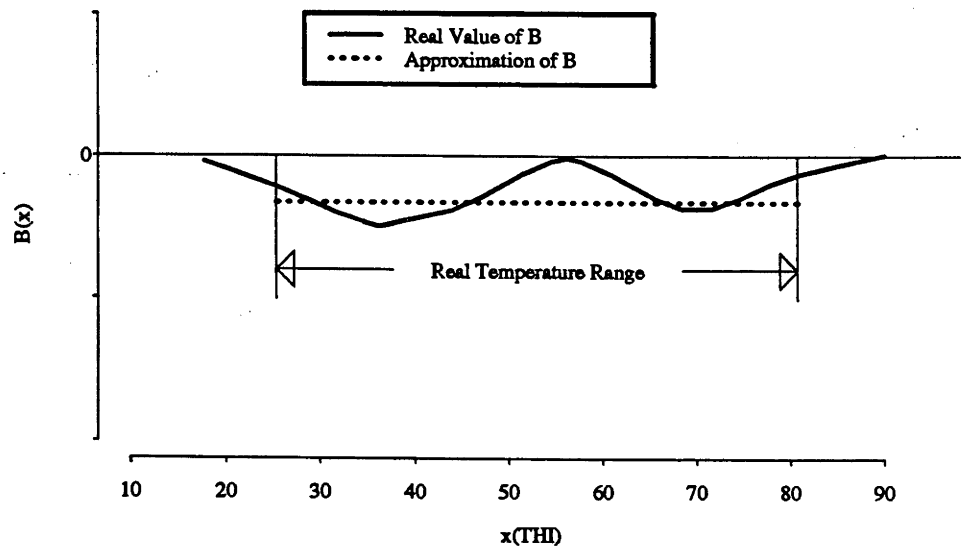


Figure 6.3: The Profile of Adjustment Term $B(x_t)$ When Cooling Load Is More Sensitive than Heating Load

average B (see dotted line in Figure 6.3). Under the above assumptions, the nonlinear load/temperature relationship can be modelled by the following expression,

$$y_t(x_t) = A + B\,x_t + C\,e^{-(x_t-D)^2/E} + \epsilon_x \tag{6.5}$$

To identify the above nonlinear equation, a nonlinear regression procedure is needed to estimate parameter set $\Theta = \{A,\ B,\ C,\ D,\ E\}$.

6.3.2 Model Estimation

The regression model (6.5) can be thought of as being made up of a deterministic part $f(x_t;\Theta) = A + B\,x_t + C\,e^{-(x_t-D)^2/E}$ and stochastic part ϵ_x. Although a wide variety of assumptions about ϵ_x are possible , the most frequent assumption about it is that it is independently and identically normally distributed with a constant but unknown variance σ^2. Unlike the linear regression model, the normal equations of a nonlinear regression model do not lead to an explicit expression for $\hat{\Theta}$. The least squares estimates $\hat{\Theta}$ are obtained only by an iterative procedure starting from some

assumed value of Θ_0. An important point is that the least squares estimator $\hat{\Theta}$ of Θ for the nonlinear model does not have the properties of a least squares estimator of a linear model. For instance, $\hat{\Theta}$ is a biased estimator of Θ for the nonlinear model, however, a least squares estimator for unknown parameter set of a linear model is an unbiased estimator. For finite samples, the general statement may be made that even though y_t may be normally distributed about its mean $f(x_t;\Theta)$ with some finite unknown variance σ^2 for all x, $\hat{\Theta}$ is not a linear combination of y_t and hence in general is not normally distributed, nor is it unbiased, nor is it a minimum variance estimator. Thus, unlike a linear least squares estimator, a nonlinear least squares estimator for Θ has essentially unknown properties for a finite sample size. Under the assumption for the disturbance term, $\epsilon_x \sim \text{NID}(0,\sigma^2)$, the nonlinear least squares estimate of the parameter set Θ in the deterministic part of a model is asymptotically unbiased, has asymptotic minimum variance and is asymptotically normal (see Chapter 12, Seber and Wild (1989) for a discussion).

However, there is often little understanding of the stochastic nature of the model. If such a model is fitted by, for example, least squares, the residuals $\hat{\epsilon}_x = y_t - f(x_t,\hat{\Theta})$ may provide some insight into the validity of the model and the error assumptions. Although such residuals can be misleading, particularly if the model is highly nonlinear, they still can be used to diagnose independence, to assess the normality of the disturbances, and whether the moments are identical. The independence can be tested via the auto-correlation of $\hat{\epsilon}_{x_t}$. The identical nature depends on the moments of the distribution of ϵ_x at each x, particularly the variance. In some cases, the magnitude of deviations from an estimated model depend on the magnitude of the response variable y. Seber and Wild (1989) (pp. 77 - 79) give an example to illustrate that a nonlinear least squares estimates of a parameter set Θ of a model may give completely misleading estimates if there is variance heterogeneity although the nonlinear least squares estimator is asymptotically unbiased, has asymptotic minimum variance and is asymptotically normally distributed.

To remedy this problem, there are two methods of modelling available: (1) the

transformation method, (2) the weighted least squares method. The main difference between the two methods is that the transformation method transforms so that transformed response variable $h(y_t) = h(f(x_t, \Theta)) + \epsilon^*$ has a different distribution as well as having a homogeneous variance structure. Whereas the weighted least square method models the variance heterogeneity but leaves unchanged the distribution of response variable y. In practice, however, the transformation chosen for the transformation method may not achieve the desired objective; also the estimated weights of the weighted least squares may depend on the parameter set Θ to be estimated. It should be noted that transformation and weighting change the values of parameter estimation and the standard errors of the estimates relative to the values obtained in the absence of weighting for finite samples. However, only when there is a marked non-homogeneity of variance will the differences between the weighted solution and un-weighted solution be considerable. For some practical data, it is often very difficult to detect, with any confidence the evidence of non-homogeneity of variance. The question of whether the distribution of ϵ_x at each x is normal is very difficult to resolve unless one has large samples. With large samples, the so-called quantile-quantile plots of the expected normal quantiles versus the residual from the fitted model may reveal departures from the normality assumption. With small samples, detection of non-normality may be tested by bootstrap techniques which usually will be extremely computer intensive and are not considered here.

Assumptions on which statistical models are built may often be based on limited knowledge. The extent to which the accuracy of a model may be affected by certain deviations from assumptions is an important general question. Another procedure, which is designed to make as few assumptions about the basic features of the data as possible, is called robust data analysis, which includes robust regression. This procedure is discussed in a book edited by Hoaglin et al. (1983). The concerns addressed by robust regression are focused on assumptions regarding the shape of distribution function of the assumed random error term ϵ_x. A robust estimate of a model can

make the model estimates robust to the effects of outliers. An outlier is an observation that is aberrant or unusually different from the rest of the observations. Often outliers arise in real data from known causes such as a change in pricing policy, a business promotion, or a labour strike. In his book, Huber (1981) established the effect of outliers on modelling in the case of an *i.i.d.* random variable.

For the nonlinear model (6.5), the normal equations for the least-squares estimate $\hat{\Theta}$ are

$$\sum_{t=1}^{n} \frac{\partial f(x_t, \hat{\Theta})}{\partial \theta_r}[y_t - f(x_t, \hat{\Theta})] = 0 \quad (r = 1, \cdots, p) \tag{6.6}$$

If there is variance heterogeneity, by using weighted least-squares estimate, the normal equation to estimate $\hat{\Theta}$ become

$$\sum_{t=1}^{n} w_t \frac{\partial f(x_t, \hat{\Theta})}{\partial \theta_r}[y_t - f(x_t, \hat{\Theta})] = 0 \quad (r = 1, \cdots, p) \tag{6.7}$$

When the weights w_t in equation (6.7) are unknown, the most common method of estimating parameter Θ in the deterministic part of model (6.5) is to apply weighted least squares with estimated weights $\hat{w}_t$. This method is called *generalized least squares* where the weights w_t are usually estimated from the variance of ϵ_x. For this reason, the modelling of the variance function is critically important to obtain the correct weighted least squares estimate of Θ. In variance function estimation, we often try to understand the structure of variances as a function of the predictor variable x_t, e.g.

$$\sigma_\epsilon^2 = \mathbf{var}(\epsilon_x) = g(x_t, \Psi) \tag{6.8}$$

On the other hand, there are many instances where the variance function is an important component of independent interest and not just an important part of estimating the deterministic component of model (6.5) since our interest is not just in estimating the response y_t from an observed variable x_t but also in the associated confidence interval. This interval is primarily determined by the variance function (see Watters (1987) for detail).

If outliers are a problem, a more robust method of estimation is needed. The outliers, which can be thought of as arising when the stochastic error term has a long tailed distribution, tend to have small robustness weights and therefore do not play a large role in the estimation of parameters of the model, By analogy with robust linear regression, an M-estimate $\tilde{\Theta}$ of Θ is the solution of (Huber (1981)) the equations

$$\sum_{t=1}^{n} \frac{\partial f(x,\tilde{\Theta})}{\partial \theta_r} \psi[\frac{y_t - f(x_t,\tilde{\Theta})}{\tilde{\sigma}}] = 0 \quad (r = 1, \cdots, p) \tag{6.9}$$

where ψ is a suitable function which down-weights or omits extreme values, and $\tilde{\sigma}^2$ is a robust estimate of dispersion (for a brief summary of robust estimation for simpler models see (Seber (1984), section 4.4). There are many robust function ψ, such as Andrews, Bisquare, Cauchy, Fair, Huber, Logistic, Talworth, Welsch (see Hampel et al. (1986) for detail). The choice of ψ is a matter for each investigator to decide based on each function's particular properties.

Therefore, if there is variance heterogeneity as well as outliers in a data set, a robust and weighted least-squares estimate should be used. There are many methods to estimate Θ in $f(x_t, \Theta)$ and Ψ in $g(x_t, \Psi)$. Basically, however, there are two kinds of methodologies (1) an iterative estimation approach (2) a simultaneous estimation approach. In the first, we estimate Θ with a given initial Ψ and using $1/g(x_t, \Psi)$ as weights at the first step; then update Ψ using the estimated Θ, $\hat{\Theta}$, which generates the residuals from the estimated $f(x_t, \hat{\Theta})$, and so on. In the later approach, Θ and Ψ are estimated simultaneously. There are many estimators for Θ and Ψ in both approaches, such as least-squares, maximum likelihood, etc. The least-squares estimator is used in our study. The advantage of the least squares estimates in the first approach is they are simple and numerically stable; the disadvantage is that the estimated Θ and Ψ are more biased than in the second approach. The advantage of the least square estimates in the second approach is there is less bias than in the first approach; the disadvantage, however, is that the non-linear least squares optimization problem can be so complicated that numerical solutions do not easily converge and may not necessarily be stable (see Carroll and Ruppert (1988) for a discussion).

It can be seen that the load/temperature relation is very scattered from Figure

6.1. For instance, the spread of load is lower when *THI* is below 50 or is beyond 70, and is higher when *THI* is in between 50 and 70, for the electricity load against *THI* at 9 o'clock. The above evidences implies that people can tolerate more temperature changes around 60 *THI* and do not in that region change their habitual electricity use as much. This is consistent as, for example, most people can wear more clothes instead of using heating devices when the temperature drops from $25C°$ to $15C°$. Nevertheless, most people cannot accept colder temperature, without using heating devices, when the temperature drops from $15C°$ or lower. In statistical terms, the evidence implies that the error variances are not identical. In addition, there are some load points far from the others. This indicates that outliers are a problem. There is always the danger that minor problems with the data or the model will destroy the good properties of the classical estimators. Even seemingly slight deficiencies in the data or small modelling errors can have disastrous effects. As many authors have pointed out, a single outlier in a large data set can overwhelm the normal-theory maximum-likelihood estimator. Also, when the errors in a regression model are close to normally distributed but with heavier tails, then the least-squares estimator can be substantially less efficient than certain alternative estimators. The latter include robust estimators using the actual distribution of the errors when this distribution is known. Overall evidence, therefore, suggests that we can use a robust and weighted least-squares nonlinear regression to handle the variance heterogeneity and those pathological pieces of data and so avoid misleading parameter estimation. Therefore, we use the first approach to estimate Θ and Ψ.

The problem is that we do not have enough information about the initial values of Θ and the formulation of the variance function $g(x_t, \Psi)$ although the formulation of the deterministic part, $f(x_t, \Theta)$, the load and weather condition relation, was constructed in the last section. We have developed a two stage estimation procedure to deal with this initial problem and to avoid the influences of variance heterogeneity and outliers.

In the first stage, a robust locally weighted smoothing developed by Cleveland

(1979) is employed to smooth the load/temperature relation. The advantage in using Cleveland's smoothing algorithm is that the algorithm does not make any assumptions about the data, and it can avoid the disadvantages of an initial assumption on the form of the variance function and on its parameter estimates since the algorithm does not estimate the model parameter Θ. Rather it estimates a robust locally weighted regression the response from a linear model, on a sectional basis, arising from an approximation of the deterministic part of the model. We assume the smoothed data are a realization of the "true" deterministic function at this stage. A nonlinear least-squares procedure is applied to the smoothed data to estimate the model function parameters. There are a great many algorithms to find the least squared estimator for non-linear models in the literature (see a review paper by Chambers (1973)), such as Gauss-Newton, Newton-Raphson, Levenberg-Marquardt, etc. The properties of those algorithms will not be explored here but we refer to Ratkowsky (1983) pp. 155 for a discussion. In our study, the Gauss-Newton algorithm is employed to give a simple solution.

There are no simple model criteria available to test and to select model functions which are based upon robust and weighted estimates where variance heterogeneity and outliers are involved. The model functions estimated based on the smoothed data are tested in the next section, although initial model estimation, testing, and other assessments based on model discrimination criteria, established under the assumption of independent, identical and normally distributed disturbances, must be affected by the smoothing of the data. The variance function construction and estimation will be discussed after the model function based on the smoothed data has been finalized.

In the second stage, the estimated model function parameters, $\hat{\Theta}$, and the variance function structure and its estimated parameter $\hat{\Psi}$ serve as the initial values for an iterative weighted and robust estimation procedure presented in section 6.4. The procedure is applied to the original data to estimate the model function parameters Θ and variance function parameters Ψ.

6.3.3 Model Selection

From Galiana's empirical nonlinear relation between the load and THI, the parameter D in the model (6.5) should be around 65. Therefore, we should ask two questions (1) Is $D = 65$ and time invariant? (2) Is the load more sensitive to cooling (or heating)? In other words, is B significantly different from zero. To answer these two questions, we start with a basic model f_1. More complexity, along with an increasing number of unknown parameters, is continually added, thereby forming a sequence of proposed models, which provides a nested model function set.

$$f_1: \quad y_t = A + C\, e^{-(x_t-65)^2/E} + \epsilon_x \tag{6.10}$$

$$f_2: \quad y_t = A + B\, x_t + C\, e^{-(x_t-65)^2/E} + \epsilon_x \tag{6.11}$$

$$f_3: \quad y_t = A + C\, e^{-(x_t-D)^2/E} + \epsilon_x \tag{6.12}$$

$$f_4: \quad y_t = A + B\, x_t + C\, e^{-(x_t-D)^2/E} + \epsilon_x \tag{6.13}$$

There are two nested relations among the models, $f_1 \subset f_2 \subset f_4$ and $f_1 \subset f_3 \subset f_4$ where $f_i \subset f_j$ implies that model function f_i is a reduction of f_j. We assume the "true" model function is one of them. For the selection of the "true" model, we consider the properties of a proposed model function $f = f(x_t, \Theta)$ where Θ is the parameter set to be estimated.

Model Criterion Function

In the case of estimated model functions, we have seen that there exist two basic quality criteria associated with the models, namely accuracy and stability (see details in Appendix C). It is obvious that the less parameters in a proposed model, the more stable the model. A stable model with a moderate RSS may be preferable to an unstable model with a small RSS. On the other hand, discrimination functions such as $RSS(f)$ or $R^2(f)$ decrease (or increase) with the number of estimated parameters and are not sensitive to model stability. For this reason a single discrimination

function needs to be constructed to consider both the accuracy and the stability of a particular model specification. This discrimination function is called a criterion function(CF) and takes into account the prospective model specification's goodness of fit and stability. The CF is used to eliminate unlikely models and so to select the best model from a set of candidate models.

A common CF comprises the sum of two terms. The first term measures the accuracy of the fitted model function to the response while the second term is a penalty function which penalizes each estimated model according to its instability. The larger the model's instability, the larger the second term should be.

Since the penalty function is to be sensitive to the stability of the proposed models, the second term is based on the stability function $SF(f)$ in Appendix C.2. It is noted that $SF(f)$ can be partitioned into linear and nonlinear parts. Hence, the second term is taken to be a function of $\alpha(x_t)$ and $\delta(x_t)$ and dependent on the location x. For stability over all locations, the penalty function is related to $\sum \alpha(x_t) = Tr[F'MF] = m$, the number of estimated parameters, and $\sum \delta(x_t)$. Thus, the linear part of the penalty is a function of the number of unknown parameters. So, the criterion function for an estimated model function $\hat{f}$ is defined as

$$CF(\hat{f}) = RSS(\hat{f}) - \hat{\sigma}^2[n-m] + \lambda\hat{\sigma}^4\hat{\Delta} \tag{6.14}$$

where λ is a constant, $\hat{\Delta} = \sum \delta(x_t)$ and $\hat{\sigma}^2$ is a consistent estimation of the residual variance, σ^2, of the true model.

It is noted that the $CF(\hat{f})$ is a combination of three parts: accuracy as measured by RSS, linear stability as assessed by $\hat{\sigma}^2[n-m]$, and nonlinear stability accounted for by $\hat{\sigma}^4\hat{\Delta}$, and where the constant λ balances the weight between the linear and nonlinear stability in the criterion function. Borowiak (1989) proved that under the true model f and normal regularity conditions, then $\lambda = 3$, $n^{-1}\mathbf{E}[CF(f)] \to 0$, as $n \to \infty$. If $\hat{\sigma}^2$ of the true model, is known and one uses $\lambda = 3$, the CF will tend to be positive large when an inappropriate model is examined. For instance, if a model is too simple(there are less parameters than needed) , it will cause RSS to be large in relation to its number of fitted parameters; if a model is too complicated(there

are more parameters than needed), $\hat{\sigma}^2[n-m]$ will be small, and the stability effect, $\hat{\Delta}$, arising from the unsuitable nonlinear portion of the model will be large although RSS may be small.

Thus, a model with large CF values relative to its competitors is unlikely to be the true model and can be eliminated from consideration. Unfortunately, the probabilities of correct rejection or selection of models in this selection procedure are hard to calculate and can be only evaluated by simulations which is beyond the scope of this thesis. In addition, $\hat{\sigma}^2$ may not be obtained before we know which model is the true model among the set of candidate models. In the Appendix C.3, we use various estimations of σ^2 to show that the model f_4 is the best model.

6.3.4 Variance Function

By using the locally robustly weighted smoothed data in the last section, we avoid the effects of variance heterogeneity and outliers in the data on the nonlinear least squares estimates of parameter set Θ in model (6.5). However, of course, the nature of any variance heterogeneity and what outliers, if any are present, are not well known. We can obtain better estimation of Θ if the variance function which describe the heterogeneity of variance is known, and if the effects of outliers are reduced or eliminated. Figure 6.4 shows the residuals from the estimated models and the real data at different times of a day. It can be seen that the plots of the residuals indicate that they are roughly symmetric, and their amplitudes tend to be small when THI is low or high. This implies that the variances tend to be small when THI is low or high and to be large when THI is in the mid-range.

Since least squares estimates are adversely affected by outlying responses (see Huber (1981)), because the method is particularly prone to degraded performance as unexpectedly large residuals will, when such residuals are squared, become very influential. Cohen (1984) noted this and suggested that a better performance can be obtained by using absolute residual (see also Judge (1985)). Harvey (1976) has suggested regressing the logarithm of the absolute residuals on the logarithm of their

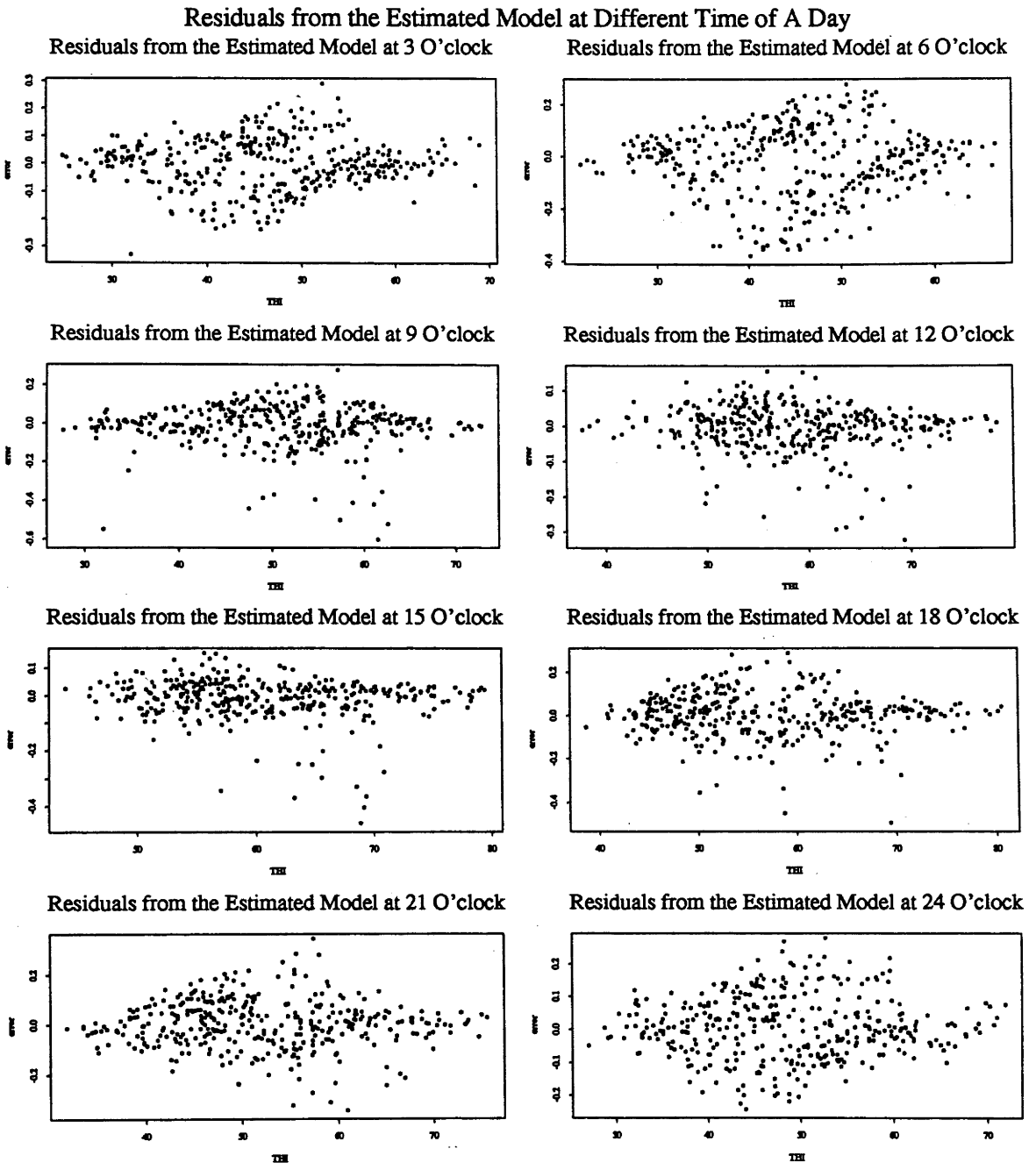


Figure 6.4: Residuals from the Estimated Model Based on the Smoothed Data

approximate expected values. Assuming that the errors are independent and identically distributed, this should produce an approximately homoscedastic regression. The calculations are easy since only an ordinary nonlinear least-squares program is required, although a practical problem can arise if any one of the residuals is very near zero, in which case taking logarithms induces a rather large and artificial outlier. By deleting those artificial outliers effects, we build a variance function model based on the logarithm of the absolute residual

The basic requirement amounts to an assumption of a different form, namely that the logarithm of absolute deviation has expected value giving by

$$\log(\mathbf{E}|y_t - f(x_t, \Theta)|) = \log(g(x_t, \Psi)) = h(x_t, \Psi) \tag{6.15}$$

Figure (6.5) shows the logarithms of the absolute residuals. With similar arguments to those given in section 6.3.1, a model for $\log(g(x_t, \Psi))$ is build as

$$h(x_t, \Psi) = a + b\,x_t + c\,\exp\{-(x_t - d)^2/e\} \tag{6.16}$$

where a is constant; b reflects the way of the variance function sensitive to temperature; c combined with a (possibly b) indicates the maximum value of the variance function; d indicates when the variance function reaches its maximum; and e reflects the flatness of the variance function.

Our primary interest centres on whether the variance function is more sensitive to low temperature than high temperature ($b < 0$) or vice versa ($b > 0$). In other words, is b significantly different from zero? To answer this question, we start with a simple basic model h_1. More complexity, along with an increasing unknown parameter b, is added, thereby forming a proposed model h_2. They form a nested model function set.

$$h_1 : \quad h(x_t, \Psi) = a + c\,\exp\{-(x_t - d)^2/e\} + \epsilon_x \tag{6.17}$$

$$h_2 : \quad h(x_t, \Psi) = a + b\,x_t + c\,\exp\{-(x_t - d)^2/e\} + \epsilon_x \tag{6.18}$$

where $h_1 \subset h_2$

With the possibility of outliers or artificial outliers caused by taking logarithms of the absolute residuals, we use a locally robust weighted regression to smooth the

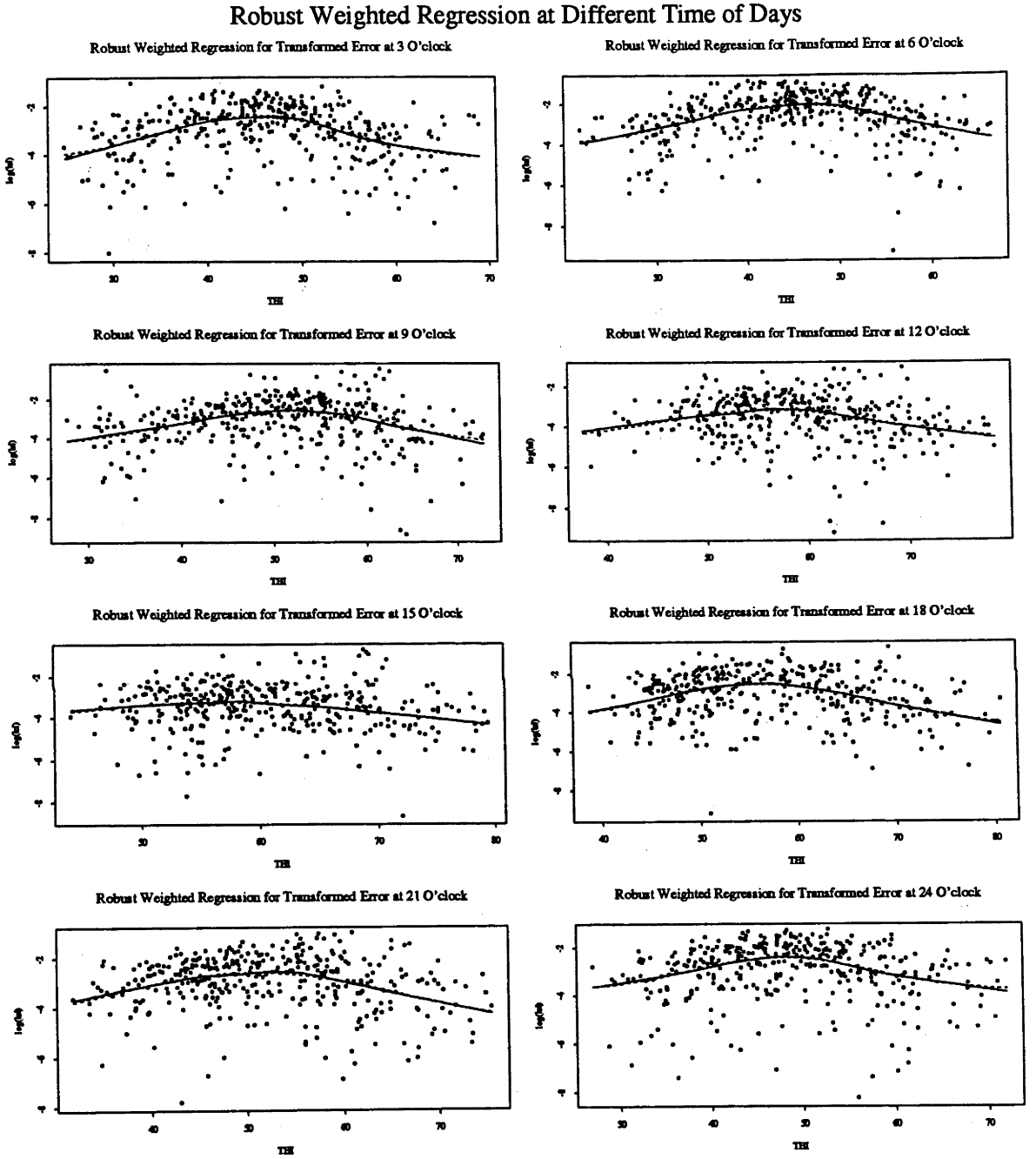


Figure 6.5: Logarithm of the Absolute Residuals from the Estimated Model Based on the Smoothed Data

logarithm of the absolute residuals and so to reduce the effects of those outliers in Appendix C.4. By using a likelihood ratio test and a model selection criterion CF, we are convinced that h_2 is the better model for the variance function.

6.4 A Robust-Weighted Nonlinear Least Squares

In section 6.3.3, we established a model relating electricity load and weather conditions, f_4, and the corresponding disturbance variance function h_2 in section 6.3.4. Bearing in mind that the parameters of these two models are least squares estimated on the locally robust smoothed data so that the effects of variance heterogeneity and data outliers are diminished in the least squares estimation. The estimated parameters may not be correct since they are based on the smoothed data which may affect the deterministic parts of the model function and the variance function. Nevertheless, the estimated parameters from the smoothed data can provide a rough approximation to the unknown parameters.

Being aware there is variance heterogeneity and there are outliers in the real data, we developed robust-weighted nonlinear least squares procedure to estimate the model parameters Θ and the variance function parameters Ψ as follows

Step 1 Using estimated parameters $\hat{\Psi}$ of the variance function $h_2(x_t, \Psi)$, and computing the estimated weights

$$u_t = 1/g^2(x_t, \hat{\Psi}) = 1/e^{2\,h_2(x_t, \hat{\Psi})} \tag{6.19}$$

calculate robustness weights by forming

$$w_t = u_t \times r_t \tag{6.20}$$

where

$$r_t = R(\frac{|(y_t - f(x_t, \hat{\Theta})|}{k\, g(x_t, \hat{\Psi})}) = R(\frac{|(y_t - f(x_t, \hat{\Theta})|}{k\, e^{h_2(x_t, \hat{\Psi})}}) \tag{6.21}$$

and k is a constant and R is the bisquare weight function defined by

$$R(x) = \begin{cases} (1-x^2)^2 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases} \tag{6.22}$$

Step 2 Let $\hat{\Theta}$ be the robust-weighted least squares estimate using the estimated robust-weights (6.20). Update the preliminary estimator by setting $\hat{\Theta}_* = \hat{\Theta}$.

Step 3 Calculate the residuals from updated function model

$$e_t(x_t, \hat{\Theta}_*) = y_t - f_4(x_t, \hat{\Theta}_*) \tag{6.23}$$

and form a logarithm transformation of the absolute residuals, i.e.

$$h_t = \log(|e_t(x_t, \hat{\Theta})|) \tag{6.24}$$

and then calculate the median of h_t, i.e.

$$s = \text{median}(h_t) \tag{6.25}$$

and calculate robustness weights by forming

$$v_t = R(\frac{h_t}{k\, s}) \tag{6.26}$$

Step 4 Using v_t as robust weights, fit h_t into model h_2 and let $\hat{\Psi}$ be the robust weighted least squares estimates and set the estimated updated $\hat{\Theta}_*$ as the preliminary estimator of the next iteration by using, $\hat{\Theta} = \hat{\Theta}_*$

Step 5 Repeat the cycle from step 1 to step 4 until there is little change in $\hat{\Theta}$ and $\hat{\Psi}$.

In steps 1 and 2 of the procedure, the model f_4 is fitted by robust-weighted least squares. The robust-weights w_t (6.20) which are designed to reduce the effects of variance heterogeneity and outliers, will distort the parameter estimates if the initial parameters are far away from their true values. Thus, the closeness of the initial

parameter values of the model function f_4 and the initial parameter values of the variance function h_2 to their true values plays an important role.

Therefore, the estimated parameters $\hat{\Theta}$ and $\hat{\Psi}$ from the robust locally weight smoothed data from section 6.3.3 and section 6.3.4 in Table C.4 and Table C.13 of Appendix C are used as the initial parameter values in the above iterative estimation procedure.

Similarly, in step 3, the robust-weights v_t (6.26) are designed to reduce the effect of outliers and artificial outliers which are produced by the logarithm transformation of the absolute residuals. The robustness parameter k in step 2 and step 3 should be chosen larger than 2. Roughly speaking, the larger k, the less robust. The choice of k is dependent on the quality of the data set. The robustness function chosen is the bisquare function (6.22) since other investigations have shown it to perform well for robust regression (see Gross (1977)). A reasonable range of k for our data is from 2 to 6 and $k = 3$ has been chosen in our study. The final estimated parameters $\hat{\Theta}$ and $\hat{\Psi}$ listed in Table 6.3 are considered to be the optimal estimates.

Figure 6.6 shows the normalized residuals, which are estimated by the robust-weighted nonlinear least squares estimation procedure with robust parameter $s = 3$ from model f_4 with variance function h_2, plotted against *THI* at different time of a day. It can be seen that the residuals have no systematic pattern and there are a few identified outliers. To support the assumption that residuals are normally distributed, Figure 6.7 displays Quantile-Quantile Plots (See Chambers (1983)) of the normalized residuals against the standard normal distribution at different time of a day. If the normalized residuals are normally distributed, then the plots(dot lines) will be approximately a straight line. It can be seen the plots are approximately straight lines except for a few identified outliers whose effects are reduced or eliminated by the robust weights. The above statistical evidence supports the claim that the model f_4 and variance function h_2 are built and estimated properly.

Model $\mathbf{E}y_t = A + B\,x + C\,e^{-(x-D)^2/E}$						
Time \Θ	A	B	C	D	E	vertex
3	12.2062	-7.0798E-03	-0.1670	55.6131	135.2751	58.48053
6	12.5431	-1.1451E-02	-0.1475	55.7049	60.9549	58.07098
9	13.4217	-1.2598E-02	-0.2107	60.1284	50.8239	61.64781
12	13.0939	-7.0320E-03	-0.2666	63.3080	94.2850	64.55146
15	12.9093	-3.8730E-03	-0.3529	64.6837	144.9248	65.47896
18	13.7179	-1.5526E-02	-0.3603	64.8314	114.1528	67.29093
21	13.4496	-1.4411E-02	-0.2762	61.8481	87.9413	64.14231
24	12.4812	-4.9326E-03	-0.3026	59.9780	245.1377	61.97596
Model $\mathbf{E}log(\lvert y - f\rvert) = a + b\,x + c\,e^{-(x-d)^2/e}$						
Time \Ψ	a	b	c	d	e	vertex
3	-3.0952	-5.0780E-03	1.1629	45.5038	96.3817	45.29337
6	-3.0832	-7.0247E-03	1.5520	45.8063	196.3287	45.36199
9	-4.1598	-1.3318E-02	2.6264	52.5150	497.3102	51.25411
12	-3.7259	-1.1452E-02	1.6560	53.3759	352.0990	52.15844
15	-5.5124	-1.3586E-02	2.4568	60.7947	3336.6332	51.56898
18	-2.0556	-2.9178E-02	1.6770	58.6597	168.9758	57.1897
21	-5.2966	8.8320E-03	2.6662	50.8843	368.0993	51.49398
24	-3.6676	-3.4793E-03	1.7976	47.6125	219.3726	47.4002

Table 6.3: Robust-Weighted Estimated Parameters

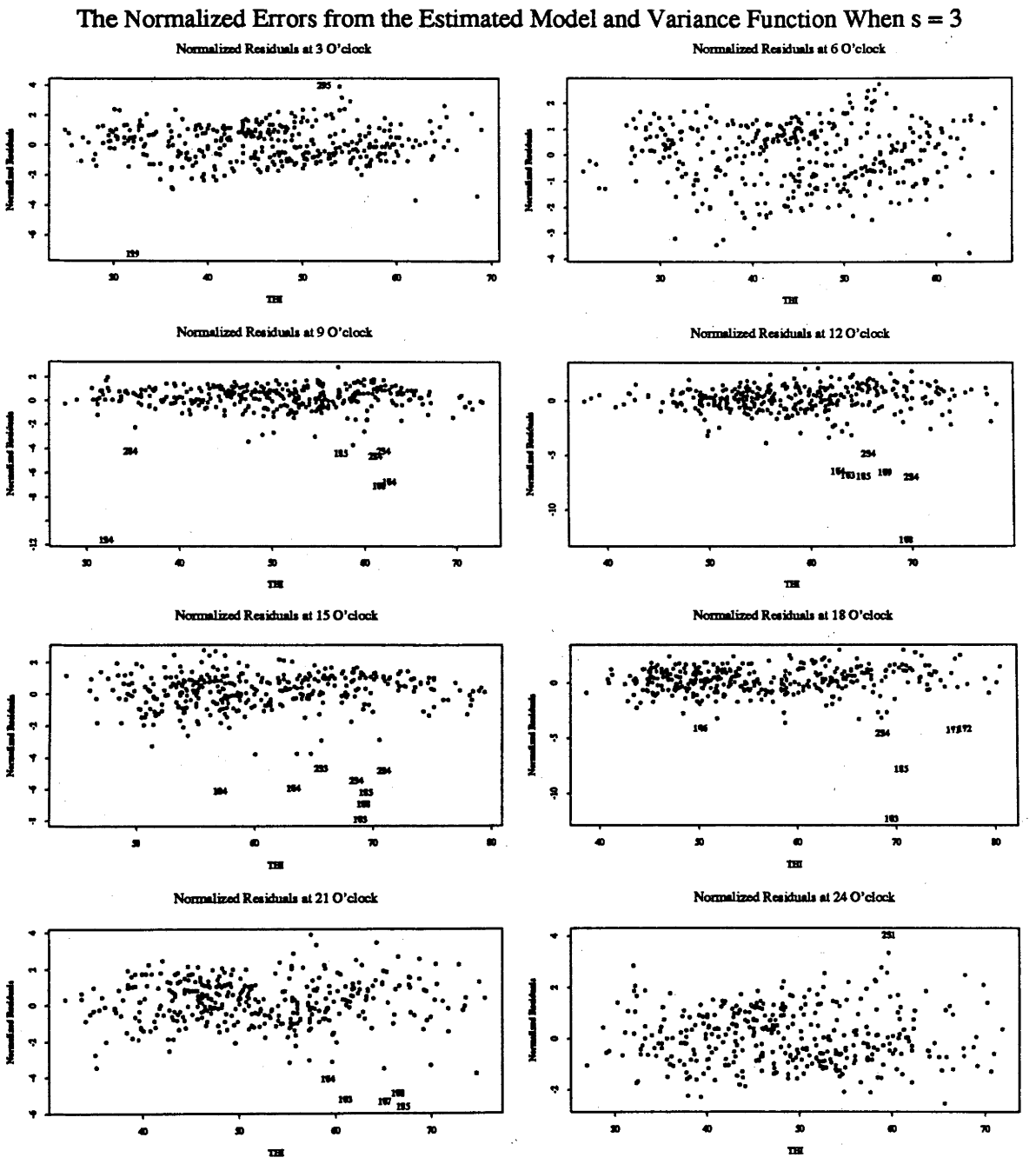


Figure 6.6: Normalized Residuals from the Estimated Model and Variance Function

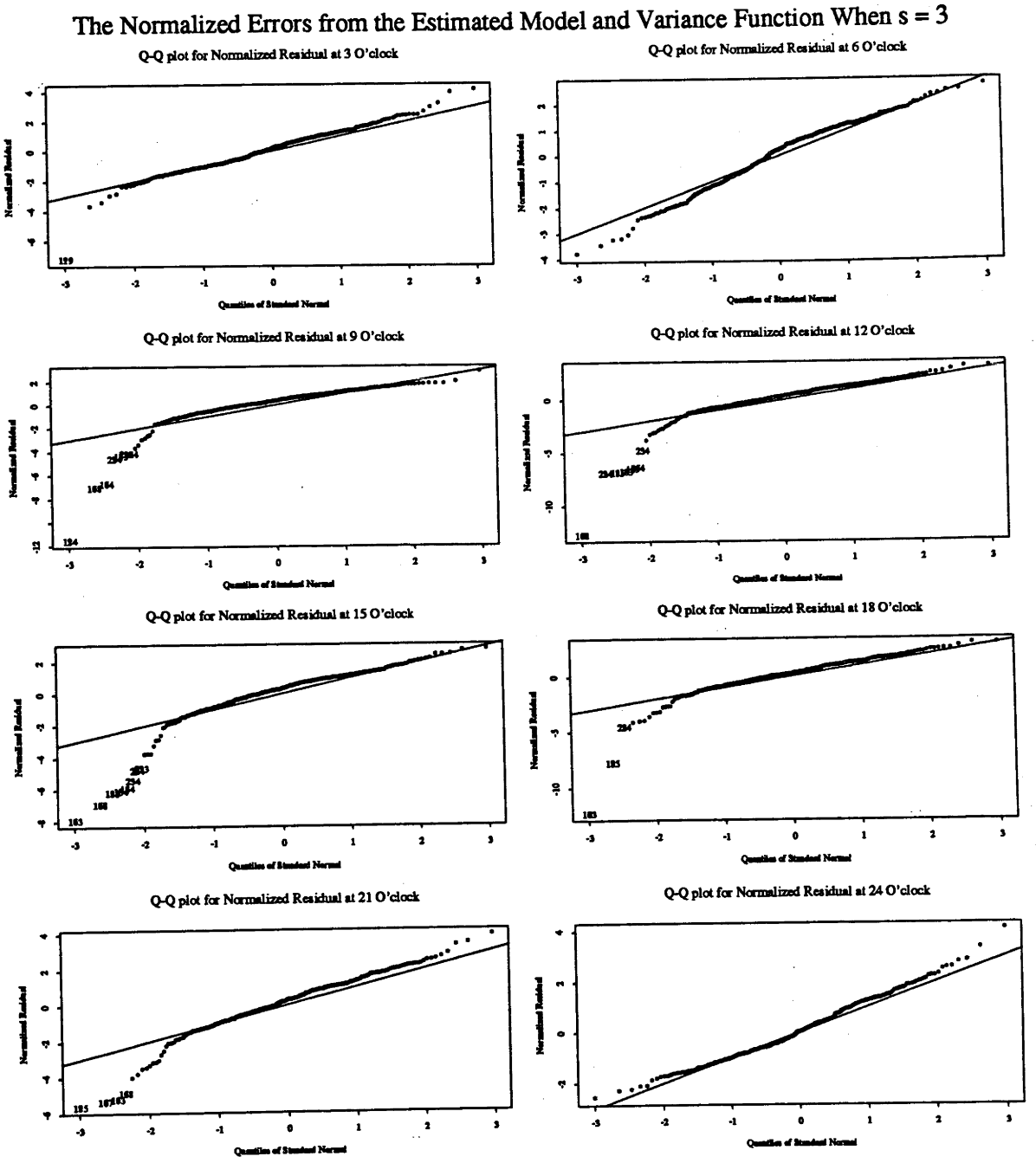


Figure 6.7: Quantile-Quantile of the Normalized Residuals

6.5 The Confidence Interval for the Estimated Model

When the stochastic term is *i.i.d*, ignoring a small bias, we have developed in Appendix C.2 that an asymptotic confidence interval for the assumed true model value $f(x_t)$ is (see equation (C.6))

$$\hat{f}(x_t) \pm [SF(\hat{f}(x_t)\sigma^2]^{1/2} z_{(1-\alpha/2)} \tag{6.27}$$

where z is the standard normal distribution and α is the significance level. However, in section 6.3.4, we have demonstrated that the stochastic term in model f_4 is not *i.i.d.* and the variance function model h_2 is an approximation to the variance heterogeneity. i.e.

$$\mathbf{E}(y_t) \approx f_4(x_t, \Theta)$$

$$\mathbf{var}(y_t) \approx \hat{\sigma}^2(x_t, \Psi) = e^{2h_2(x_t,\Psi)}$$

Hence, the asymptotic confidence interval for model $f_4(x_t, \Theta)$, with variance function $\hat{\sigma}^2(x_t, \Psi)$ replacing σ^2 in (6.27), is given by

$$f_4(x_t, \hat{\Theta}) \pm [SF(f_4(x_t, \hat{\Theta}))\, \hat{\sigma}^2(x_t, \Psi)]^{1/2} z_{(1-\alpha/2)} \tag{6.28}$$

On the other hand, the width of a confidence interval for $y_t - \mathbf{E}y_t = y_t - f_4(x_t, \Theta)$, with confidence coefficient $1-\alpha$, is $2\hat{\sigma}(x_t, \Psi) z_{(1-\alpha/2)}$. Thus, for a fixed α the range on $|y_t - f_4(x_t, \Theta)|$ is $\hat{\sigma}(x_t, \Psi) z_{(1-\alpha/2)}$. Since

$$\begin{aligned} |y_t - f_4(x_t, \hat{\Theta})| &= |(y_t - f_4(x_t, \Theta)) + (f_4(x_t, \Theta) - f_4(x_t, \hat{\Theta}))| \\ &\leq |y_t - f_4(x_t, \Theta)| + |f_4(x_t, \Theta) - f_4(x_t, \hat{\Theta})| \end{aligned} \tag{6.29}$$

the confidence interval of the estimated model $f_4(x_t, \hat{\Theta})$ is

$$\begin{aligned} & f_4(x_t, \hat{\Theta}) \pm \{\hat{\sigma}(x_t, \Psi) z_{(1-\alpha/2)} + [SF(f_4(x_t, \hat{\Theta}))\, \hat{\sigma}(x_t, \Psi) z_{(1-\alpha/2)}\} \\ \approx\ & f_4(x_t, \hat{\Theta}) \pm \{1 + [SF(f_4(x_t, \hat{\Theta}))]^{1/2}\} e^{h_2(x_t,\Psi)} z_{(1-\alpha/2)} \end{aligned} \tag{6.30}$$

Figure 6.8 shows the 95% confidence intervals of the estimated models at different time of a day. In each graph for the different time of a day, the dots are the original

data plot; the solid curve is the estimated model function, $f_4(x_t, \hat{\Theta})$, for the relation of the load and weather condition variable *THI*; a pair of dash curves are the 95% confidence interval of the estimated model function. It can be seen that the most dots are just surrounded by the 95% confidence interval except for a few outliers in each graph. This fact verifies that the estimated model function $f_4(x_t, \hat{\Theta})$ and variance function $h_2(x_t, \hat{\Psi})$ are properly built and estimated. The 95% confidence intervals from midnight to the early morning period (from 0(24) to 6 o'clock) are much wider than the intervals for the other times of a day. The physical explanation will be given in the next section.

6.6 Summary

In section 6.3.1, the model building of the relation between load and *THI*, and the variance function model building are based on limited knowledge and derived from a careful empirical study and approximation. Within the class of non-linear models, we look at a set of candidate models each one with a different number of parameters and attach the title "true" model to that candidate model which is nearest to reality. After the model accuracy, stability and criterion function study in Appendix C.1, and C.2, section 6.3.3, and 6.3.4 for the proposed model sets, we have enough statistical evidence to choose f_4, and h_2 as an optimal model for the load/weather relation and variance function from the proposed model sets based on locally robust weighted smoothed data. Nevertheless, it may not be appropriate to claim that f_4, and h_2, with their parameters Θ and Ψ based on the smoothed data, are part of the correct model, since there are many ways of producing smoothed data. In section 6.4, to reduce the outliers effects on the estimation of the proposed models for our data, a proposed robust-weighted nonlinear least squares estimation procedure with initial estimates of Θ and Ψ is applied to the real data to obtain the refined estimators of Θ and Ψ. The final estimated parameters Θ for f_4 and Ψ for h_2 are listed in Table 6.3 on page 189.

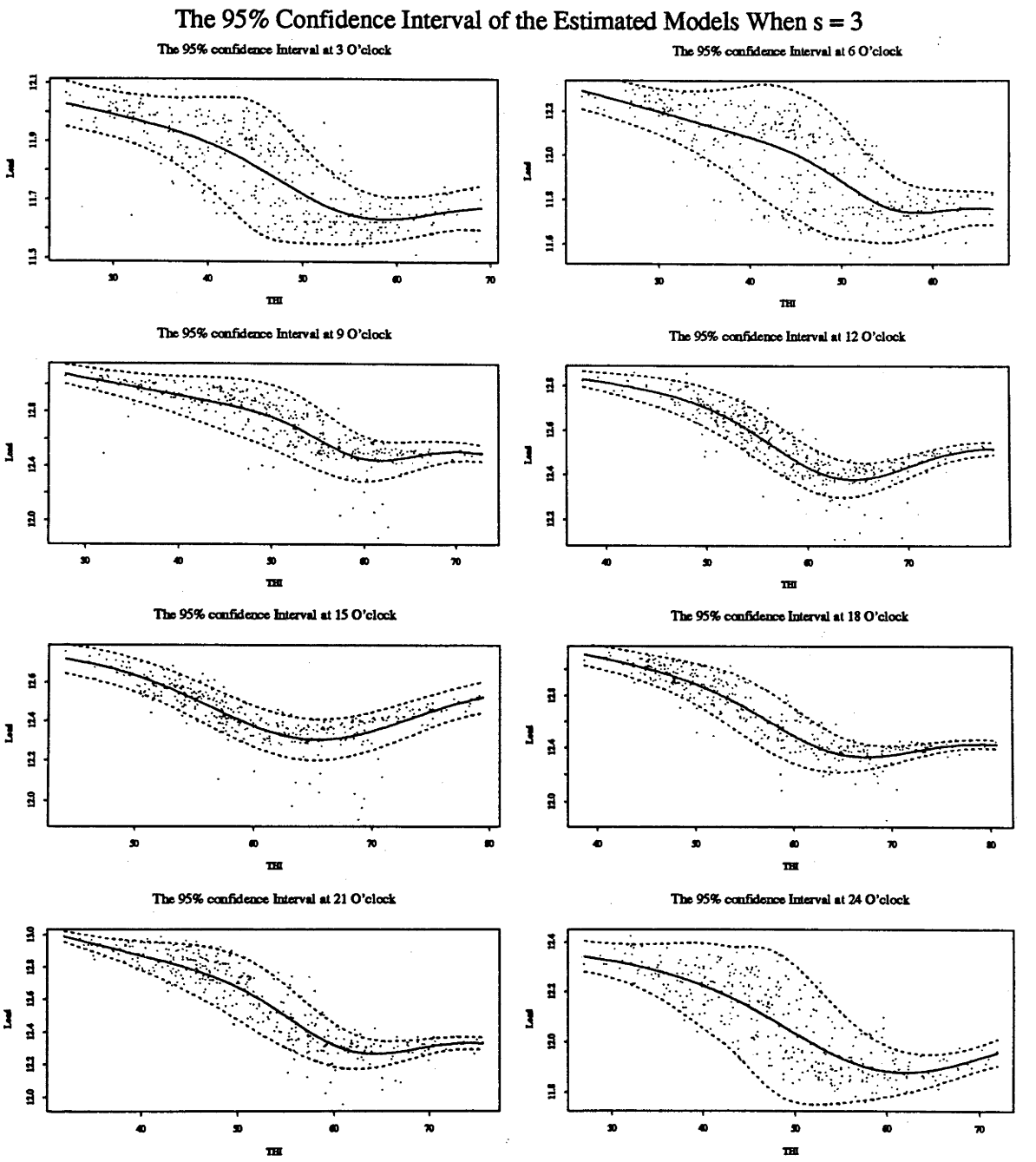


Figure 6.8: 95% Confidence Interval of the Estimated Model f_4

Utilization of the estimated model f_4 leads to decomposition of the load into a weather sensitive load component and a weather insensitive component. Since the weather insensitive load should be the minimum value of f_4, we set the first derivative of f_4 with respect to x equal to zero to obtain the expression for the weather insensitive load.

$$\begin{aligned}\frac{\partial f_4}{\partial x} &= B - 2C\frac{(x-D)}{E}e^{-(x-D)^2/E} \\ &= B - 2C\frac{(x-D)}{E} + O((x-D)^3/E^2) = 0 \\ &\Rightarrow \quad x \doteq D + \frac{BE}{2C}\end{aligned}$$

The selected model function f_4 can be re-written as

$$\begin{aligned}f_4: \; y(x) &= \{A + B(D + \frac{BE}{2C}) + C\,e^{-(\frac{BE}{2C})^2/E}\} \\ &\quad + \{B(x - D - \frac{BE}{2C}) + C\,(e^{-(x-D)^2/E} - e^{-(\frac{BE}{2C})^2/E})\} \qquad (6.31)\end{aligned}$$

Recalling the additive model (6.1), the model function f_4 decomposes the deterministic part of the load into the weather insensitive basic load and the weather sensitive load. It is obvious that the first part of (6.31) $\{A + B(D + \frac{BE}{2C}) + C\,e^{-(\frac{BE}{2C})^2/E}\}$ is weather unrelated, ie. the weather insensitive load; the second part of (6.31) $\{B(w - D - \frac{BE}{2C}) + C\,(e^{-(w-D)^2/E} - e^{-(\frac{BE}{2C})^2/E})\}$ is the weather sensitive load. The consistent negative values for B in model f_4 at different times indicate that the weather sensitive load is more sensitive to cold weather than to hot weather.

In order to explain explicitly the influences of weather conditions (THI) on the load, the profiles of the total load, the weather insensitive load and the weather sensitive load, we define the following terminology:

- *Non-weather Related Load THI*: the *THI* value at which the load has no weather sensitive load component.
- *Maximum Variance THI*: the *THI* value at which the variance of the error term reaches maximum.
- *Peak Load*: the load which is larger than the load at adjacent times of a day.

- *Peak Load Time*: the time of a day corresponding to the peak load.

The *Non-weather Load THI* and and *Maximum Variance THI* from model f_4, are listed in the 7th(vertex) column of the first and the second part of Table 6.3 on page 189. Figure 6.9 displays the profiles of the *non-weather related load THI* and the maximum variance *THI*. It can be seen that the *non-weather related load THI* changes over the times of a day. For our three hour interval weather information data, the highest *non-weather related load THI* occurs at 18 o'clock, the lowest occurs at 6 o'clock, and the values of *THI* range from about 58 to 67 *THI*. Recall Galiana's non-linear relation between load and temperature in Table 6.1, where it shows the *non-weather related load temperature* is between 60 to 70 Fahrenheit. Comparing with the estimated proposed model f_4, it is obvious that Galiana's non-linear relation is only a rough approximation to the load and temperature relation since it ignores the time factor in the relation. The time factor plays an important role in the relation between load and temperature since it is also an index variable of human life patterns and activity levels over a day. Figure 6.9 plots the profile of the *THI* associated with a weather insensitive load and the profile of the *THI* at which the variance function reaches maximum for three hour intervals of a day. From this figure, we can see that *non-weather related load THI* is higher during working hours (9 to 17 o'clock) and lower in the remaining hours especially in the very early morning. This fact indicates that the *non-weather related load THI* is related directly to social and family activities and it may also be associated with the normal human body temperature patterns over a day, a matter which will not be pursued further here. Figure 6.9 also shows that the *maximum variance THI* is roughly parallel to the *non-weather related load THI* and is lower by about 12 *THI* [1]. This shows that the most variability in the weather sensitive load occurs at a *THI* lower than the *non-weather related load THI* by about 12*THI*, and implies that people change their consumption pattern of electric heating more uncertainly when *THI* is lower than the *non-weather related load THI* by about

[1]This fact can also be observed if we look at the 95% confidence interval of the estimated model function in Figure 6.8. The widest 95% confidence interval occurs at about 12 *THI* below the *THI* at which the load reaches the lowest level at all three hour intervals of a day.

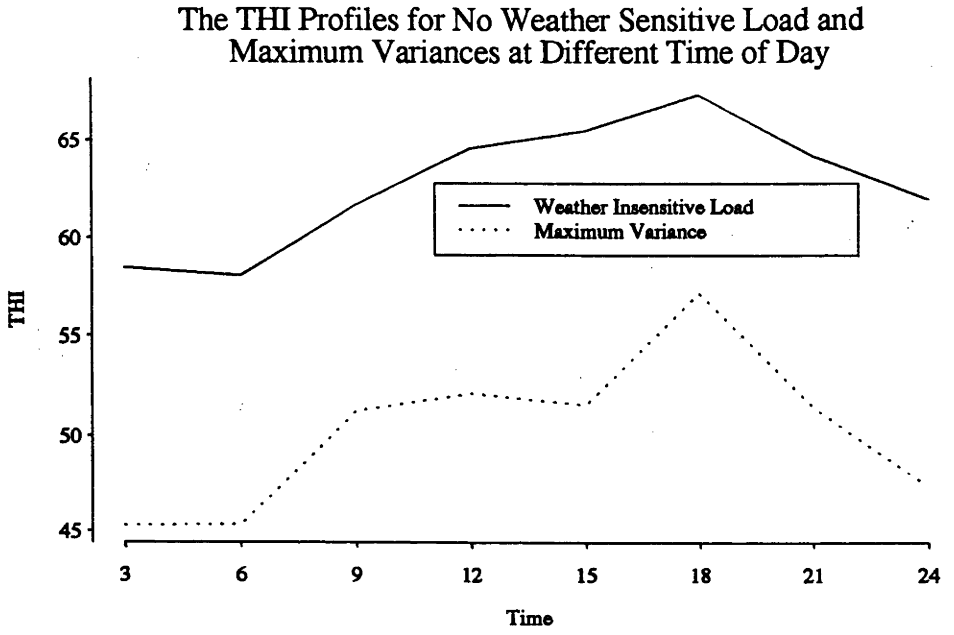


Figure 6.9: THI Profiles of the Non-weather Related Load and Maximum Variances

$12\,THI$ although people are more sensitive to cold weather than hot weather (since $B < 0$ at all times of a day).

On the other hand, the disturbance variances do not increase when the temperature is higher than the the *non-weather related load THI*. One explanation is that people may keep their consumption pattern for electric cooling steadier when *THI* is higher than the *non-weather related load THI*. Another explanation is that the capacity of household electric cooling appliances may be much smaller than the capacity of household electric heating appliances. The use of cooling appliances, therefore, does not produce as wide disturbance variances on the load as heating appliances do. The two explanations are speculations and cannot be verified unless information about capacity of household cooling and heating appliances becomes available.

Figure 6.10 shows the load profile, the variance function profile, the weather sensitive load profile and the weather insensitive load profile. The perspective graph of the load profile (see The Load Profile in Figure 6.10) displays how the weather condition(THI) and time of a day jointly affect the total load. Roughly speaking, the

load peak of a day occurs at 18 o'clock, 9 o'clock and 15 o'clock when the temperatures are lower, close to, or higher than the comfortable temperatures at each time respectively.

The variance function profile (see The Variance Function in Figure 6.10) shows the effects of *THI* and time to the disturbance variances at three hourly intervals of a day. The largest variances tend to be around the temperature below 12 *THI* of the *non-weather related load THI* at each three hourly intervals of a day, and the variances tend to be large in early morning and in the evening rather than during the day.

The weather sensitive load profile (see The Weather Sensitive Load Profile in Figure 6.10) shows that the weather sensitive load is not only dependent on *THI* but also on the time of a day. When the *THI* values are lower than the *non-weather related load THI*, the heating load is concentrated at 9 o'clock and 18 o'clock; when *THI* are higher than the *non-weather related load THI*, the cooling load is concentrated at 9 to 15 o'clock (may be at 14 o'clock since maximum temperature of a day usually occurs at 14 o'clock. Greater detail is available if the weather data is measured for a shorter interval. It is obvious that the load peaks are closely associated with weather sensitive load peaks.

The weather insensitive load profile (see The Weather Insensitive Load Profile in Figure 6.10) presents the load which excludes the additional heating or cooling components(weather sensitive load) from the load. As expected, the weather insensitive load has two peaks at 9 o'clock(first *peak time*) and 18 o'clock(second *peak time*).

Figure 6.11 portrays the details of the *load peak* location and the weather sensitive peak location respectively since these details cannot be seen in perspective from Figure 6.10. The thin solid curves represent the contours of the load or the weather sensitive load against the time of a day and *THI*. It can been seen that there are, at most, two *load peaks* and two *peaks of weather sensitive load*. The thick solid and dot curves represent the first and second peaks of the load and the weather sensitive load where first and second mean the largest and second largest peaks instead of early and

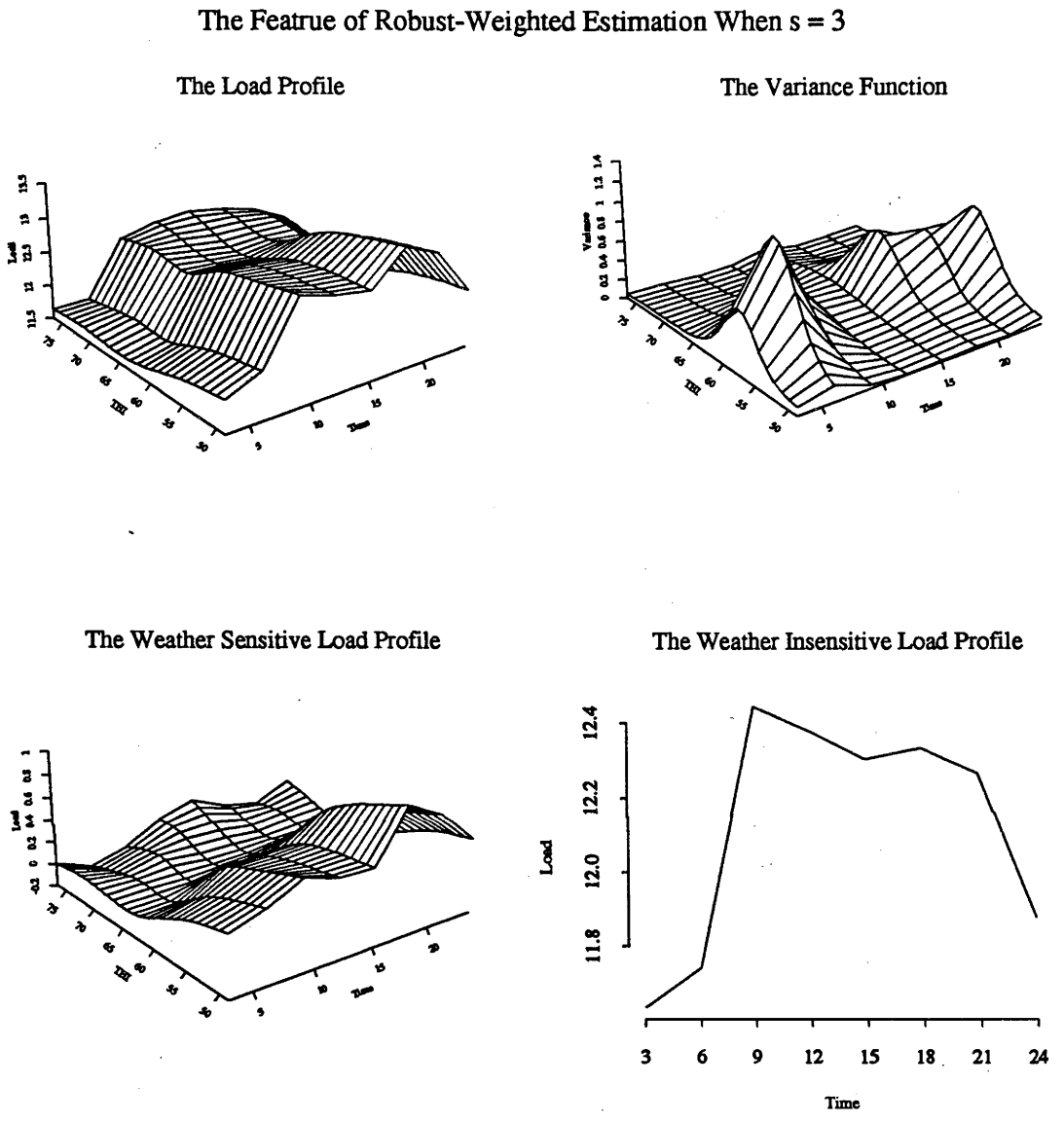


Figure 6.10: Profiles of Load — Decomposition and Variance Function

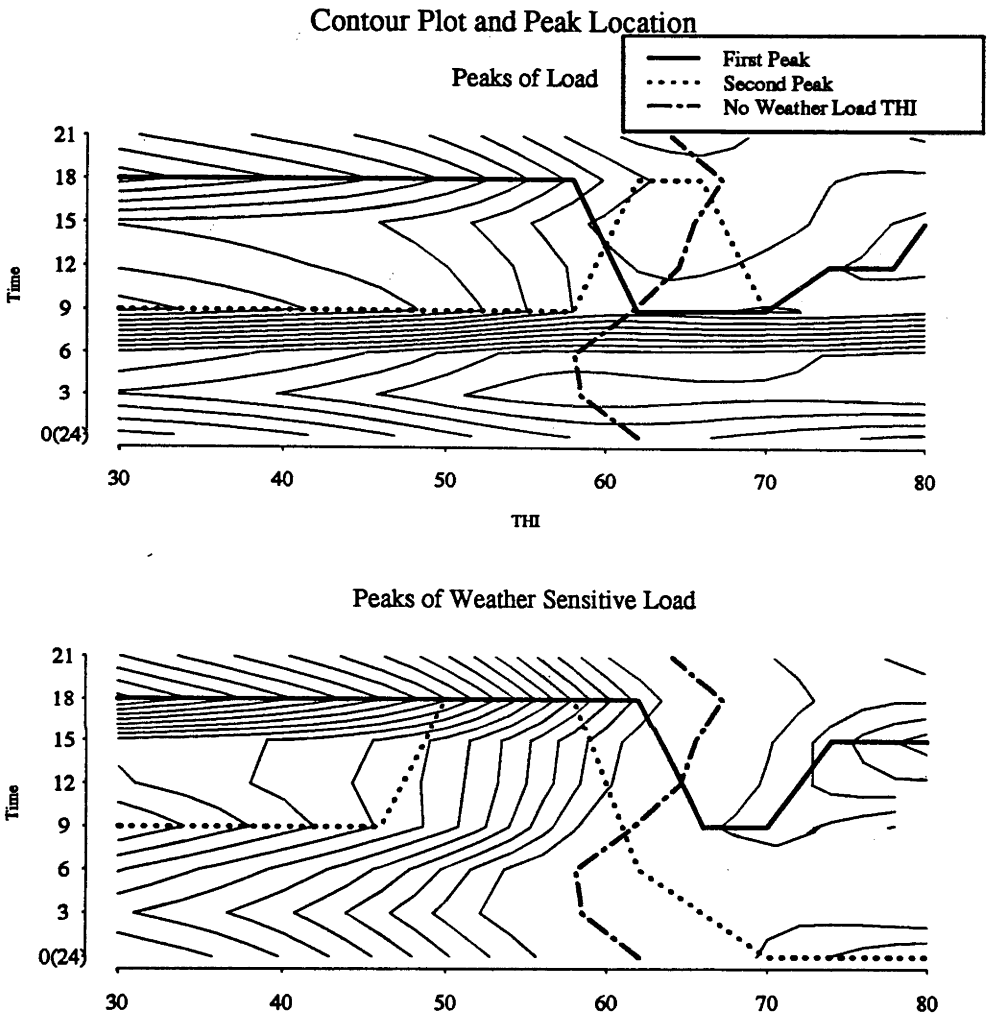


Figure 6.11: Load and Weather Sensitive Load Peaks

later peaks. The thick dash curves represent the *non-weather related load THI* at different time of a day. The thick solid and dot curve illustrate the evolution of first and second peaks due to the changes of time of day and *THI*. It is noted that the variation of the first and second *peak time* are right across the non-weather related load *THI* curve(the thick dash curve). This illustrates how the cool and hot weather conditions affect the peak time.

By noting that temperature changes over a day, and that the changing patterns

are different from day to day, we, therefore, randomly chose four days from different seasons to illustrate the relation between the *peak load time* and temperature pattern of a day in Figure 6.12. The thin curves represent the temperature profile of the days. The intersections of temperature curves and peak curves are the expected *peak load time*. In our example, it is clear that the first *peak load time* is not the time of the lowest temperature in a day of winter or spring(temperature is lower than the *non-weather THI* at each time); however, the first *peak load time* is consistent with the highest temperature in a day of summer(temperature is higher than the *non-weather THI* at each time). This fact indicates that cooling load responds to the hot temperature quickly and there is no time lag as far as the three hour data is concerned. However, the heating *load peak* seems not to have a certain fixed pattern(see the graph of peaks of weather sensitive load in Figure 6.12). The first heating load peak times are always at 18 o'clock when the temperature is lower than the *non-weather load THI* (i.e. there is no cooling load). The second heating load peak time will be at 9 o'clock when temperature is really cold(below 45 *THI*). This heating load peak can be explained as follows: When the weather is really cold at 9 o'clock, extra heating is needed to warm buildings (such as offices, shopping centres etc.) before temperatures rise outside. There will be a second cooling load peak when temperature is higher than 70 *THI* at midnight(it will rarely happen). This cooling peak can be explained as follows: When *THI* (temperature and humidity) is too high at midnight, extra cooling is needed to help people to sleep.

On the other hand, we are also concerned with the *load peaks* which are jointly affected by the weather insensitive load and the weather sensitive load; the first and second weather insensitive load peak times are 9 o'clock and 18 o'clock, respectively(see graph of the weather insensitive load profile in Figure 6.10). However, for the weather sensitive load, the peak load time changes with weather conditions. The first and second peak load times are at 18 o'clock and 9 o'clock respectively when *THI* is lower than 60; the first and second peak load exchange their times when *THI* is close to the *non-weather related load THI*; there is only one load peak when *THI*

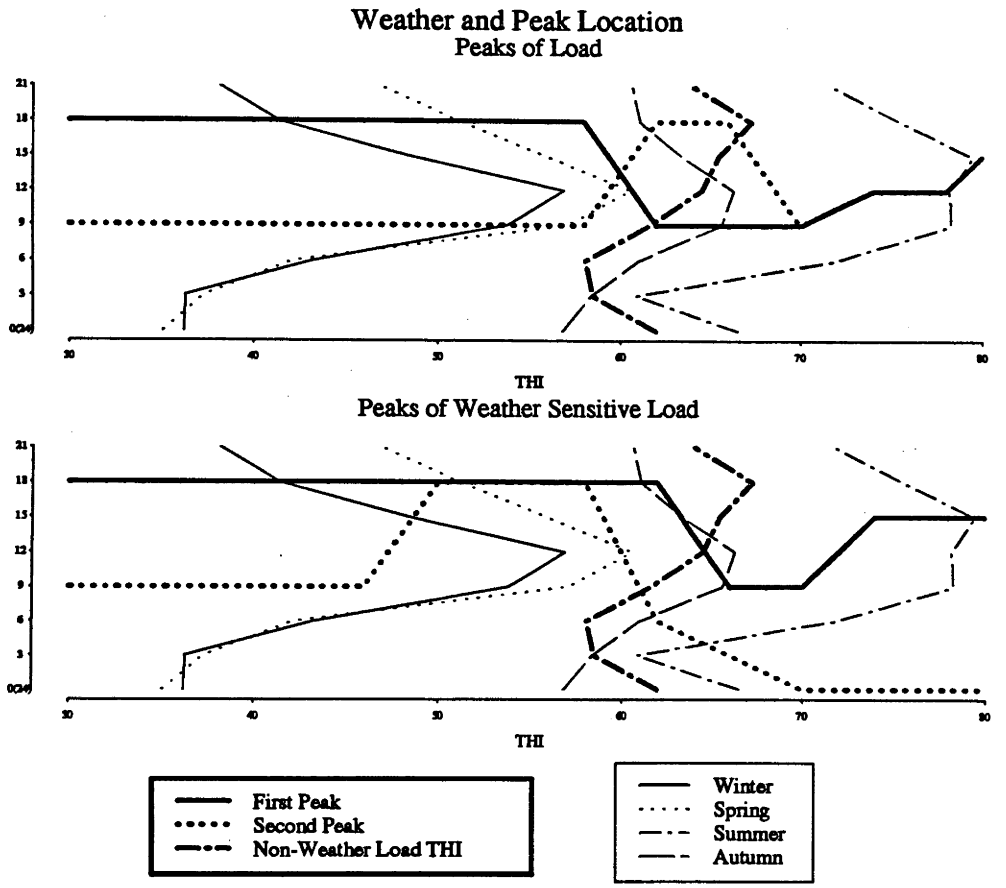


Figure 6.12: An Example of Weather Patterns and Corresponding Peaks

is higher than 70.

From the weather sensitive load peak discussed in the last paragraph, we conclude that the first load peak at 18 o'clock is mostly contributed to by the heating load when *THI* is lower than 60; the first load peak at 9 o'clock is mostly contributed to by the weather insensitive load when *THI* is between 60 and 70; the first load peak at 12 to 15 o'clock is mostly contributed to by the cooling load. The evolution of the load peak time, for both weather insensitive and weather sensitive load, can be summarized in Table 6.4 where the boxes indicate the consistency between the load peak and the weather insensitive load or the weather sensitive load. It can be seen that the load peak times are consistent with the weather sensitive load peak time when temperature is lower than 60 *THI* or higher than 70 *THI*.

THI	Order	Load Peak Time	Weather Insensitive Load Peak Time	Weather Sensitive Load Peak Time
< 45	1st	18	9	18
	2nd	9	18	9
$45 \rightarrow 60$	1st	18	9	18
	2nd	9	18	
$60 \rightarrow 70$	1st	9	9	$18 \rightarrow 9$
	2nd	18	18	$9 \rightarrow 0$
$70 \rightarrow 80$	1st	$9 \rightarrow 15$	9	$9 \rightarrow 15$
	2nd		18	0

Table 6.4: Relation among the Peaks

It seems likely that the business and industry load dominate the overall load during working hours, and domestic load occupies a large proportion of the load in the early morning and night. The domestic heating load, therefore, leads the load peak at 18 o'clock when temperature is lower than 60 *THI*. On the other hand, the business and industry heating load during the day time is not so sensitive to low temperature. Nevertheless, it appears that the business and industry cooling load is so sensitive to hot weather and the cooling load leads to the load peak at 12 to 15 o'clock when *THI* is higher than 70.

On the other hand, the 95% confidence intervals (see Figure C.6) from late morning to early evening are much narrower. It can be explained that the load is dominated by business and industry during day time and that source has a consistent pattern although changing with the temperature, since cooling or heating devices are thermostatically controlled. In the early evening, the domestic heating or cooling electric load is also consistent with temperature due to "building effects"[2]. For instance, in summer, the temperature in dwellings is still higher than outside in the evenings, therefore people use their air-conditioners to ensure the temperature in their homes is brought down to a comfortable level.

In winter, people use electric heating appliances to warm their home for sleeping, and some of them turn off the heating appliances after sleep, some of them have the heating appliances controlled by thermostats. The heating appliances which are controlled by thermostats will turn on more frequently since the temperature is cooler in the early morning. As a result, the changes in electric heating and cooling load is more consistent with temperature before midnight than the changes after midnight and in the early morning.

The above analyses suffer a little since only three hourly interval weather information is available for analysis. Therefore, the load, weather insensitive, and weather sensitive load peak times are only as accurate as is possible with data based on three hourly intervals. As we see in Table 6.3 the parameter Θ for f_4 and Ψ for h_2 are time related. The interval between the available weather information is just too large to establish clearly how Θ and Ψ change with time. Θ and Ψ can be modelled as functions of time, and so, the pattern of the peak times can be refined more effectively if shorter time interval weather information is available.

After closely examining the load and temperature relation model f_4, and the decomposition of the load into weather insensitive and weather sensitive load by utilizing f_4(see equation (6.31)), it is not difficult to find that the stochastic term of model f_4 should comprise two parts. One part is contributed to by the stochastic

[2]The building effect is described as a phenomenon that the change of temperature in a building is always naturally delayed behind the change of temperature outside.

behaviour of the weather insensitive load. Another part is derived from the stochastic behaviour of the weather sensitive load. If we can assume that the two stochastic parts are independent, the variance function for the stochastic term in model f_4 is the sum of the variance functions of the two stochastic parts. Therefore, the real confidence interval of weather sensitive load should be narrower than f_4's (see equation (6.30)). Model f_4 ignores the stochastic nature of the weather insensitive load, and treats it as a deterministic function of time which may not be the case in reality. So, the stochastic behaviour of the weather insensitive load neeeds to resolve. For this reason, model f_4 may not predict future load accurately even when future weather forecasting information is available.

Nevertheless, f_4 provides a non-linear transformation between load and weather information(THI) since model f_4 fits the non-linear relation between the load and temperature data. Utilization of this non-linear transformation converts weather information data into a weather sensitive load variable which is linearly related with the load data and can also be predicted by the second part of equation (6.31) from weather information. Using this variable as an exogenous variable in a linear system which describes the stochastic behaviour of the weather insensitive load, such as an ARX or a State Space model, etc. It is expected that in this way the load prediction can be made more accurate than a prediction which uses f_4 only.

Chapter 7

Dynamic Models for Daily Load

7.1 Introduction

In this chapter, we are going to use dynamic modelling approaches to model the daily electricity demand. The load data and the corresponding weather information in Canberra region, Australia from January, 1985 to July, 1988 have been provided for our study.

We suppose that the total load is the sum of the weather insensitive load (or base component) and the weather sensitive load components. As mentioned in chapter 1, most two stage models assume that the weather insensitive load component is deterministic and the stochastic nature of the total load lies in the weather sensitive load component. This assumption simplifies the model structure because only one stochastic error term is included. In our opinion, this assumption is not realistic because it ignores the dynamic nature of the weather insensitive load component, as would arise for example from an evolving trend, seasonal and weekly periodic components.

Since only three years of daily data is available, there is not enough data to correctly model those "global" long term components, such as the long term trend and the annual seasonal behaviour. However, there is evidence that the annual seasonal behaviour is mainly contributed to by the seasonal changes of climate and the annual Christmas and New-Year holiday period. Therefore, providing weather forecasts are

available, our models will put emphasis on a "local" model instead of a "global" model. In other words, we are interested in using a reasonably small sample data set to model the load approximately and thus obtain reliable short run forecasting.

Two dynamic models, an ARMAX and a structural state space model, and the associated model building, identification, and hypothesis testing are presented in this chapter.

7.2 Weather Sensitive Load Variable

In the first instance, we are interested in the relationship between the load and the corresponding weather conditions and so to extract the weather sensitive component, since we assume that the load is dependent on the weather conditions. A linear regression model has been employed to fit the "dependent" load against the "explanatory"(or exogenous) weather variables. The measure of fit will serve as an indication of how well the "dependent" load can be linearly explained by the "explanatory" weather variables, and the model provides a description of the salient features of the "dependent" load. During the Christmas–New Year holiday period there will be an effect which alters the profile of the weather insensitive component completely and does not have a markable impact on the profile of the weather sensitive component. There is no similar impact on the weather insensitive load profile for any other periods of the year. We therefore use a sample set from January 6, 1985 to December 11, 1985 to model the load/temperature relationship to avoid the possible model mis-estimation arising from mis-specifying the effect of the holiday period.

Among the different linear models with various weather related explanatory variables, we find that the daily maximum temperature(x_1), minimum temperature(x_2), evaporation(x_3) and weekend dummy variables for Saturday(sat) and Sunday(sun) have significant effects on the "dependent" load. The regression model is as follows

$$y(t) = \alpha_0 + \alpha_1 t + \alpha_2 sat + \alpha_3 sun + \alpha_4 x_1 + \alpha_5 x_2 + \alpha_4 x_3 \tag{7.1}$$

with the results listed in Table 7.1.

Residual Standard Error = 0.0949, R^2 = 0.8615				
	coef	std.err	t.stat	p.value
α_0	12.5539	0.0231	542.3437	0
α_1	-0.136E-3	0.5033E-4	-2.7020	0.0072
α_2	-0.1044	0.0144	-7.2376	0
α_3	-0.1448	0.0145	-9.9910	0
α_4	-0.0197	0.0013	-15.5043	0
α_5	-0.0149	0.0013	-11.7795	0
α_6	-0.0107	0.0020	-5.2382	0

Table 7.1: Linear Regression — Daily Load on Weather Variables

The value of the Multiple R-Square in Table 7.1 leads us to conclude that 86.15% of the variability in the load can be explained directly by the included weather variables and weekly periodic dummy variables. This model assumes that the daily weather insensitive load for weekdays is identical (actually the differences between the daily load for weekdays are insignificant), and the weekly periodic component of the load is supposed to be deterministic. The model does not fit well and the estimated trend slope α_1 is quite misleading since the trend slope is expected to be positive.

Recall that in the last chapter, the relationship between the three hourly electricity load and the relevant temperature is nonlinear. Therefore, it may not be appropriate that the daily maximum and minimum temperature serve as regressors directly in the regression model. In a similar way, the analysis in the last chapter is used to find that the nonlinear function, of the f_4 form, is still suitable for the relation between the load and the maximum temperature, or the minimum temperature, i.e.

$$y(x_i) = A_i + B_i x_i + C_i e^{-(x_i - D_i)^2/E_i} \quad (i = 1, 2) \tag{7.2}$$

The estimated model parameters are listed in Table 7.2. The estimated weather sensitive load variables W_1, W_2 from the maximum and minimum temperatures can be derived from equation (6.31) respectively as

$$W_i(x_i) = B_i\,(x_i - D_i - \frac{B_i E_i}{2C_i}) + C_i\,(e^{-(x_i - D_i)^2/E_i} - e^{-(\frac{B_i E_i}{2C_i})^2/E_i}) \quad (i = 1, 2) \tag{7.3}$$

Because the transformed maximum and minimum temperature variables, W_1, W_2

Daily Load & Maximum Temperature					
Parameters	A_1	B_1	C_1	D_1	E_1
	12.16	-0.02	-0.30	23.34	72.72
Variance	0.00	0.00	0.00	0.24	286.14
Daily Load & Minimum Temperature					
Parameters	A_2	B_2	C_2	D_2	E_2
	12.21	-0.02	-0.22	11.12	44.98
Variance	0.00	0.00	0.02	2.31	1609.84

Table 7.2: Nonlinear Relationship Between the Daily Load and Temperature

are approximately linearly related to the daily load, the linear regression model (7.1) is modified by replacing x_1, x_2 with W_1, W_2 as follows

$$y(t) = \beta_0 + \beta_1 t + \beta_2 sat + \beta_3 sun + \beta_4 W_1 + \beta_5 W_2 + \beta_6 x_3 \tag{7.4}$$

Using the least squares estimator, we have the results listed in Table 7.3.

Residual Standard Error = 0.0693 , R^2 = 0.9262				
	coef	std.err	t.stat	p.value
β_0	11.7744	0.0139	844.9950	0
β_1	0.4053E-4	0.3536E-4	1.1462	0.2525
β_2	-0.1035	0.0105	-9.8285	0
β_3	-0.1438	0.0106	-13.6123	0
β_4	0.7458	0.0288	25.9351	0
β_5	0.3283	0.0299	10.9655	0
β_6	-0.0087	0.0015	-5.9214	0

Table 7.3: Linear Regression — Daily Load on Transformed Weather Variables

Comparing the values of R-squared for the above two regression models, the latter model is obviously superior. This suggests that the nonlinear relation between the temperature and the load is essential to achieve the improvement. We can create a new weather sensitive variable W by forming[1],

$$W = \beta_4 W_1 + \beta_5 W_2 + \beta_6 x_3 \tag{7.5}$$

[1]The weather sensitive variable is a function of weather conditions instead of time although the weather conditions are related to time. We denote W as the weather sensitive variable and treat it as a constant with respect to time if the corresponding weather conditions do not change.

and then regress $y(t)$ on a linear time trend, W, and weekly periodic dummy variables *sat*, *sun* again to obtain a linear regression model

$$y(t) = \beta_0 + \beta_1 t + \beta_2 sat + \beta_3 sun + \beta_W W \tag{7.6}$$

which assumes the response consists of a deterministic (level component + trend component + weekly periodic component + weather dependent component) part and a stochastic part. This model is clearly equivalent in structure to model (7.4) and the least squares estimation results are listed in Table 7.4.

Residual Standard Error $\hat{\sigma}^2_{reg} = 0.0693$, $R^2 = 0.9262$				
	coef	std.err	t.stat	p.value
β_0	11.7744	0.0079	1491.8365	0
β_1	0.4053E-4	0.3536E-4	1.1511	0.2525
β_2	-0.1035	0.0105	-9.8285	0
β_3	-0.1438	0.0106	-13.6123	0
β_W	1.0	0.01569	63.7304	0

Table 7.4: Linear Regression — Daily Load on Transformed Weather Variables

However, the deterministic assumption for trend and weekly periodicity may not be realistic because it ignores the stochastic nature of the trend and weekly periodic components. These are possible reasons which limit the ability of the regression model to model the data.

We have discussed the conventional additive model and adaptive additive model in chapter 5, and explained why an adaptive additive model may achieve a better fit. To incorporate the external weather information into an adaptive additive model to describe the behaviour of the daily electricity load, we employ two major adaptive models, the ARMAX and the State Space models, in the following sections.

7.3 Application of ARMAX Model

In the linear dynamic system literature, a linear system with observed output $y(t)$ and observed input $u(t)$ has been modelled in the time domain for different purposes

by many researchers in statistics, econometrics, and systems engineering, etc. One of the model types is called an ARMAX model and has the following representation for the univariate case,

$$\sum_{j=0}^{p} a(j)y(t-j) = \sum_{j=0}^{q} b(j)\epsilon(t-j) + \sum_{j=0}^{r} c(j)u(t-j) \tag{7.7}$$

where $a(j)$, $b(j)$, $c(j)$ are unknown constants; the roots of $\sum_{j=0}^{p} a(L) = 0$ are outside the unit circle and $\epsilon(t)$ is the disturbance term, where it is assumed that $\mathbf{E}(\epsilon(t)) = 0$, $\mathbf{var}(\epsilon(t)) = \sigma^2$, $\forall t$, and $\mathbf{E}(\epsilon(s), \epsilon(t)) = 0$, $s \neq t$.

The basic idea of the ARMAX model is to use the dynamic behaviour of an exogenous variable $u(t)$ within an ARMA model to describe a time series in a more effective way than the usual ARMA model. In an ARMAX model, there are two stochastic exciting forces entering the system; through the disturbance $\epsilon(t)$ and the exogenous variable $u(t)$. and they are dynamically linked with their response by model (7.7). The explanatory part $\sum_{j=0}^{r} c(j)u(t-j)$ is generated by the exogenous variable and produces an exogenous dependent component of $y(t)$, and the disturbance part $\sum_{j=0}^{q} b(j)\epsilon(t-j)$ produce a further stochastic component of $y(t)$.

The above ARMAX model works well in a wide range of circumstances; its properties have been explored in the time series and linear systems literature, and will not be repeated in this thesis. The maximum likelihood estimation procedure will be used to estimate the parameters of an ARMAX model.

One of ARMAX models for nonstationary periodic series which can be parsimoniously structured is called a multiplicative periodic ARIMAX model with a form which is similar to a Box and Jenkins multiplicative seasonal model, and is specified as follows,

$$\begin{aligned} \phi_p(B)\Phi_{P_1}(B^{s_1})\Phi_{P_2}(B^{s_2})\Delta^d\Delta_{s_1}^{D_1}\Delta_{s_2}^{D_2}y(t) &= \theta_q(B)\Theta_{Q_1}(B^{s_1})\Theta_{Q_2}(B^{s_2})\epsilon(t) \\ &\quad + \sum_{k=1}^{K} \lambda_k(B)u_k(t) + \rho h(t) \end{aligned} \tag{7.8}$$

where it is assumed that $\mathbf{E}(\epsilon(t)) = 0$, $\mathbf{var}(\epsilon(t)) = \sigma^2$, $\forall t$, and $\mathbf{E}(\epsilon(s), \epsilon(t)) = 0$, $s \neq t$. K is the number of exogenous variables $u_k(t)$ $(k = 1, \cdots, K)$; $h(t)$ is the Christmas and

New-Year dummy variable; s_1 and s_2 are constants to represent the annual seasonal period and a weekly periodic effect, respectively. The symbol $\Delta = 1 - B$ represents the difference operator; $\Delta_{s_1} = 1 - B^{s_1}$ the annual seasonal difference operator; $\Delta_{s_2} = 1 - B^{s_2}$ the weekly periodical difference operator; ϕ_p, Φ_{P_1}, Φ_{P_2},θ_q, Θ_{Q_1}, Θ_{Q_2}, λ_k are polynomial functions with order p, P_1, P_2, q, Q_1, Q_2 and $r(k)$ respectively; and all the roots of $\phi_p(L) = 0$, $\Phi_{P_1}(L) = 0$ and $\Phi_{P_2}(L) = 0$ are outside the unit circle.

Among the many proposed ARIMAX models, and for the different combinations of exogenous variables $u_k(t)$, e.g. daily maximum and minimum temperatures, evaporation, wind speed, and the weather sensitive load variable W etc., we find that

1. the coefficients of $\Phi_{P_1}(B^{s_1})$ and $\Theta_{Q_1}(B^{s_1})$ are insignificant i.e. $\Phi_{P_1}(B^{s_1}) = 1$, $\Theta_{Q_1}(B^{s_1}) = 1$, and $D_1 = 0$ because the annual seasonal component of the load are mainly contributed to by the exogenous weather variables

2. the multiplicative periodic ARIMA$(1,1,1) \times (0,1,1)_7$ with the exogenous variable W is the optimal model [2] with respect to the model selection criterion AIC, i.e. we have a specification for sample data sets which do not include Christmas and New-Year period,

$$(1 - \phi B)(1 - B)(1 - B^7)y(t) = (1 - \theta B)(1 - \Theta B^7)\epsilon(t) + \lambda W \qquad (7.9)$$

For the daily data from January 6, 1985 to December 11, 1985, the model ARIMAX$(1,1,1) \times (0,1,1)_7$ has been identified by maximum likelihood estimation as

$$\begin{aligned}(1 - 0.547B)(1 - B)(1 - B^7)y(t) &= (1 - 0.867B)(1 - 0.833B^7)\epsilon(t) \\ &\quad +0.469W \qquad (7.10)\end{aligned}$$

The model diagnostic checks are plotted in figures 7.1 and 7.2 on page 241, 242. Comparing the estimated disturbance variance of the regression model (see Table 7.4) and the above ARIMAX$(1,1,1) \times (0,1,1)_7$ model, we can clearly see the latter model achieves a significant improvement.

[2] All experimental models and their results are not presented in this thesis. The copy of all the results can be obtained from the author on request.

AIC = -1201.575, σ^2 = 0.001468296			
COV	ϕ_1	θ_1	Θ_7
ϕ_1	0.008417665	0.004783661	-0.002104029
θ_1	0.004783661	0.003631949	-0.001621765
Θ_7	-0.002104029	-0.001621765	0.001648345

Table 7.5: Covariance Matrix of Estimated Parameters of ARIMAX$(1,1,1)\times(0,1,1)_7$ Model

The attraction of the ARIMAX class of models is that they provide a general framework for forecasting time series with the specification of a model within the class is determined by the data. On the other hand, to view all the models within that class as potential candidates and then to select the best one is ineffective. The selected model in general is unquestionably an arbitrary one and may not be appropriate unless one has *a priori* knowledge of the models which are likely to be most useful.

7.4 State Space Model

A commonly used time invariant state space model has the following form

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) + \delta(t) & \text{Transition equation} \\ y(t) = Cx(t) + Dv(t) + \epsilon(t) & \text{Observation equation} \end{cases} \tag{7.11}$$

where $x(t)$ is called the state vector which may unobservable and may appear to be only tenuously connected with the data $y(t)$. The first equation is called the state transition equation. The second equation is the observation equation which specifies the relation between the data and the newly introduced auxiliary state vector $x(t)$. $u(t)$, and $v(t)$ are exogenous input vectors. $\delta(t)$, and $\epsilon(t)$ are disturbance terms for the transition equation and the observation equation. It is assumed that the disturbance term $\delta(t)$ is a noise vector with zero mean and fixed finite covariance matrix; $\epsilon(t)$ is also white noise with zero mean and fixed finite variance.

The prominent advantage in using the state space model is that one can directly apply the Kalman filter which can produce stable dynamics even when the original

dynamics are unstable when certain conditions are met. These conditions are intensively discussed in Anderson and Moore (1979), Caines (1988), Aoki (1987), and Harvey (1989), etc. Some extensions and generalizations of these conditions have been developed in chapter 3 of this thesis. Therefore, the properties of the Kalman filter are not discussed in detail here. We concentrate on the application of the state space model and the Kalman filtering to our daily electricity load data.

Once one decides to use a state space model to represent a time series, the first problem one faces is how to construct a suitable model for the time series. In the following section we discuss some state space model construction schemes and their advantages and disadvantages.

7.4.1 State Space Model Construction

Canonical State Space Modelling

Akaike's canonical correlation method (see Akaike (1975), Aoki (1987), Hannan and Deistler (1988)) for constructing a state space model from a data set is based upon the Hankel matrix of the observed data set. The main idea is based on the canonical analysis of the Hankel matrix, and to gather as much information as possible from the observed data using those canonical vectors whose corresponding eigenvalues are significantly different from zero. Therefore, the minimum order of the state space model can be determined by the number of the Hankel matrix eigenvalues which are judged to be significantly different from zero. After determining the minimum order, the system matrix can be calculated by a UV-decomposition of the Hankel matrix (see Akaike (1975), Aoki (1987) for details).

Because the UV-decomposition is not unique, there would be many different minimal dimensional state space representations for the same time series. The difference between the two different minimal dimensional state space representations are the coordinates of the state vector of the two representations. If we assume the system matrices of the two different minimal state-space representations are $\{A_1, B_1, C_1\}$ and $\{A_2, B_2, C_2\}$, and state vectors are $x_1(t)$ and $x_2(t)$, there must exist a non-singular

matrix P satisfying $x_2(t) = Px_1(t)$, $A_2 = PA_1P^{-1}$, $B_2 = PB_1$, $C_2 = CP^{-1}$.

The advantage of this method is that it can directly construct an optimal state space model without evaluating many models and their fit to the data set. The disadvantage is that, first, the dynamic exogenous variables are not easy to include in this state space modelling procedure; secondly, the system matrices are to be estimated as unknown parameters, therefore, this construction scheme may not be effective for prediction, especially, for high order statistical models; thirdly, as far as the observed data is concerned, the created state vector $x(t)$ is an extraneous theoretical construction. This vector may not have any obvious physical explanation. In practice, people are often interested in extracting particular components of the observed data, such as a representation of the trend, or the seasonal components.

Conversion of ARMAX Models to State-Space Models

There is a representation of an ARMAX model in state-space form which particularly connects with the Kalman filter. For example, an ARMAX model with order (p,q,r) has a state-space representation as follows

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) + K\epsilon(t) & \text{Transition equation} \\ y(t) = Cx(t) + \epsilon(t) & \text{Observation equation} \end{cases} \tag{7.12}$$

where

$$A = \begin{pmatrix} -a(1) & -a(2) & \cdots & -a(p-1) & -a(p) & b(1) & \cdots & b(q-1) & b(q) & c(1) & \cdots & c(r-1) & c(r) \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

$$B' = (\underbrace{0, \cdots, 0}, \underbrace{0, \cdots, 0}, \underbrace{1, \cdots, 0})$$

$$C = (\underbrace{-a(1), -a(2), \cdots, -a(p)}, \underbrace{b(1), \cdots, b(q)}, \underbrace{c(1), \cdots, c(r)})$$

$$K' = (\underbrace{1, 0, \cdots, 0}, \underbrace{1, 0, \cdots, 0}, \underbrace{0, \cdots, 0})$$

The above representation is called an observable canonical model since the system matrices A and C are a pair of observable system matrices which guarantee that the state vector covariance matrix converges to a steady state when the initial state vector covariance matrix is set as a positive definite matrix. Of course, the structure of the state space representation for an ARMAX model is not unique. Among the different representations, those with minimal dimension are called *minimal dimension state space representations*. Generally speaking, representations with non-minimal dimension are to be avoided because such representations are over parameterized, i.e. redundant information or irrelevant information is embedded in their state vectors which will affect the efficiency of the optimization computations, and create other technical deficiencies during model filtering, and model identification. Conversion of the more general traditional ARMAX model into a state space representation can be found in Kailath (1980) for example. After models are thus converted, their observability and reachability should always be verified to ensure minimal representation, since these conditions are necessary and sufficient for minimality.

In a similar way, the minimum order for an ARMAX model can be determined by the number of non-zero eigenvalues of the Hankel matrix. Furthermore, Hannan (see (viii) of Theorem 2.5.3, pp 63 of Hannan and Deistler (1988)) proves that the estimation of state space and ARMAX parameters is really the same thing.

Although an ARMAX model has a state space representation which can use the Kalman filter to estimate the model parameters, the true ARMAX model is usually unknown in practice, and therefore its state space representation will be unknown as well. Thus, a state space representation of an ARMAX model can only help in estimating the parameters of the specified model and in smoothing, filtering and predicting through the Kalman filter. It cannot however suggest a suitable structure for the ARMAX models directly. For example, an ARMAX model with over-estimated orders has its minimal dimension state space representation which is larger in size

than would result from an AR MAX model based on the true orders.

Structural State Space Modelling

Based on the assumption that the observed data comprises different components in additive form, and that those components are of interest, Engle (1978), Gersch and Kitagawa (1983), Harvey and Todd (1983) and others propose another mechanism to construct a state space model called the *Structural State Space Model* (SSSM) for economic time series because economists are usually interested in extracting the various components of economic data. The Kalman filter can be used to estimate non-stationary time series models. The SSSM consists of many micro-state-space models for each component, such as, a stochastic linear trend, seasonal, cyclical components, and the disturbance term. The SSSM approach is to choose some suitable micro-state-space models which may describe the behaviour of the different components of the observed data, and combine them into a main state-space model framework. The constructed model should be at least detectable[3] because we are interested in the "energy" distribution of the output among the different components (the state vector). If the main state-space model is both detectable and stabilizable the state covariance matrix converges to steady state and produces a plausible model at the outset.

The major advantage of using the SSSM is that this model can produce the components of interest (the state vector) in the data, and provides sensible forecasts if the model is accepted. While diagnostic checking is common in both structural state space and ARMAX model building, the way in which models are initially specified is quite different. The disadvantage of the SSSM scheme is that the state vector may not necessarily be minimized. A remedy for this problem can be found by simplifying the component models or even dropping some component models from the main state space model framework as long as the model adequacy criteria is met.

The model selection methodology for structural models is more akin to those

[3]the reconstructibility or the observability are sufficient for the detectability

adopted in econometrics where a tentative model is formulated on the basis of knowledge of the nature of the variables and the hypothesized relationship expected between them. If a SSSM appears to be inadequate in diagnostic checking, its specification is changed and the process of estimation and diagnostic checking is repeated. If the model survives diagnostic checking, it is either accepted or an attempt is made to simplify it. Simplification occurs where, for example, small values are set equal to zero, or perhaps by dropping a component completely.

7.4.2 A Structural State Space Model for Daily Electricity Load

As mentioned in the introduction to this chapter and the summary of chapter 1 of this thesis, the main use of a state space model is to model the stochastic behaviour of the weather insensitive load component since the deterministic model, i.e. regression model in the last section, failed to represent the weather insensitive load component well. This conclusion has been supported by the model adequacy diagnostic tests.

In the construction of a SSSM, we assume that

1. The weather insensitive load comprises a stochastic linear trend, weekly periodic components, and a disturbance component.

2. The weather sensitive variable W (see equation (7.5)) is an exogenous variable which does not affect the weather insensitive component.

3. The daily load is the sum of the weather insensitive load and the weather sensitive load which is a linear function of the estimated weather sensitive variable W.

Using the above assumptions, we construct basic structural state space models for the trend and weekly periodic components of the weather insensitive load and the weather sensitive load component.

Local Linear Trend Model

$$\begin{cases} m(t) = m(t-1) + \beta(t) + \eta(t) \\ \beta(t) = \rho\,\beta(t-1) + \zeta(t) \\ y_{trend}(t) = m(t) \end{cases} \tag{7.13}$$

where $m(t)$, and $\beta(t)$ are the level and slope of the trend, respectively; ρ is a damping factor for the slope, $0 \leq \rho \leq 1$, and $\eta(t)$ and $\zeta(t)$ are mutually uncorrelated with $\eta(t) \sim \text{NID}(0, \sigma_\eta^2)$, $\zeta(t) \sim \text{NID}(0, \sigma_\zeta^2)$. The effect of $\eta(t)$ is to allow the level of the trend to shift randomly up and down, while $\zeta(t)$ allows the slope to change in a similar way. The larger the variances of $\eta(t)$ and $\zeta(t)$ the greater the stochastic movements in the trend.

The state space representation of the above local linear trend model is

$$\begin{cases} x(t) \triangleq \begin{pmatrix} m(t) \\ \beta(t) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & \rho \end{pmatrix} x(t-1) + \begin{pmatrix} \eta(t) \\ \zeta(t) \end{pmatrix} \\ y(t)_{trend} = (1,0)x(t) \end{cases} \tag{7.14}$$

This model is a general local linear trend which can produce the following forms as special cases:

IF $\rho = 1$: The trend is a stochastic local linear trend, when $\sigma_\eta^2 > 0$, $\sigma_\zeta^2 > 0$. The forecasting function is $y_{trend}(T+l|T) = \widehat{m(T)} + \widehat{\beta(T)}\,l$

When $\sigma_\zeta^2 = 0$: The slope of the linear trend is a constant. The forecasting function then becomes $y_{trend}(T+l|T) = \widehat{m(T)} + \hat{\beta}\,l$

When $\sigma_\eta^2 = \sigma_\zeta^2 = 0$: The linear trend is deterministic. The forecasting function becomes $y_{trend}(T+l|T) = \hat{m} + \hat{\beta}\,(T+l)$

IF $\rho < 1$: The trend is a stochastic local linear damped trend when $\sigma_\eta^2 > 0$, $\sigma_\zeta^2 > 0$. The forecasting function is $y_{trend}(T+l|T) = \widehat{m(T)} + [(1-\rho^l)/(1-\rho)]\widehat{\beta(T)}$

When $\sigma_\zeta^2 = 0$: The slope of the linear trend is deterministic and exponentially decreasing. The forecasting function becomes $y_{trend}(T+l|T) = \widehat{m(T)} + [(1-\rho^l)/(1-\rho)]\rho^T\hat{\beta}$

When $\sigma_\eta^2 = \sigma_\zeta^2 = 0$: The trend is a deterministic exponential curve. The forecasting function becomes $y_{trend}(T+l|T) = \hat{m} + (1-\rho^{T+l})/(1-\rho)]\rho^T\hat{\beta}$

IF $\beta(t) = 0$: The trend model collapses to a local level model, i.e. a simple random walk plus noise

$$\begin{cases} m(t) = m(t-1) + \eta(t) & \eta(t) \sim \text{NID}(0, \sigma_\eta^2) \\ y_{trend}(t) = m(t) \end{cases}$$

The trend itself follows a random walk. The forecasting function becomes $y_{trend}(T+l|T) = \widehat{m(T)}$

When $\sigma_\eta^2 = 0$: The level is a constant. The forecasting function becomes $y_{trend}(T+l|T) = m$

Weekly Periodic Models

There are two commonly used periodic models,

$$\sum_{j=0}^{s-1} r(t-j) = \omega(t) \text{ or } S(B)r(t) = \omega(t) \tag{7.15}$$

$$r(t) = r(t-s) + \omega(t) \text{ or } \Delta_s r(t) = \omega(t) \tag{7.16}$$

where s is the length of the period; $S(B) = 1 + B + \cdots + B^{s-1}$, $\Delta_s = 1 - B^s$; $r(t)$ is a weekly periodic effect at time t; $\omega(t) \sim \text{NID}(0, \sigma_\omega^2)$. The disturbance term $\omega(t)$ allows periodic pattern changing.

Because $\Delta_s = \Delta S(B)$, where $\Delta = 1-B$, Δ_s and Δ have a unit root in common. It implies that the real periodic component in model (7.16) is confounded with the trend component because the factor $(1-B)$ is also associated with a long-run trend. The sum of the "periodic" effects, which is modelled by (7.16), will not, in general, sum to zero over the periods. Furthermore, in this model, it is not possible to separate the trend and periodic components from the data. Therefore, model (7.16) is not suitable for the purpose of separately modelling the periodic component.

An alternative way of modelling a weekly periodic pattern apart from model (7.15) is by a set of trigonometric terms at the weekly periodic frequencies, $\lambda_j = 2\pi j/s$ (See details in Hannan et al. (1970)).

The weekly periodic effect at time t is

$$r(t) = \sum_{j=1}^{[s/2]} (r_j \cos \lambda_j t + r_j^* \sin \lambda_j t) + \omega(t) \tag{7.17}$$

Provided that the full set of trigonometric terms is included in model (7.17), this form is equivalent to model (7.15) and the estimated weekly periodic patterns will be identical because the homogeneous solution of equation (7.15) has the same form as equation (7.17). Furthermore, if the weekly periodic patterns change relatively smoothly, some higher-order frequencies can reasonably be dropped to reduce the unknown parameter numbers (i.e. reduce the state vector dimension if the model is to be put into a state space representation). As an example see Abraham and Box (1978).

However, our experience with daily electricity load shows that there is little evidence for dropping some higher-order frequencies. The use of model (7.17) is more complicated and has no advantage over model (7.15). Therefore, the periodic model (7.15) which has following state space representation is employed.

$$\begin{cases} x(t) \triangleq \begin{pmatrix} r(t) \\ r(t-1) \\ \vdots \\ r(t-s+1) \end{pmatrix} = \begin{pmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{pmatrix} x(t-1) \\ \qquad\qquad + (\omega(t)\ 0\ \cdots\ 0)' \\ y_{wp}(t) = (1, 0, \cdots, 0)x(t) \end{cases} \tag{7.18}$$

where $\omega(t) \sim \text{NID}(0, \sigma_\omega^2)$ allows for the periodic pattern changing over time.

Weather Sensitive Component Model

In section 7.2, we have constructed a weather sensitive load variable W from a regression model (deterministic additive model). However, this weather sensitive load variable is deterministic and may not reflect the stochastic nature of the weather sensitive load components. In other words, the effect of the weather sensitive load variable W on the load y should be adaptive and allow for change.

To solve this problem, we assume that the stochastic weather sensitive load $u(t)$ satisfies

$$\begin{cases} \alpha(t+1) = \alpha(t) + \xi(t) \\ u(t) = \alpha(t)W(t) \end{cases} \tag{7.19}$$

where $\xi(t) \sim \text{NID}(0, \sigma_\xi^2)$. By introducing a random walk $\alpha(t)$, the $u(t)$ is allowed to change with time. $\alpha(t)$ will be a constant when $\sigma_\xi^2 = 0$.

7.4.3 A Basic Structural State Space Model

Based on the models developed for the weather insensitive trend, weekly periodic components and the weather sensitive component, we establish a basic structural state space model for the daily electricity load as follows

$$\begin{cases} x(t+1) = Ax(t) + \delta(t) & \text{Transition equation} \\ y(t) = C(t)x(t) + \epsilon(t) & \text{Observation equation} \end{cases} \tag{7.20}$$

where

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(j) \\ x_3(t) \\ \vdots \\ x_9(t) \\ x_{10}(t) \end{pmatrix} = \begin{pmatrix} m(t) \\ \beta(j) \\ r_1(t) \\ \vdots \\ r_6(t) \\ \alpha(t) \end{pmatrix}, \quad \delta(t) = \begin{pmatrix} \delta_1(t) \\ \delta_2(t) \\ \delta_3(t) \\ \\ \vdots \\ \delta_9(t) \\ \delta_{10}(t) \end{pmatrix} = \begin{pmatrix} \eta(t) \\ \zeta(t) \\ \omega(t) \\ 0 \\ \vdots \\ 0 \\ \xi(t) \end{pmatrix}$$

$$A(t) = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & \rho & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & -1 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, C(t) = (1, 0, 1, 0, \cdots, 0, W(t))$$

The model parameter set Θ contains six unknown parameters, $(\rho, \sigma_\eta^2, \sigma_\zeta^2, \sigma_\omega^2, \sigma_\xi^2, \sigma_\epsilon^2)$, and the system matrix $C(t)$ is a time varying matrix.

7.4.4 Off-line Parameter Estimation

There are various algorithms in the literature for the model parameter estimation according to different criteria, such as, maximum likelihood, minimum prediction error, etc. The estimation can also be obtained in time domain or frequency domain (see discussion of details in Harvey (1989)). In our study, maximum likelihood estimation in the time domain is employed.

Maximum Likelihood Estimation

As mentioned in chapter 3, Maximum likelihood (ML) estimation for the parameters of the model can be based upon the Kalman filter. It is easy to verify that the basic structural state space model is detectable and stabilizable if and only if σ_ζ^2 and σ_ω^2 are strictly positive which ensures the state variance matrix converges to its steady state value exponentially fast.

The general form of the maximum likelihood function for a univariate state space model can be represented as follows

$$\log\ L = -\frac{T}{2}\log\ 2\pi - \frac{T}{2}\log\ \sigma_\epsilon^2 - \frac{1}{2}\sum_{t=1}^{T}\log f(t) - \frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T} v^2(t)/f(t) \qquad (7.21)$$

where $v(t) = y(t) - \hat{y}(t|t-1)$ and $f(t) = \mathbf{var}[v(t)]$ are obtained from the Kalman filter.

The univariate model can be reparameterized so that $\Phi = [\Phi_*', \sigma_*^2]'$ where Φ_* is a vector containing $n-1$ parameters and σ_*^2 is one of the disturbance variances in the model. The variances of the disturbance can then be expressed as $\mathbf{var}[\epsilon(t)] = \sigma_*^2 h$ and $\mathbf{var}[\delta(t)] = \sigma_*^2 Q$ where h is positive and Q is positive semi-definite. As a rule, h or one of the diagonal elements in Q will be set to equal to 1. The reparameterization of the model enables σ_*^2 to be concentrated out of the general likelihood function and the remaining disturbance variances are called *relative variances*. The prediction error decomposition yields

$$\log L = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma_*^2 - \frac{1}{2}\sum_{t=1}^{T}\log f(t) - \frac{1}{2\sigma_*^2}\sum_{t=1}^{T} v^2(t)/f(t) \qquad (7.22)$$

and since $v(t)$ and $f(t)$ are independent of σ_*^2, differentiating the above equation with respect to σ_*^2 gives

$$\hat{\sigma}_*^2(\Phi_*) = \frac{1}{T}\sum_{t=1}^{T}\frac{v^2(t)}{f(t)} \qquad (7.23)$$

The notation $\hat{\sigma}_*^2(\Phi_*)$ indicates that it is the ML estimator of σ_*^2 conditional on a given value of Φ_*. Concentrating out a parameter reduces the dimension of the search involved in the numerical optimization procedure. The difficulty arises because the param space of the variances includes zero. The likelihood function has a singularity if some variance goes t Also, care needs to be taken in choosing one variance as the scalar σ_*^2 because the algebraic manipula is not viable if σ_*^2 is zero.

The solution usually proposed is to bound the relative variances away from zero. As a special case, Gersch and Kitagawa (1983) point out that the ratios $q_{11} = \sigma_\eta^2/\sigma_\epsilon^2$, $q_{22} = \sigma_\zeta^2/\sigma_\epsilon^2$, $q_{33} = \sigma_\omega^2/\sigma_\epsilon^2$ and $q_{44} = \sigma_\xi^2/\sigma_\epsilon^2$ are signal-to-noise measures, trade-off parameters or hyperparameters. They set $\sigma_*^2 = \sigma_\epsilon^2$ and compute the likelihoods, with each $q_{ii} = 2^k$, $k = 0, 1, 2, \cdots$, by Kalman filtering. Their procedure is an approximation to maximum likelihood estimation in order to avoid the numerical optimization procedure for unknown σ_η^2, σ_ζ^2, σ_ω^2, σ_ξ^2, under the assumption that σ_η^2, σ_ζ^2, σ_ω^2, σ_ξ^2 are always greater than σ_ϵ^2.

In general, however, if the chosen σ_*^2 is close to zero, numerical problems may arise as the relative variances will tend to become very large. However,

the general likelihood function does not have this problem. Without any prior knowledge about the disturbance variances of different components in a state space model, one can use the unconcentrated likelihood function to estimate the model parameters (including the disturbance variances) instead of considering the choice of σ_*^2 in the concentrated form of the likelihood function. Therefore the general form of likelihood function is employed in our initial model identification.

Initial Parameters Estimation

As Engle (1978) points out the Kalman filtered estimates perform very poorly at the beginning of the sample period. This result is traceable directly to the estimation of the initial state. The model parameters are very sensitive to these estimates. Furthermore, the initial state has a substantial effect on the estimates for many periods. Considerable effort should therefore be directed to developing better methods for beginning the Kalman filtering procedure.

As discussed in chapter 3, the initial state covariance plays an important role in ensuring that the state covariance converges quickly to its steady state. The accuracy of the starting value of the initial state covariance is extremely important when the available sample size is small and the model is not stabilizable under these conditions the state covariance matrix may not converge approximately to the steady state at the end of the sample. Shephard and Harvey (1990) have paid special attention to the local linear trend component. Their Monte Carlo study showed that the diffuse prior initial state variance leads to an estimated σ_ζ^2 with a higher probability of being non-zero than the fixed unknown initial state with zero variance does, if in fact $\sigma_\zeta^2 > 0$. On the other hand, if $\sigma_\zeta^2 = 0$, the fixed initial approach estimates σ_ζ^2 to be zero with higher probability than does the diffuse initial approach.

Theorem 3.7 in chapter 3 proves that an over estimated initial variance-covariance for the state vector leads to quicker convergence than does an under-estimated one. By the error sensitive analysis in chapter 3, Theorem 3.9 and Corollary 2, we also

show that a conservative initial state covariance matrix plus a conservative state disturbance covariance matrix provides a high convergence speed for the state covariance matrix. Furthermore, the RDE with an over estimated initial state covariance matrix will converges to the steady state covariance matrix (the solution of the ARE) without the stabilizable condition on the specified model. However, this condition is a necessary condition for the RDE with an under-estimated initial state covariance matrix to converge. Although a diffuse state variance and relatively large variance elements in the state disturbance covariance matrix can always be used initially to allow the parameters to be estimated by the maximum likelihood approach through the Kalman filter, a more effective estimation can be obtained by using prior information about the state and state disturbance covariance matrices.

First of all, we suppose the basic structural model (7.20) is an adequate model when the trend is a local linear trend, i.e. $\rho = 1$, and the coefficient of the exogenous input, W, is set so that $\alpha = 1$ and $\sigma_\xi^2 = 0$, i.e. the weather sensitive load is deterministic.

Under the above conditions, $y(t)$ has the following form based on model (7.20)

$$y(t) = y_{trend}(t) + y_{wp}(t) + y_{ws}(t) + \epsilon(t) \tag{7.24}$$

where $y_{trend}(t)$, $y_{wp}(t)$, $y_{ws}(t)$ represent trend, weekly periodic, and weather sensitive components.

From the local linear trend model (7.13), the slope can be expressed as $\beta(t-1) = \zeta(t-1)/\Delta$ and substituted into $m(t) = m(t-1) + \beta(t-1) + \epsilon(t)$. The local linear trend, therefore, can be expressed as

$$y_{trend}(t) = \eta(t)/\Delta + \zeta(t)/\Delta^2 \tag{7.25}$$

Similarly, from the weekly periodic model (7.15), the weekly periodic component $y_{wp}(t) = r(t)$ can be expressed as

$$S(B)r(t) = \omega(t) \tag{7.26}$$

from the weather sensitive component model (7.19), the coefficient $\alpha(t)$ can be expressed as

$$\alpha(t) = \xi(t)/\Delta \tag{7.27}$$

Under the specified conditions, therefore, $y(t)$ can be expressed as driven by four different disturbance components.

$$y(t) = \frac{\eta(t)}{\Delta} + \frac{\zeta(t)}{\Delta^2} + \frac{\omega(t)}{S(B)} + \frac{\xi(t)}{\Delta}W + \epsilon(t) \tag{7.28}$$

Furthermore, $y(t)$ can be made stationary by the operator $\Delta\Delta_s$ because

$$\Delta\Delta_s(y(t)) = \Delta_s\eta(t) + S(B)\zeta(t-1) + \Delta^2\omega(t) + \Delta_s\xi(t)W + \Delta\Delta_s\epsilon(t) \tag{7.29}$$

If we assume $\sigma_\xi^2 = 0$, it can be verified that the auto-covariances of $\Delta\Delta_s y(t)$ satisfy the following relations

$$\begin{cases} c(0) = & 2\sigma_\eta^2 & +s\sigma_\zeta^2 & +6\sigma_\omega^2 & +4\sigma_\epsilon^2 & \\ c(1) = & & (s-1)\sigma_\zeta^2 & -4\sigma_\omega^2 & +2\sigma_\epsilon^2 & \\ c(2) = & & (s-2)\sigma_\zeta^2 & +\sigma_\omega^2 & & \\ c(\tau) = & & (s-\tau)\sigma_\zeta^2 & & & \tau = 3, \cdots, s-2 \\ c(s-1) = & & \sigma_\zeta^2 & & +\sigma_\epsilon^2 & \\ c(s) = & -\sigma_\eta^2 & & & -\sigma_\epsilon^2 & \\ c(s+1) = & & & & \sigma_\epsilon^2 & \\ c(\tau) = & 0 & & & & \tau > s+1 \end{cases} \tag{7.30}$$

If the weather sensitive load is not deterministic and W is a constant, i.e. the variance of the disturbance $\xi(t)$, σ_ξ^2 is nonzero, σ_η^2 in equation (7.30) will be replaced by $\sigma_\eta^2 + W^2\sigma_\xi^2$. Therefore, under the assumption of a deterministic weather insensitive load, from equation (7.30), the initial estimation for $\hat{\sigma}_\eta^2, \hat{\sigma}_\zeta^2, \hat{\sigma}_\epsilon^2$ can be obtained from the auto-covariances of $\Delta\Delta_s y(t)$ and no matter how W changes, the true value of σ_η^2 could be expected always to be smaller (or at least not larger) than its initial estimated value. i.e.

$$\hat{\sigma}_\eta^2 = \sigma_\eta^2 + W^2\sigma_\xi^2 \geq \sigma_\eta^2 \tag{7.31}$$

It is also noticed that the mis-specification of the weather insensitive load affects the level of the local linear trend.

On the other hand, from equation (7.31), for any W, we also have $\hat{\sigma}_\eta^2 \geq W^2\sigma_\xi^2$, and then $\sigma_\xi^2 \leq \frac{\hat{\sigma}_\eta^2}{[\max(W)]^2} \leq \frac{\hat{\sigma}_\eta^2}{[\mathrm{E}(W)]^2}$. Because, from equation (7.5) for the weather conditions of the samples, the maximum value of W does not exceed 0.685 and the sample mean of W is $\bar{W} = 0.233$, $\hat{\sigma}_\xi^2 = \frac{\hat{\sigma}_\eta^2}{\bar{W}^2}$ is a reasonable initial conservative estimate of σ_ξ^2.

The initial variances of the four different disturbances are calculated and set out in Table 7.6. With these initial estimates, the model estimation procedure, which includes a fixed point smoothing for the initial state as proposed in chapter 3, can be applied to the proposed SSSM.

$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\zeta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\xi^2$	$\hat{\sigma}_\epsilon^2$
1.909E-4	1.4324E-4	10.1161E-4	35.1637E-4	22.3720E-4

Table 7.6: Estimated Initial Disturbance Variances

7.4.5 Model Testing

If the basic state space model is adequate, what we are interested is follows:

For The Trend Component

- Is the trend a linear trend ? i.e. is $\rho = 1$
  1. Is the slope of the linear trend constant ? i.e. is $\sigma_\zeta^2 = 0$
  2. Is the linear trend deterministic ? i.e. is $\sigma_\eta^2 = \sigma_\zeta^2 = 0$
- Is the trend a linear damped trend ? i.e. is $\rho < 1$
  1. Is the slope of the damped trend constant ? i.e. is $\sigma_\zeta^2 = 0$
  2. Is the damped trend deterministic ? i.e. is $\sigma_\eta^2 = \sigma_\zeta^2 = 0$
- Is the trend a simple random walk plus noise ? i.e. is $\beta(t) = 0$

1. Is the trend itself a random walk ? i.e. is $\sigma_\epsilon^2 = 0$
2. Is the trend a constant ? i.e. is $\sigma_\epsilon^2 = \sigma_\eta^2 = 0$

For The Weekly Periodical Component

- Is the weekly periodic component deterministic ? i.e. is $\sigma_\omega^2 = 0$

For The Weather Sensitive Component

- Is the weather sensitive component deterministic ? i.e. is $\sigma_\xi^2 = 0$

Basically, there are three tests which are suitable to test a structural state space model estimated by the maximum likelihood method. These are the *Likelihood Ratio*(LR) test, *Lagrange Multiplier* (LM) test and *Most Powerful Invariant* (MPI) test based on the theory of invariance by Lehmann (1959).

Our basic structural model contains five variance parameters $(\sigma_\eta^2, \sigma_\zeta^2, \sigma_\omega^2, \sigma_\xi^2, \sigma_\epsilon^2)$, and all of them are constrained to be non-negative since they are variances. Thus a value of zero is on the boundary of the parameter space. The LR based tests can be applied to test some of the parameters provided that the maximum likelihood estimator of the parameters in question are asymptotically normally distributed when the restriction is not applied. The LR test has following form

$$\text{LR} \approx (T - d) \log(\hat{\sigma}_{p0}^2 / \hat{\sigma}_{p1}^2) \tag{7.32}$$

where T is the number of data points; d is the dimension of the model; $\hat{\sigma}_{p0}$, $\hat{\sigma}_{p1}$ are variances of one step prediction from the null and alternative hypothesis respectively.

The LR test statistic has a χ^2 distribution with m degree of freedom where m is the number of parameters under the alternative hypothesis if the null hypothesis is true. It is noticed that when some of variance parameters $(\sigma_\eta^2,\ \sigma_\zeta^2,\ \sigma_\omega^2,\ \sigma_\xi^2, \sigma_\epsilon^2)$ are zero (i.e. they lie on the boundary of the parameter space) under the null hypothesis, the LR test statistic does not have the usual asymptotic χ^2 distribution because of the one sided restriction.

For instance, when one of variance parameters in the state space model is zero under the null hypothesis and to be estimated under the alternative hypothesis, the LR test statistic will satisfy a χ^2 mixture distribution such that

$$LR = \begin{cases} \chi_1^2 & p = 1/2 \\ 0 & p = 1/2 \end{cases} \tag{7.33}$$

If there are more than two variance parameters lying on the parameter space boundary under the null hypothesis, the LR statistic is dependent on the information matrix of the parameter vector and could be distributed with a very complicated χ^2 mixture. Further details can be found in Gourieroux et al. (1982). However, under any significance level α, it is sufficient to reject the null hypothesis when the LR statistic is greater than $\chi_m^2(\alpha)$ where m is the number of parameters to be estimated under the alternative hypothesis because the value of the true mixture χ^2 distribution, with significance level α, is always less than $\chi_m^2(\alpha)$.

Actually for the basic structural model, the LR test can be used provided that σ_ζ^2, σ_ω^2, σ_ξ^2 are all strictly positive, i.e. the structural model is stabilizable and detectable since the maximum likelihood estimated parameters are asymptotically normally distributed.

Once one of σ_ζ^2, σ_ω^2, σ_ξ^2 is constrained to be zero, the other maximum likelihood estimated parameters may not be asymptotically normally distributed. The LR based test is then not valid anymore.

One solution is to use the LM test because the LM test statistic still has the usual asymptotic χ^2 distribution even though some variance parameters lie on the boundary of the parameter space. However, it takes no account of the one-sided nature of the alternative, and therefore has lower power compared to other tests. Rogers (1986) develops a modified LM test to handle the one-sided alternatives.

Another partial solution to this situation is an MPI test which sets up a test prior to estimating the model in which some parameters values under both the null and the alternative are pre-specified and related. An MPI test can be constructed as follows

$$b = \frac{S(\Phi_1^*)}{S(\Phi_0^*)} \tag{7.34}$$

where Φ_0^*, Φ_1^* are estimated parameter sets under the null and alternative hypothesis respectively; $S(\Phi_0^*)$, $S(\Phi_1^*)$ are the corresponding generalized residual sum of squares from the models under the null and alternative hypothesis.

It is easy to verify that the generalized residual sum of squares is then given by the sum of squares of the standardized one step ahead prediction errors, i.e.

$$S(\Phi^*) = \sum_{t=s+2}^{T} v^2(t)/f(t) \tag{7.35}$$

where $v(t)$ is the one-step ahead prediction error at time t and $f(t) = \mathbf{var}[v(t)]$.

It is expected that the statistic b obeys the relation $b < c$, where c is a constant, and this relation will be the basis for testing the null hypothesis. As to the efficiency and power of the MPI test, Franzini and Harvey (1983) suggest that a pre-specified parameter $q > 0$ is chosen so that a prespecified relationship will occur between the relative variances. The critical value c is dependent on T and s for the MPI test. The critical values calculated using the method of Imhof (1961) for quarterly data are given in Franzini and Harvey (1983).

We refer to the statistic b in equation (7.34) as the FH statistic when the pre-specified parameter is determined in the way Franzini and Harvey suggested and the corresponding test is referred to as the FH-MPI test.

Nyblom (1986) studied the test for a deterministic trend and suggested that the pre-specified noise ratio between the slope and the level, i.e. $q = \sigma_\zeta^2/\sigma_\eta^2$, should be chosen as $375.1/(T-2)^2$ to gain maximum power and Pitman efficiency. The statistic then has the following form

$$b_* = (T-2)[1 - S(\Phi_1^*)/S(\Phi_0^*)]/375.1 \tag{7.36}$$

where the relation $b_* < c_*$ gives the critical region for the test of the null hypothesis where c_* is a constant. The table of critical values for the test can be calculated by Imhof's method (see Nyblom (1986) for details). Similarly, we refer to the statistic b_* in equation (7.36) as the N statistic when the pre-specified parameter is determined by the approach of Nyblom and the corresponding test is known as the N-MPI test.

Since there are five variance parameters to be tested, in other words, there are 32 possible hypotheses to test. It is not feasible practically to conduct all the possible hypothesis tests. Therefore, we designed a scheme to test the most interesting hypotheses sequentially as follows: In the first stage, we assume that for $\forall t$, $\sigma_\xi^2 = 0$ and $\alpha(t) = 1$. Therefore, we can use $y(t) - W$ as the observed weather insensitive load, and apply the SSSM (7.20) without the last dimension (weather sensitive dimension) to test the nature of the trend and weekly periodic components. Taking the accepted model from one test, we test further alternatives in the chosen model and so on. In the second stage, then we test the alternatives of the optimal model from the first stage without the above assumptions. The LR and MPI tests are employed in the testing procedure.

Deterministic Trend and Weekly Periodicity Test

The null hypothesis is that the trend and weekly periodic components are deterministic, i.e. $\mathbf{H_0}$: $\sigma_\eta^2 = \sigma_\zeta^2 = \sigma_\omega^2 = 0$ and $S(\Phi_0^*) = S(0,0,0,0,\sigma_{*0}^2)$ where $\sigma_{*0}^2 = \sigma_\epsilon^2$. This model collapses to the regression model.

The alternative hypothesis is more general; stochastic linear trend and weekly periodic components, i.e. $\mathbf{H_1}$: σ_η^2, σ_ζ^2, σ_ω^2, $\sigma_\epsilon^2 > 0$. Because the model under the null hypothesis is not stabilizable and detectable, the hypothesis can be tested by constructing a "special" alternative model with constraints $\sigma_\eta^2 = \sigma_\omega^2 = q\sigma_\epsilon^2$, $\sigma_\zeta^2 = 0$, $\sigma_{*1}^2 = \sigma_\epsilon^2$, and $S(\Phi_1^*) = S(q\sigma_{*1}^2, 0, q\sigma_{*1}^2, 0, \sigma_{*1}^2)$. The rationale behind the special alternative model is that if there is variation in the trend, of any kind, it will tend to show up in a test against σ_η^2. An FH-MPI test is employed with $q = 0.0234$ to test the null hypothesis. The test results listed in Table 7.7 reject the null hypothesis.

Deterministic Trend Test

The null hypothesis is that the trend component is deterministic and the weekly periodic component may not be deterministic. i.e. $\mathbf{H_0}$: $\sigma_\eta^2 = \sigma_\zeta^2 = 0$ and $S(\Phi_0^*) = S(0,0,\sigma_\omega^2,0,\sigma_\epsilon^2)$.

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0}$: $\sigma_\eta^2 = \sigma_\omega^2 = 0$	$\sigma_{*0}^2 = \sigma_\epsilon^2$	$\sigma_{*0}^2 = 1.442824$
	$\sigma_\eta^2 = \sigma_\zeta^2 = \sigma_\omega^2 = 0$	$S(\Phi_{*0}) = 340$
		$\sigma_p^2 = 4.2436\text{E-}3$
$\mathbf{H_1}$: $\sigma_\eta^2 > 0,\ \sigma_\omega^2 > 0$	$\sigma_{*1}^2 = \sigma_\epsilon^2$	$\sigma_{*1}^2 = 1.220561$
	$\sigma_\eta^2 = \sigma_\omega^2 = q\sigma_{*1}^2$	$S(\Phi_{*1}) = 282.234$
	$\sigma_\zeta^2 = 0$	$\sigma_p^2 = 3.91\text{E-}3$
Test Statistic	$b = 0.8301$ (FH-MPI)	
5% Critical Value	$c = 0.8671$	
	$b < c$, $\mathbf{H_0}$ is rejected	

Table 7.7: Test Trend and Weekly Periodicity — Deterministic

The alternative hypothesis is the more general stochastic linear trend, i.e. $\mathbf{H_1}$: $\sigma_\eta^2,\ \sigma_\zeta^2 > 0$. A "special" alternative model is set up with constraints $\sigma_\eta^2 = q\sigma_*^2$, $\sigma_\zeta^2 = 0$, $\sigma_*^2 = \sigma_\epsilon^2$ and $S(\Phi_1^*) = S(\sigma_\eta^2, 0, \sigma_\omega^2, 0, \sigma_\epsilon^2)$. Again, because the model under the null hypothesis is not stabilizable and detectable, the N-MPI test is conducted with $q = 375.1/(T-2)^2$ and result is given in Table 7.8, and the null hypothesis is rejected by the N-MPI test.

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0}$: $\sigma_\eta^2 = \sigma_\zeta^2 = 0$	$\sigma_{*0}^2 = \sigma_\epsilon^2$	$\sigma_{*0}^2 = 3.7852\text{E-}3$
	$\sigma_\eta^2 = 0$	$\sigma_\omega^2 = 0.0$
	$\sigma_\zeta^2 = 0$	$S(\Phi_{*0}) = 339.3488$
		$\sigma_p^2 = 4.103\text{E-}3$
$\mathbf{H_1}$: $\sigma_\eta^2 > 0,\ \sigma_\zeta^2 > 0$	$\sigma_{*1}^2 = \sigma_\epsilon^2$	$\sigma_{*1}^2 = 2.4983\text{E-}3$
	$\sigma_\eta^2 = q\sigma_\epsilon^2$	$\sigma_\omega^2 = 0.0$
	$\sigma_\zeta^2 = 0$	$S(\Phi_{*1}) = 329.1507$
		$\sigma_p^2 = 2.7529\text{E-}3$
Test Statistic	$b_* = 0.027079$ (N-MPI)	
5% Critical Value	$c_* = 0.035$	
	$b_* < c_*$, $\mathbf{H_0}$ is rejected	

Table 7.8: Test Trend — Deterministic

Deterministic Weekly Periodicity Test

In a similar manner to the deterministic trend test adopted above, the null hypothesis is that the weekly periodic component is deterministic and the trend component may not be deterministic. i.e. $\mathbf{H_0}$: $\sigma_\omega^2 = 0$ and $S(\Phi_0^*) = S(\sigma_\eta^2, \sigma_\zeta^2, 0, 0, \sigma_\epsilon^2)$.

The alternative hypothesis is a more general random walk weekly periodic component, i.e. $\mathbf{H_1}$: $\sigma_\omega^2 > 0$. Since $\sigma_\eta^2, \sigma_\zeta^2$ are not known, we construct "special" testing models by constraining $\sigma_\eta^2 = \sigma_\zeta^2 = q\sigma_\epsilon^2$ with $q = 375.1/(T-2)^2$ for the null and alternative hypothesis respectively. The N-MPI test is conducted and the results show that the null hypothesis is rejected (see Table 7.9). In other words, the weekly periodic component is not deterministic.

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0}$: $\sigma_\omega^2 = 0$	$\sigma_{*0}^2 = \sigma_\epsilon^2$	$\sigma_{*0}^2 = 1.425\text{E-}3$
	$\sigma_\omega^2 = 0$	$S(\Phi_{*0}) = 319.75$
	$\sigma_\eta^2 = \sigma_\zeta^2 = q\sigma_\epsilon$	$\sigma_p^2 = 2.4017\text{E-}3$
$\mathbf{H_1}$: $\sigma_\omega^2 > 0$	$\sigma_{*1}^2 = \sigma_\epsilon^2$	$\sigma_{*1}^2 = 1.2578\text{E-}3$
	$\sigma_\eta^2 = \sigma_\zeta^2 = \sigma_\omega^2 = q\sigma_{*1}^2$	$S(\Phi_{*1}) = 319.0204$
		$\sigma_p^2 = 2.3830\text{E-}3$
Test Statistic	$b_* = 0.002057$ (N-MPI)	
5% Critical Value	$c_* = 0.035$	
	$b_* < c_*$, $\mathbf{H_0}$ is rejected	

Table 7.9: Test Weekly Periodicity — Deterministic

Partial Deterministic Trend and Weekly Periodicity Test

The null hypothesis is that the trend and weekly periodic components are partially deterministic, i.e. $\mathbf{H_0}$: $\sigma_\zeta^2 = \sigma_\omega^2 = 0$, and $\sigma_\epsilon^2 = 0$. This model implies that the trend is a random walk plus drift and the weekly periodic component is fixed.

The alternative hypothesis is a more general model, i.e. $\mathbf{H_1}$: σ_ζ^2, σ_ω^2, $\sigma_\eta^2 > 0$ and $\sigma_\epsilon^2 = 0$. The FH-MPI test model will be $\sigma_\zeta^2 = \sigma_\omega^2 = q\sigma_\eta^2$, $S(\Phi_1^*) = S(\sigma_{*1}^2, q\sigma_{*1}^2, q\sigma_{*1}^2, 0, 0)$,

where $\sigma^2_{*1} = \sigma^2_\eta$. The FH-MPI test is conducted with $q = 0.00784$ and results listed in Table 7.10 show that the null hypothesis should be rejected.

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0} : \sigma^2_\eta > 0$	$\sigma^2_{*0} = \sigma^2_\eta$	$\sigma^2_{*0} = 1.8521\text{E-}3$
$\sigma^2_\zeta = \sigma^2_\omega = \sigma^2_\epsilon = 0$	$\sigma^2_\zeta = \sigma^2_\omega = \sigma^2_\epsilon = 0$	$S(\Phi_{*0}) = 327.7981$
		$\sigma^2_p = 1.9214\text{E-}3$
$\mathbf{H_1} : \sigma^2_\epsilon = 0,\ \sigma^2_\eta > 0$	$\sigma^2_{*1} = \sigma^2_\eta,\ \sigma^2_\epsilon = 0$	$\sigma^2_{*1} = 1.4567\text{E-}3$
$\sigma^2_\zeta > 0,\ \sigma^2_\omega > 0$	$\sigma^2_\eta = \sigma^2_\omega = q\sigma^2_{*1}$	$S(\Phi_{*1}) = 294.4885$
		$\sigma^2_p = 1.7022\text{E-}3$
Test statistics	$b = 0.8984$ (FH-MPI)	
5% Critical Value	$c = 0.9034$	
	$b < c$, $\mathbf{H_0}$ is rejected	

Table 7.10: Test Trend and Weekly Periodicity — Partial Deterministic

From the above tests, we are convinced that the trend and weekly periodic components are stochastic. However, the stochastic trend component can take on different forms, such as, (1) a random walk plus drift($\sigma^2_\zeta = 0$, fixed slope), (2) a simple random walk plus noise, or (3) a random walk.

Random Walk Plus Drift Trend Test

The null hypothesis is that the slope of the trend is deterministic i.e. $\mathbf{H_0} : \sigma^2_\zeta = 0$ and the alternative is more general, i.e. $\mathbf{H_1} : \sigma^2_\zeta \neq 0$. Under the null hypothesis, the properties of stabilizability and detectability are lost, so, the MPI test has to be applied to the null hypothesis with $\sigma^2_\zeta = 0$, $\sigma^2_\eta = \sigma^2_\omega = q\sigma^2_\epsilon$, and $S(\Phi^*_0) = S(q\sigma^2_{*0}, 0, q\sigma^2_{*0}, 0, \sigma^2_{*0})$ where $\sigma^2_{*0} = \sigma^2_\epsilon$ against the alternative with $\sigma^2_\eta = \sigma^2_\zeta = \sigma^2_\omega = q\sigma^2_\epsilon$, and $S(\Phi^*_1) = S(q\sigma^2_{*1}, q\sigma^2_{*1}, q\sigma^2_{*1}, 0, \sigma^2_{*1})$ where $\sigma^2_{*0} = \sigma^2_\epsilon$. The FH-MPI test is conducted with $q = 0.007841$ and the result listed in Table 7.11 rejects the alternative hypothesis.

It is noted that the variance of the estimated one step ahead predictions under the null hypothesis σ^2_{p0} is less than σ^2_{p1} obtained under the alternative hypothesis and the FH-MPI test statistic is on the edge of 5% level. This fact indicates that the model

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0}: \sigma_\zeta^2 = 0$	$\sigma_{*0}^2 = \sigma_\epsilon^2$	$\sigma_{*0}^2 = 1.3843\text{E-}3$
	$\sigma_\eta^2 = \sigma_\omega^2 = q\sigma_{*0}^2$	$S(\Phi_{*0}) = 307.8817$
	$\sigma_\zeta^2 = 0$	$\sigma_{p0}^2 = 2.2822\text{E-}3$
$\mathbf{H_1}: \sigma_\zeta^2 > 0$	$\sigma_{*1}^2 = \sigma_\epsilon^2$	$\sigma_{*1}^2 = 1.2217\text{E-}3$
	$\sigma_\eta^2 = \sigma_\zeta^2 = \sigma_\omega^2 = q\sigma_{*1}^2$	$S(\Phi_{*1}) = 281.8578$
		$\sigma_{p1}^2 = 2.3903\text{E-}3$
Test Statistic	$b = 0.9155$ (FH-MPI)	
5% Critical Value	c = 0.9034	
	$b > c$, $\mathbf{H_1}$ is rejected	

Table 7.11: Test Trend — Random Walk Plus Drift

under the null hypothesis may be superior to the alternative one. However, because the fixed relation between the variances of different components may not be realistic, the test statistic implies that the variance of the slope, σ_ζ^2, may be relatively smaller than the other component variances.

Random Walk Plus Noise Trend Test

When $\rho = 0$, the local linear trend model becomes a random walk plus noise model, i.e.

$$\begin{cases} m(t) = m(t-1) + \eta_*(t) \\ y_{trend}(t) = m(t) \end{cases} \tag{7.37}$$

where $\eta_*(t) = \eta(t) + \zeta(t)$, and $\sigma_{\eta_*}^2 = \sigma_\eta^2 + \sigma_\zeta^2$. It can be seen that the random walk plus noise model is stabilizable and detectable if σ_η^2 or σ_ζ^2 is strictly positive.

The null hypothesis is that the trend component is a local linear trend model i.e. $\rho = 1$, against an alternative of a random walk plus noise model, i.e. $\rho = 0$. Since both trend models still make the basic structural model stabilizable and detectable when σ_η^2, σ_ζ^2, $\sigma_\omega^2 > 0$, the LR test can be employed to test the null hypothesis. The test result listed in Table 7.12 shows that the local linear trend hypothesis is rejected by the LR test. It is noted that the trend component of the estimated alternative

model is actually a pure random model since the value of the disturbance variance, σ_ζ^2, is estimated to be zero. This indicates that the trend component is identified as a pure random walk under the condition that the variance of the weather effect coefficient is zero.

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0}: \rho = 1$	$\sigma_{*0}^2 = \sigma_\epsilon^2$	$\sigma_{*0}^2 = 1.0579\text{E-}3$
	$\sigma_\eta^2 > 0$	$\sigma_\eta^2 = 3.6442\text{E-}5$
	$\sigma_\zeta^2 > 0$	$\sigma_\zeta^2 = 2.0400\text{E-}6$
	$\sigma_\omega^2 > 0$	$\sigma_\omega^2 = 2.5820\text{E-}5$
		$S(\Phi_{*0}) = 291.3755$
		$\sigma_{p0}^2 = 2.4543\text{E-}3$
$\mathbf{H_1}: \rho = 0$	$\sigma_{*1}^2 = \sigma_\epsilon^2$	$\sigma_{*1}^2 = 8.0714\text{E-}4$
	$\sigma_\eta^2 > 0$	$\sigma_\eta^2 = 1.0797\text{E-}4$
	$\sigma_\zeta^2 > 0$	$\sigma_\zeta^2 = 0.0$
	$\sigma_\omega^2 > 0$	$\sigma_\omega^2 = 1.2247\text{E-}5$
		$S(\Phi_{*1}) = 315.342$
		$\sigma_{p1}^2 = 2.0372\text{E-}3$
Test Statistic	$\chi_* = 61.65386$ (LR)	
5% Critical Value	$\chi_{0.05}^2(1) = 3.8410$	
	$\chi_* > \chi_{0.05}(1)$, $\mathbf{H_0}$ is rejected	

Table 7.12: Test Trend — Random Walk Plus Noise

Deterministic Weather Effect Coefficient Test

The above tests under the condition of a zero disturbance variance for the weather effect coefficient, suggest strongly that the trend and weekly periodic components are pure random walks, i.e. $\sigma_\eta^2 > 0$, $\sigma_\zeta^2 = 0$, $\sigma_\omega^2 > 0$, $\rho = 0$. However, the validity of the pre-specified variance is still in question because there is evidence that the weather effect was not deterministic when we created the weather sensitive component W at the beginning of this chapter. Now, we use original observed load data $y(t)$ and the state space model (7.20) to verify that the weather effect coefficient is not deterministic by constructing the following hypothesis test based on the derived models for the trend and weekly periodic components.

The null hypothesis is that the structural model has a deterministic weather effect coefficient, i.e. $\sigma_\xi^2 = 0$; the general alternative is that the structural model has stochastic weather effect coefficient, i.e. $\sigma_\xi^2 > 0$. Because the model is not stabilizable and detectable under the null hypothesis, an FH-MPI test is conducted with a pre-specified parameter $q = 0.007841$. The results listed in Table 7.13, FH-MPI test rejects the null hypothesis. This confirms that the weather effect is not deterministic.

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0}: \sigma_\xi^2 = 0$	$\sigma_{*0}^2 = \sigma_\epsilon^2$	$\sigma_{*0}^2 = 2.4425\text{E-}3$
	$\sigma_\eta^2 = \sigma_\omega^2 = q\sigma_{*0}^2$	
	$\sigma_\zeta^2 = \sigma_\xi^2 = 0$	$S(\Phi_{*0}) = 343.8454$
		$\sigma_{p0}^2 = 2.8938\text{E-}3$
$\mathbf{H_1}: \sigma_\xi^2 > 0$	$\sigma_{*1}^2 = \sigma_\epsilon^2$	$\sigma_{*1}^2 = 2.0940\text{E-}3$
	$\sigma_\eta^2 = \sigma_\omega^2 = \sigma_\xi^2 = q\sigma_{*0}^2$	$S(\Phi_{*1}) = 307.5784$
	$\sigma_\zeta^2 = 0.0$	$\sigma_{p1}^2 = 2.6033\text{E-}3$
Test Statistic	$b = 0.89453$ (FH-MPI)	
5% Critical Value	c = 0.9034	
	$b > c$, $\mathbf{H_0}$ is rejected	

Table 7.13: Test Weather Coefficient — Deterministic

Measurement Error Test

Now, the only parameter that has not been tested is the variance of the disturbance, σ_ϵ^2, in model measurement equation. The physical explanation of σ_ϵ^2 is the measurement error which occurs when the state (or transition) model is correct, and, is expected to be zero if we assume that there is no observation error.

A test can be conducted using an LR test based on the assumptions about the nature of the random walk trend, non-deterministic weekly periodic component and weather effect coefficient, i.e. σ_η^2, σ_ω^2 and σ_ξ^2 are strictly positive and $\sigma_\zeta^2 = \rho = 0$. Under the above conditions, the null hypothesis is $\sigma_\epsilon^2 = 0$, and the alternative is $\sigma_\epsilon^2 > 0$. In Table 7.14, we observe that the alternative hypothesis is rejected.

Discussion:

Hypothesis	Test Model	Estimation of Test Model
$\mathbf{H_0}: \sigma_\epsilon^2 = 0$	$\sigma_\zeta^2 > 0$	$\sigma_\zeta^2 = 1.2483\text{E-}3$
$\sigma_\eta^2 = 0$	$\sigma_\omega^2 > 0$	$\sigma_\omega^2 = 1.9779\text{E-}6$
$\rho = 0$	$\sigma_\xi^2 > 0$	$\sigma_\xi^2 = 3.0871\text{E-}4$
		$\sigma_{p0}^2 = 1.6713\text{E-}3$
$\mathbf{H_1}: \sigma_\epsilon^2 > 0$	$\sigma_\epsilon^2 > 0$	$\sigma_\epsilon^2 = 7.7534\text{E-}4$
$\sigma_\eta^2 = 0$	$\sigma_\zeta^2 > 0$	$\sigma_\zeta^2 = 2.1589\text{E-}4$
$\rho = 0$	$\sigma_\omega^2 > 0$	$\sigma_\omega^2 = 1.3603\text{E-}5$
	$\sigma_\xi^2 > 0$	$\sigma_\xi^2 = 1.1693\text{E-}3$
		$\sigma_{p1}^2 = 1.8682\text{E-}3$
Test Statistic	$\chi^* = -36.86467$ (LR)	
5% Critical Value	$\chi^2_{0.05}(1) = 3.841$	
	$\chi^* < \chi^2_{0.05}(1)$, $\mathbf{H_1}$ is rejected	

Table 7.14: Test of Measurement Error

Summarizing all the tests carried out above, we conclude that the most suitable model for the selected sample data in the structural model framework is the model with random trend, stochastic weekly periodic and weather effect components, i.e. $\rho = \sigma_\eta^2 = \sigma_\epsilon^2 = 0$. This model has a clear and natural interpretation, i.e. the trend, weekly periodicity and weather effect components are all stochastic. The trend component is identified as $(1-B)m(t) = \eta(t)$, which is a random walk type stochastic variable. The identified model shows that the trend prediction function is just the last estimation of the trend component. The load profile has its main contribution from the weekly periodic and weather effect patterns which change slowly over time.

The natural increasing trend and annual seasonal pattern cannot be identified from one year's daily data. It is also ineffective to use a huge amount of daily data to catch the long term trend; especially, as our present interest focuses on only a few days prediction using a small sample data set. Therefore, the identified state space model performs worse than that of the ARIMAX model because the identified trend component may not be realistic. If a smaller sample set is employed to identify a structural model, the "trend" is composed of a natural increasing trend and an annual cycle. The local linear trend model may describe the "trend" effectively.

In the following section, we show, firstly, that the identified structural state space model is an adequate model although it fits the sample data less well than ARIMAX model. Secondly, five smaller sample sets are used to show that the structural state space model's performance is much better than the ARIMAX model in both fitting sample data and post sample prediction.

7.5 Diagnostic Checking and Model Selection

Diagnostic checking for the selected ARIMAX and structural state space models is based on three tests: (1) residual series correlation test (2) Box-Ljung goodness of fit test (3) cumulative periodogram test. Many other diagnostic tests give similar information to the above three diagnostic tests. It is possible to directly perceive the likely result of diagnostic checks from figures 7.1 and 7.2 for the ARIMAX model, and figures 7.3 and 7.4 for the structural state space model.

It is obvious that the two models both pass the diagnostic checks. However, the estimated one step ahead prediction error from the state space model ($\hat{\sigma^2} = 0.00167$) is greater than that from the ARIMAX model ($\hat{\sigma^2} = 0.0014$). By a close examination of both models' performance, we found that the state space model performs less well than its rival mainly because the specified trend component appears not to cope well with the trend behaviour.

To support the above explanations, we randomly choose five sample sets of daily data of size 70 for model building and a further 28 sample points for post sample prediction to compare the performances of the ARIMAX and state space models. The first sample set starts from January 6 (Sunday, summer), 1985; the second sample set starts from March 18 (Sunday, autumn), 1985; the third sample set starts from May 27 (Sunday, autumn to winter), 1985 where there was a sudden weather condition change; the fourth sample set starts from August 5 (winter to spring), 1985; the fifth sample set starts from October 14 (Sunday, spring to summer), 1985.

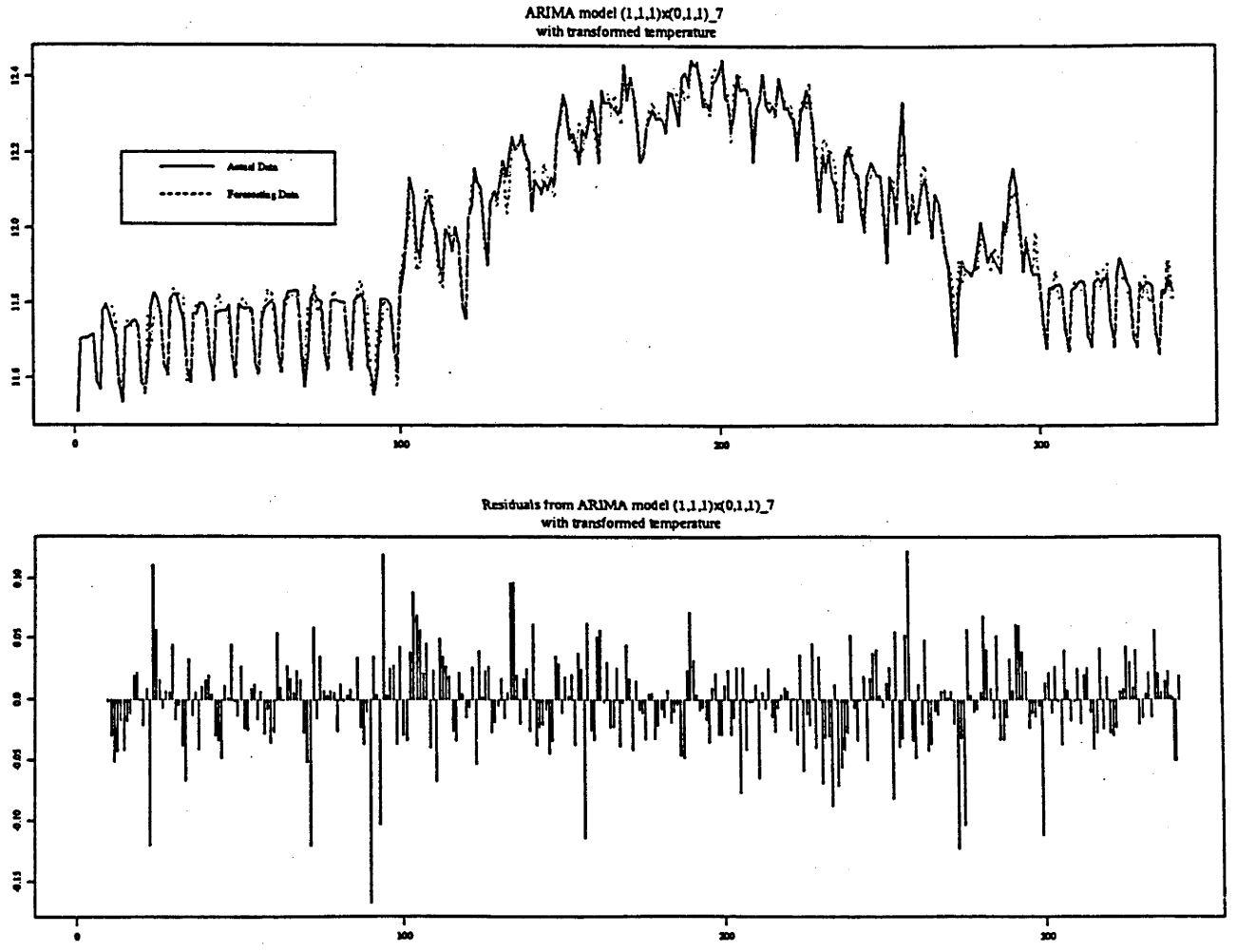


Figure 7.1: Sample Fitting of ARIMAX Model

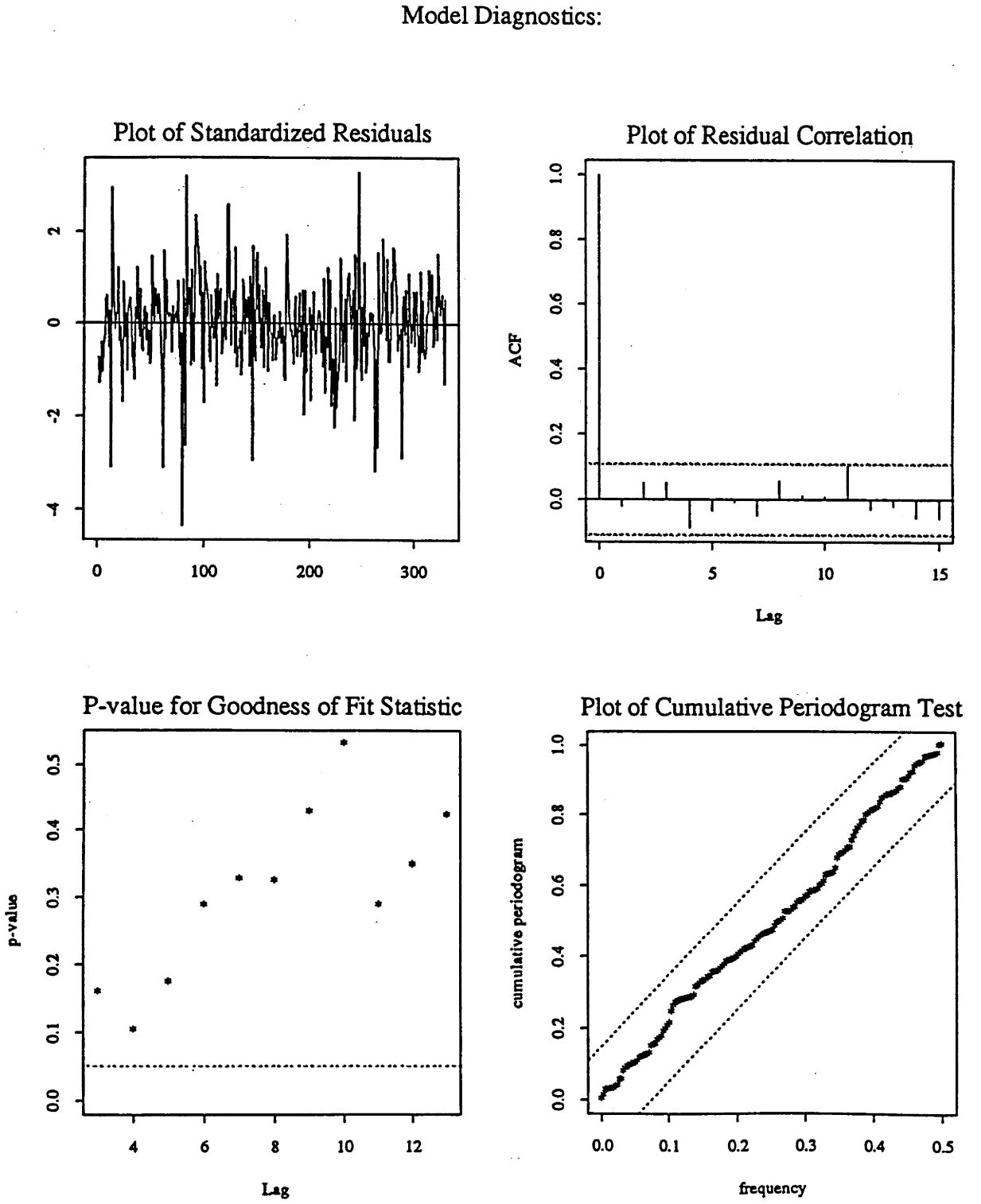


Figure 7.2: Model Diagnostic Check for ARIMAX Model

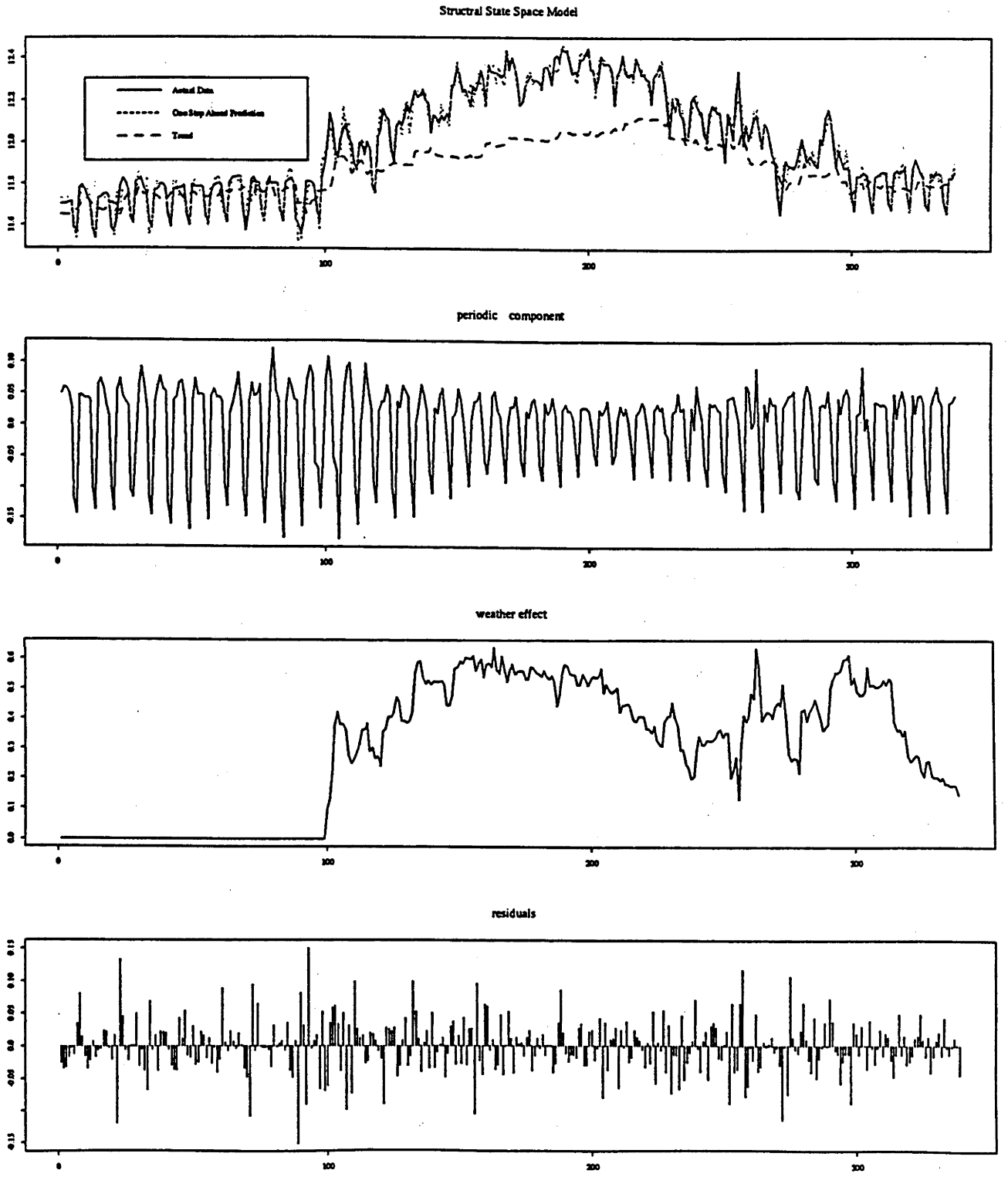


Figure 7.3: Sample Fitting of Structural State Space Model

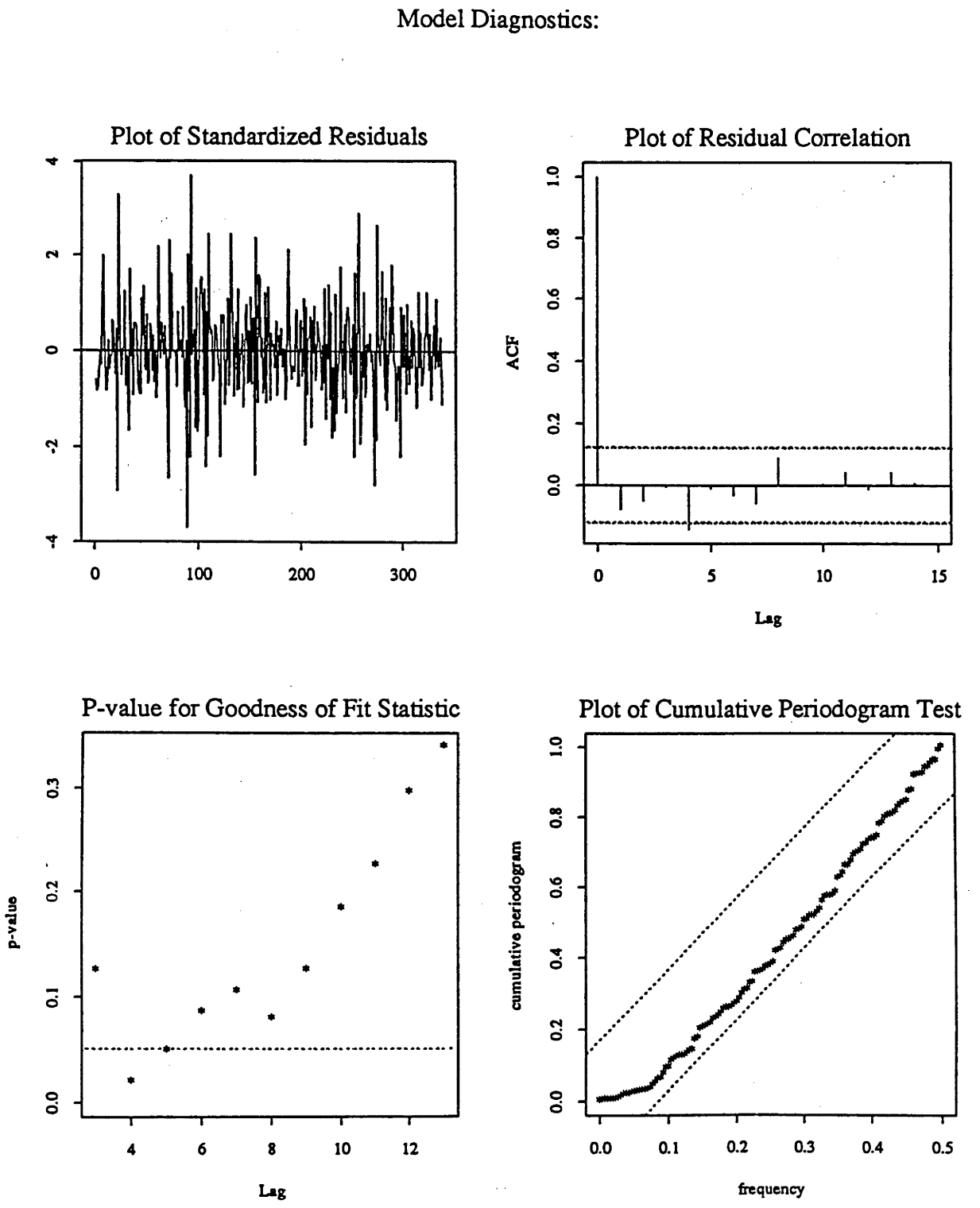


Figure 7.4: Model Diagnostic Check for Structural State Space Model

We choose (1) the variance of one step ahead prediction error, σ^2, (2) the probability of the Box-Ljung statistic at lag 15, $p[Q(15)]$ and (3) the Akaike information Criterion (AIC) to compare the fitting performance within the samples. To compare the post sample prediction performance of the two models, we choose (1) the sum of squared errors over 28 post sample predictions, PSSE(28), and (2) the Chow statistic for sample post prediction.

In building the ARIMAX model for the above sample data sets, we did not find any major model structural changes from ARIMAX$(1,1,1)\times(0,1,1)_7$. i.e. the Model ARIMAX$(1,1,1)\times(0,1,1)_7$ is an adequate model for the five different sample sets although there are differences in the parameter values. However, for the structural state space model, after the sequential model tests described in the previous section, we find there are different model forms for the structural model components for the five different sample sets. The models' performance in fitting within the samples and for post sample predictions are tabulated in Table 7.15.

Comparing the one step predictions of the ARIMAX model and the SSSM model for each sample data set, we find that the estimated one step prediction variances of the ARIMAX models are always greater than those from the SSSM models. Checking the probabilities for the Box-Ljung statistic for the two models, we find that the $p[Q(15)]$ from ARIMAX models are always greater than those from the SSSM models although the two models both pass the test at the 5% significance level. These two statistics seem to be in conflict when choosing the better model.

The reason for the conflict is that the weather effect is considered to be deterministic in the ARIMAX model, and so, the average weather effect for each sample data set is estimated. However, the significant differences between the estimated weather effect coefficients λ for the five sample data sets (see the sixth row of Table 7.15 indicate the weather effect is not deterministic. On the other hand, this effect is allowed to change in the SSSM model, as a result, the one step ahead prediction errors of the SSSM model are expected to be smaller than ARIMAX model's. In general, the

Parameter	sample 1	sample 2	sample 3	sample 4	sample 5
	ARIMAX$(1,1,1) \times (0,1,1)_7$ Model				
ϕ_1	0.5105	-0.3799	0.1515	0.3553	0.4793
θ_1	0.9888	-0.0563	0.7502	0.7538	0.3783
θ_7	0.9920	0.8390	0.9925	0.9873	0.7808
λ	0.0055	0.5385	0.6213	0.5531	0.2705
	Statistic				
σ^2	8.426E-4	2.521E-3	1.087E-3	2.026E-3	8.856E-4
$p[Q(15)]$	0.81	0.12	0.45	1.0	0.201
AIC	-489.53	-412.82	-471.70	-428.12	-486.05
PSSE(28)	0.1525	0.0932	0.2230	0.1464	0.2818
Chow	9.9754	1.4789	10.120	3.2484	11.624
	Structural State Space Model				
ρ	0.8	0.7073	0.1745	0.0	0.4772
σ_η^2	8.865E-5	6.174E-4	0.0	0.0	0.0
σ_ζ^2	0.0	0.0	4.9526E-5	7.1373E-4	1.9728E-4
σ_ω^2	0.0	6.6501E-6	0.0	1.7114E-6	3.5299E-6
σ_ξ^2	4.133E-5	2.752E-3	3.515E-4	0.0	0.0
σ_ϵ^2	4.396E-4	6.532E-4	3.983E-4	5.027E-4	1.8006E-4
	Statistic				
σ^2	7.949E-4	2.236E-3	8.884E-4	1.716E-3	8.602E-4
$p[Q(15)]$	0.42	0.101	0.12	0.83	0.08
AIC	-491.61	-417.21	-483.83	-439.74	-486.08
PSSE(28)	0.1147	0.0886	0.0127	0.0794	0.3623
Chow	5.1550	1.3433	0.7136	1.6534	15.044

Table 7.15: The Comparison of ARIMAX model and SSSM model for 5 Sample Sets

SSSM can be considered as a general form of the ARIMAX model with a stochastic coefficient for the exogenous variable and some restrictions (see equation (7.29)). These restrictions constrain the flexibility of the SSSM model to fit the sample data to some degree and cause the one step ahead prediction errors lack of normality although the overall variance of the one step ahead prediction error is smaller than those of the ARIMAX model.

The Akaike information criterion (AIC) can be used to trade off the conflict in model selection. Comparing the AIC values for the two different models in each column of Table 7.15, we can see that the SSSM's always have a lower value than the value from ARIMAX model for each selected sample set. This suggests that the state space model is a better model in the AIC sense.

The sum of squared errors for the post sample multi-step ahead, PSSE(28), and corresponding Chow statistics also show that the SSSMs are superior to its rival ARIMAX because the PSSE(28) produced by the SSSMs are less than that from ARIMAX for the first four sample sets. The reason for the different result in the fifth sample set is that the post sample period includes the Christmas and New Year holidays. Therefore, the post sample statistic from the fifth sample cannot easily be taken into account in judging the post sample prediction performance of the two models.

Discussion:

From the estimated ARIMAX$(1,1,1) \times (0,1,1)_7$ models and the estimated SSSM models for the five different sample data sets, it can also be seen that they share some similar model interpretations.

For instance, θ_7 of ARIMAX$(1,1,1) \times (0,1,1)_7$ for both sample data set 1 and 3 are close to 1 (see Table 7.15). As we discussed in chapter 5, this indicates that the weekly periodicity is very close to a weekly deterministic harmonic. This claim is also verified by the corresponding SSSM models as the estimated variance of the disturbance for the weekly periodic component is zero.

Although there is no obvious similarity between the trend from ARIMAX model and the trend from SSSM model, they both have the detrending filter $(1-B)^2$ in the models. From the SSSM model, we can see that the trend variation mainly comes from the level disturbance, $\sigma_\eta^2 > 0, \sigma_\zeta^2 = 0$ in the first half of the year (the sample data sets 1 and 2); whereas the trend variation is mainly contributed to from the slope, $\sigma_\eta^2 = 0, \sigma_\zeta^2 > 0$ in the second half of the year (the sample data sets 4 and 5).

Examining the variances of the weather effect component, σ_ξ^2, in the estimated SSSMs, we can see that the electricity load is not so dependent on weather condition $(\sigma_\xi^2 > 0)$ in the first half of a year as it is in the second half of a year. This implies that people, generally, are not consistently sensitive to temperature when it is getting cold in the first half of a year. In a contrary way, they are consistently sensitive to temperature when it is getting warm in the second half of a year. This is also reflected in the coefficient of the weather exogenous variable, λ, in the estimated ARIMAX models. This indicates that people try to save on their electricity bills by delaying as long as possible the use of their electrical heating appliances when the weather becomes cold as autumn/winter approaches; and by halting the use of their heating appliances as soon as possible when the weather becomes warmer as spring/summer arrives.

7.6 Summary

In this chapter, we mainly use a state space model based on the structural modelling approach proposed by Harvey (1989) for our daily load and weather data. The initial values of the state vector and its covariance matrix, which are often ignored, have warranted special attention by applying the techniques developed in chapter 3 to improve the model parameter estimation via the Kalman filter for a small sample data set. A hypothesis test scheme is proposed to specify the optimal state space model for the daily load data.

Judging by the overall performance of the ARIMAX and SSSM models both in

the sample and post sample data, we can conclude that the SSSM is the better model for the smaller sample size while the ARIMAX is the better model when the sample size is large. The reason for this is that the long term trend (natural growth trend and annual cycle) behaviour cannot be well approximated by a local linear model for a large sample of data, say several months.

Theoretically, the SSSM can be modified to suit the annual seasonal pattern but with very high system order when several years of daily data are available. Computing problems will however occur since it will requires considerable CPU time. It also seems a very inefficient way to produce a relatively short term prediction.

When we looked at how an ARIMAX model could be converted approximately to a structural state space form, it could be carefully noted that the associated state space model lays in a restricted domain. It appears that within that constrained domain it may not possible to model well the data investigated. This suggests that further research may be carried out on an ARIMAX model where the parameters are allowed to evolve over time and that this more general ARIMAX model may be more effective although it still has the disadvantage in lack of natural interpretation as the conventional ARIMAX model. This topic is beyond the content of this thesis.

Chapter 8

Conclusion and Suggestion

This thesis has presented a review of short-term electricity load modelling and forecasting in the literature. We divide the various models into two categories, namely, load data only, and load and weather data models. In each category, the different methods can be categorized into two approaches, i.e. the two stage time series approach and the state space approach.

Analysis and comparisons of different methods lead us to believe that it is very fruitful to divide the load into different components and for the load to comprise different components in additive form. In constructing a model for short-term load, the following main aspects have been investigated in this thesis.

1. Daily and weekly multiperiodic processes for the base component
2. A parsimonious model for the stochastic component
3. The relationship between weather variables and load demand, and hence the weather sensitive load
4. The influence of initial conditions in a state space model on model identification

In the theoretical work of this thesis, we focus on a subset AR model selection procedure in chapter 2 and the some properties of a state space model in chapter 3.

In chapter 2, the effects of deleting a lag from a full AR model, and sequentially, deleting a lag from a subset AR model have been intensively analyzed. We show

that the effects of deleting a lag are determined by the magnitude of its coefficient and by its representability by other remaining lags. As a result, we concluded that a balance should be struck between them to search for an optimal subset AR model efficiently. An efficient search procedure is proposed to find an optimal subset AR model. The results from applying this procedure to simulated data and real data show that our concern for "balance" is necessary. How to obtain an optimal balance needs still further study.

From Kalman filter theory, it is well known that both a diffuse initial state covariance matrix and a fixed initial state vector asymptotically lead to the same steady state vector covariance, i.e. a constant state covariance matrix if the state space model is detectable and controllable. The effect of an initial state vector and its covariance matrix on the convergence rate is neglected in the literature. However, the convergence rate does affect the model parameter estimation in a sample set with a limited data span. A faster convergence rate will lead to more accurate model parameter and state vector estimation than that associated with a slower convergence rate, where the sample set is not very large.

In chapter 3, we show that the state covariance matrix convergence rate from an over-estimated initial value is faster than that from an under-estimated initial state covariance matrix. A procedure for the estimation of the initial state condition is proposed in section 3.4 based on fixed point smoothing, which ensures that we are able to identify the state space model from a small sample data set. After analyzing the effect of the error-sensitivity of the state disturbance term, we conclude that conservative initial estimates for a state vector covariance matrix, the covariance matrix of the disturbance terms of the state equation and the observation equation, yield a conservative state covariance if the system matrices are known. Therefore, we proposed in section 3.6 an off-line recursive procedure to estimate for a state space model the system parameters (the covariance matrices of disturbance terms), the state vector and its covariance matrix. This procedure and the way we specify the initial values avoid those problems which may cause model mis-identification. For

general cases, there are some unknown elements in the system matrices. Therefore, a further study should be carried out to investigate how the initial estimates for the unknown elements in the system matrices affect the estimation of the state vector, its conditional covariance matrix via Kalman filtering, and the convergence of the maximum likelihood estimation for these elements and the disturbance covariance matrix. This investigation safeguards the correctness of system identification by maximum likelihood via Kalman filtering (i.e. estimation of unknown elements of the system matrices, the state vector and its covariance matrix, the covariance matrices of the state and observation disturbance terms).

In the practical work aspect of this thesis, we propose a new modelling procedure based only on the load data in chapter 4 under the assumption that the load variation follows an inherent law and is only dependent on time when weather information is not available (i.e. the exogenous weather effect is ignored). This model divides the short-term load into three components. They are trend, periodic, and stationary stochastic components. Beginning with this frame-work, a cointegration –"error correction" regression model has been proposed to estimate and to forecast the trend behaviour. This cointegration model uses relatively long-term (weekly) data to help in modelling the short-run trend. Theorem 4.1 is a theoretical result for the cointegration – "error correction" regression model proposed by Engle et al. (1989). It also proves that our cointegration – "error correction" model is a generalization of the Engle et al. (1989) model and has a better short-run prediction performance.

The detrended data, obtained by subtracting the estimated trend component from the load, is a periodic stationary time series. It presents different daily periodical patterns for weekdays and weekend days. A new modelling procedure developed in chapter 5 treats data for weekdays and weekend days separately. The link between them is a transition process so that both daily and weekly periods in the load process can be accounted for. The advantages of this approach are: (1) avoiding the use of very large sample data sets and many frequency components to model the daily and weekly periodic patterns in a discrete harmonic frequency model; (2) giving a better

understanding of the load profiles in weekdays and weekend days and how weekday load profiles evolve to weekend load profiles and vice versa. The innovation series are obtained by subtracting the estimated periodic component from the detrended data. As a result, the innovation series has no significant daily and weekly periodicity and can be assumed to be a stationary series.

The stationary innovation series can then be approximately identified automatically by the procedure proposed in chapter 2 as a parsimonious subset AR model with an order less than the daily periodic order.

The whole procedure of modelling and its identification is processed automatically after a few instructions are given. This feature means that the identified model is adequate and parsimonious and also means that the user only has to make minimal a priori assumptions to initial the procedure.

After applying the proposed new approach to New Zealand half-hourly electricity load data and utilizing the associated modelling procedures for different components in chapter 5, the overall performance of the proposed modelling procedure is very promising in both sample fitting and post sample forecasting by comparison with other popular methods.

For the load and weather model, when weather information is available, we intensively investigate the relationship between load and weather variables in chapter 6. Based on the evidence of the non-linearity between them, a non-linear functional relationship between the load and a temperature-humidity index is established. The model accuracy and stability had to be well balanced to identify this relation. As a result, a weather sensitive load variable is derived from the estimated relations and is linearly related to the load. The newly created variable is now more appropriate than the temperature-humidity index as an exogenous variable in any linear system describing the load behaviour.

We also found that the parameters of the established functional relationship depend on the time of a day and the type of day in a week from 3 hourly based data. If we wish to resolve accurately how this relation evolves over the day, we must have

available weather data measured at time intervals much shorter than 3 hours.

A structural state space model has been built for daily electricity demand in chapter 7. A procedure is provided to estimate the model's initial state which has not been given much attention in the literature. The estimated state using this procedure provides assistance in speeding up convergence of the Riccati difference equation for a limited data span. The model parameters and the state vector are estimated quickly and accurately. To test and specify an optimal model for the load data, a hypothesis testing procedure has been proposed.

After applying the proposed structural state space model to the data of daily electricity load for Canberra, Australia, and comparing its overall performance, both in sample and post sample, with an ARIMAX model, we conclude that the proposed structural state space model is better for the smaller sample size although ARIMAX is better when the sample size is large. The reason for this is that the behaviour of the trend (which includes the natural growth trend and the annual cycle) cannot be handled properly by a local linear model in structural state space, when the sample data size is sufficiently large that it entails considerable seasonal variation.

A state space model for very short term load data (i.e. quarter-hourly, half-hourly or hourly data) must have a very high order to cope with daily and weekly periodic variation, there are therefore likely to be too many parameters to be estimated. Great computing difficulties, related to memory size, CPU time and cost etc will arise. It is also a very inappropriate way to create relatively short term predictions, since such predictions must have wide prediction confidence intervals given the many unknown parameters involved. We suggest that state space models be developed for data sets observed at different intervals, such as hourly, daily, and weekly, etc. A model based on one time interval, say weekly, can be utilized to establish the "long term trend" in a model based on data observed daily. This same method can also be used when a daily model is now utilized to establish the "long term trend" for a model using hourly data. In this way, a state space model with multiple time scales, which contains several sub-state space models interacting with each other, can be setup to integrate

the load variations as observed at different time intervals.

It should be mentioned that modelling stochastic load is, in general, a quite complicated procedure. Therefore, good intuition is needed in searching for a good model specification. For instance, the form of the micro model for different components of load in a structural state space model is not easily specified. In fact, we have to choose one form from a group candidate models. The most obvious problem with such a structural state space modelling approach is that it is not easy to specify the covariance matrix structure of the disturbance term of the state vector. A complicated structural state space model creates great difficulty in testing all the hypotheses needed to obtain an optimal model structure. The sequential hypothesis testing procedure presented in chapter 7 handles the most interesting models. As all possible models are not handled we obtain the best model of those considered using a particular criterion. This means there must be an element of subjectivity in this aspect.

Besides the improvement of modelling techniques for a regional short-term electricity load, the major source of further improvement of short-term forecasting will come from the proper knowledge of the various sources of demand and the generation of the load. For example, industrial, business, agriculture and residential load demands have their own particular characteristics associated with weather, time of a day, day of a week, season of a year and stochastic properties. If the load data from the major consumer categories in a region were provided, we believe that the opportunity to deal with them separately would allow better understanding of the total load and thus achieve more accurate forecasting.

Appendix A

An Example for Chapter 3

For example the random walk plus noise model

$$\begin{cases} x(t+1) = x(t) + d^{1/2}\xi(t+1) \\ y(t) = x(t) + \epsilon(t) \end{cases} \tag{A.1}$$

where both $\xi(t)$ and $\epsilon(t)$ are scalar white noise disturbances with zero mean and variance 1, and they are independent.

The RDE is

$$\sigma^2(t+1) = [1 - k(t)]^2\sigma^2(t) + k(t)^2 + d \tag{A.2}$$

where $k(t) = \sigma^2(t)[\sigma^2(t) + 1]^{-1}$, $\sigma^2(t) = E(x(t) - x(t|t-1))^2$. The model is always observable (detectable) but it is not stabilisable when $d = 0$. The corresponding ARE is

$$\begin{aligned} & \sigma^2 = [1 - k]^2\sigma^2 + k^2 + d, \text{ where } k(t) = \sigma^2[\sigma^2 + 1]^{-1} \\ \Rightarrow \quad & (\sigma^2)^3 + (1-d)(\sigma^2)^2 - 2d\sigma^2 - d = 0 \end{aligned} \tag{A.3}$$

It can easily be proved that the ARE equation (A.3) has an unique positive solution when $d > 0$ as follows

$$\sigma^2 = 2\sqrt{-q}\cos(\frac{1}{3}\theta) - \frac{1}{3}(1-d) \tag{A.4}$$

where $q = -\frac{3d+(1-d)^2}{9}$, $r = \frac{9d+18d^2+2d^3}{54}$, and $\cos(\theta) = r/\sqrt{-q^3}$.

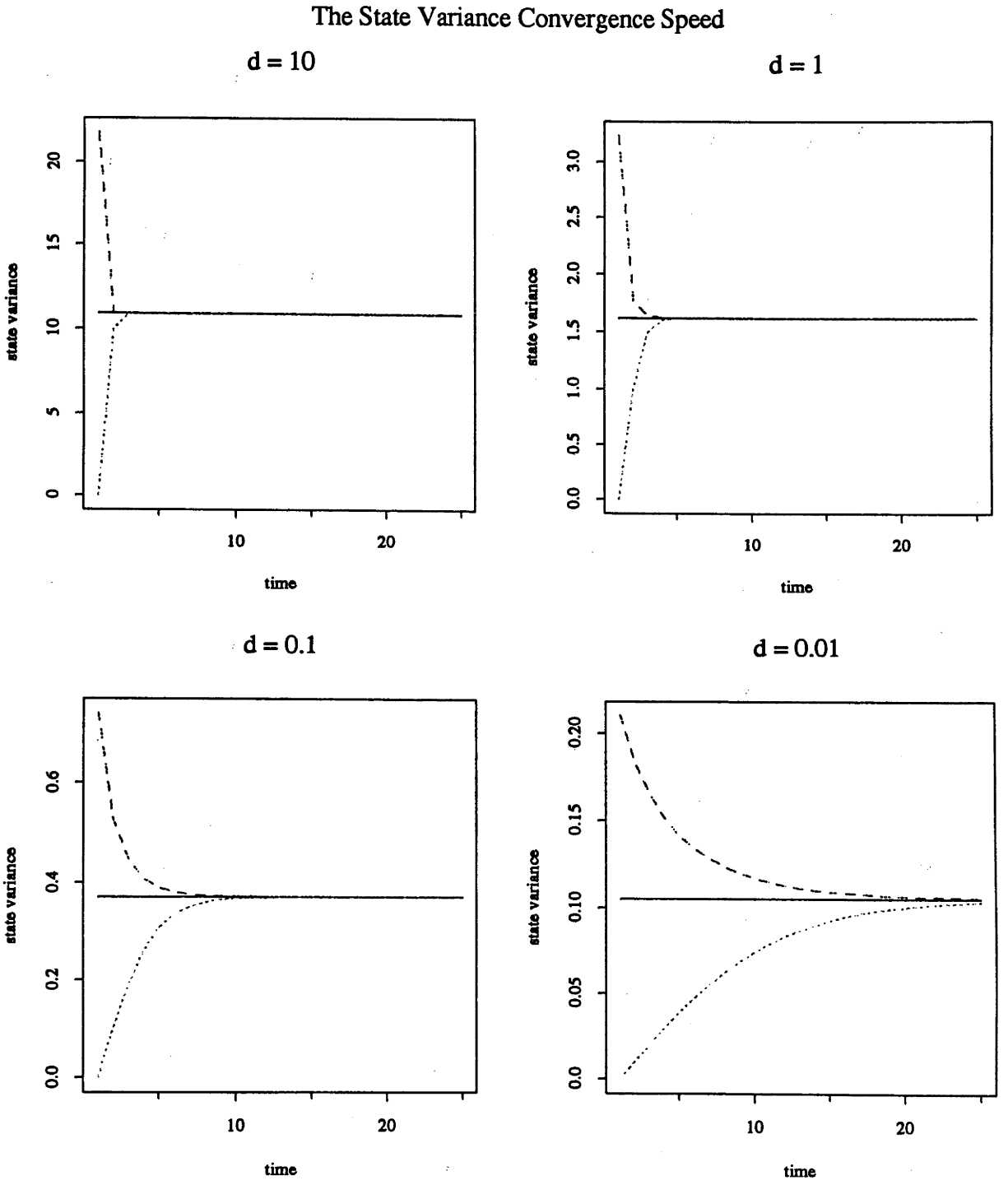


Figure A.1: The Convergence Speed of $\Sigma(t)$ when $\hat{\Sigma}(0)$ is Over- or Under-estimated

That the over-estimated initial state variances cause faster convergence than the under-estimated ones is clearly shown in figure A.1 when $d = 10,\ 1,\ 0.1,\ 0.01$.

Therefore, we can conclude that the partial information on initial conditions of the state space model (3.1) can be used to speed up the convergence to the steady state if the model is stabilizable and detectable. In general, a large initial state covariance matrix converges faster than a small initial one. Therefore, an initial conservative estimate for the state disturbance variance is reasonable and applicable.

Appendix B

Tables and Figures for Chapter 5

B.1 Model 1

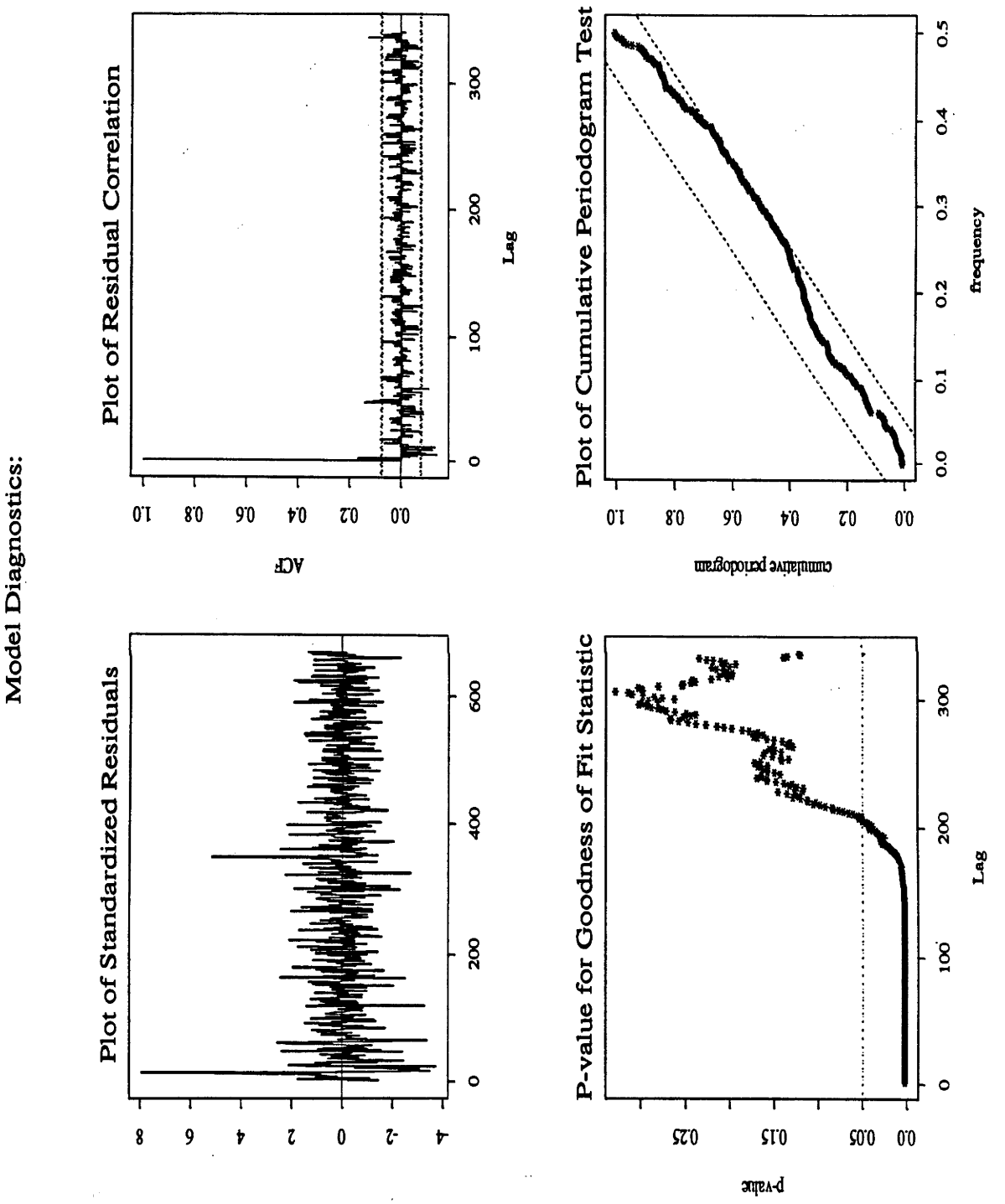


Figure B.1: Model 1: Model Fit Diagnostics for the Autumn Data Set

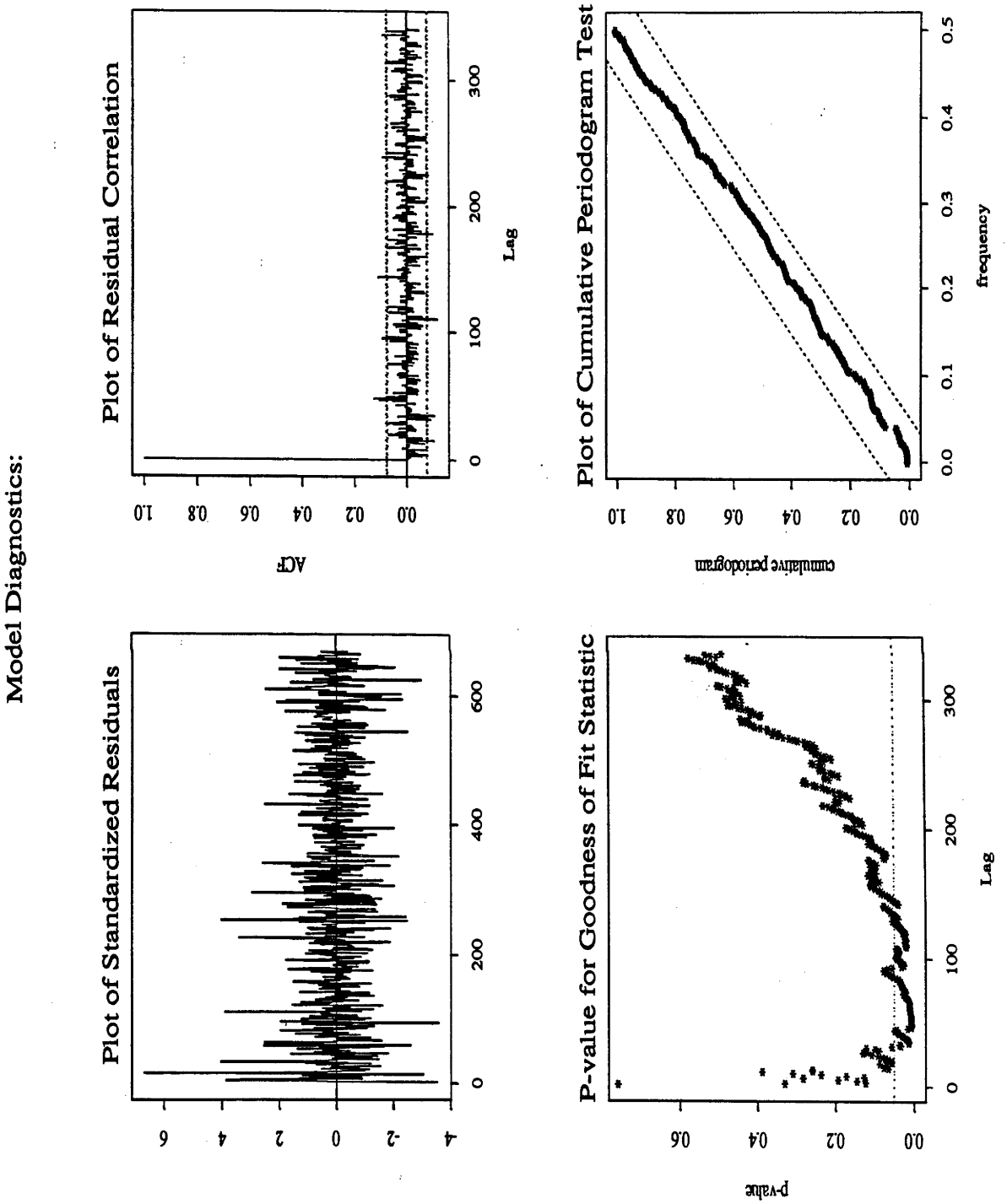


Figure B.2: Model 1: Model Fit Diagnostics for the Winter Data Set

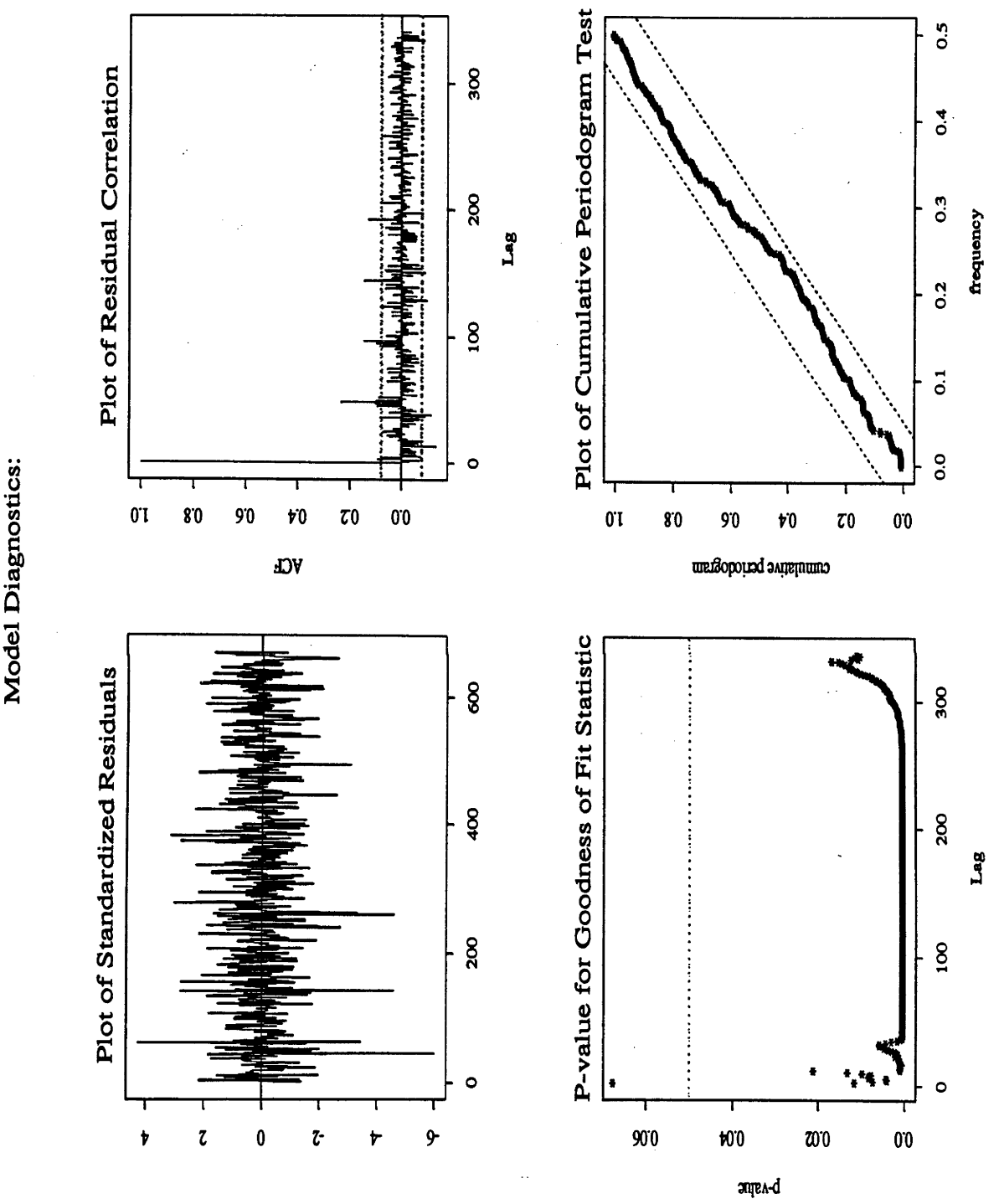


Figure B.3: Model 1: Model Fit Diagnostics for the Spring Data Set

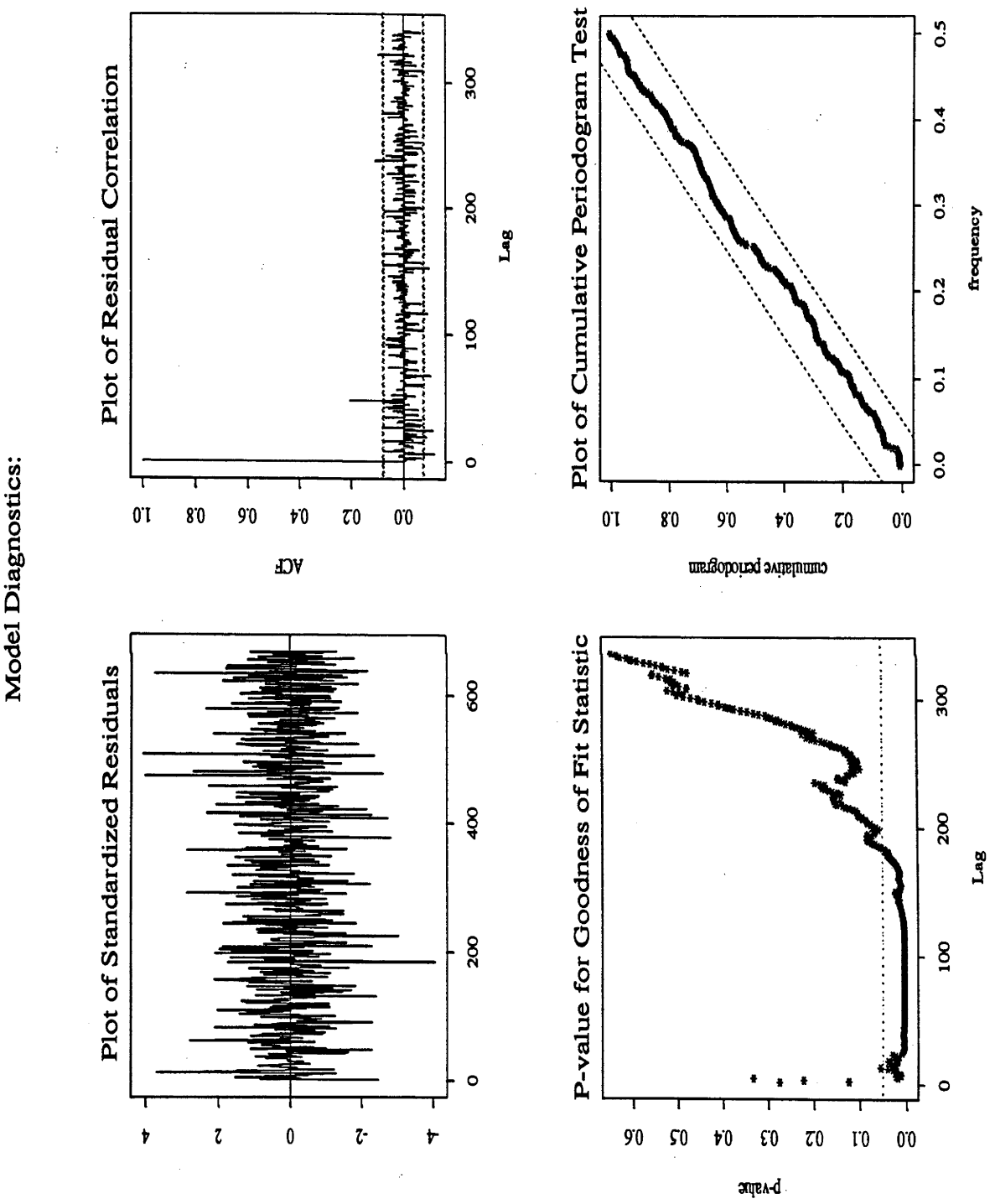


Figure B.4: Model 1: Model Fit Diagnostics for the Summer Data Set

B.2 Model 2

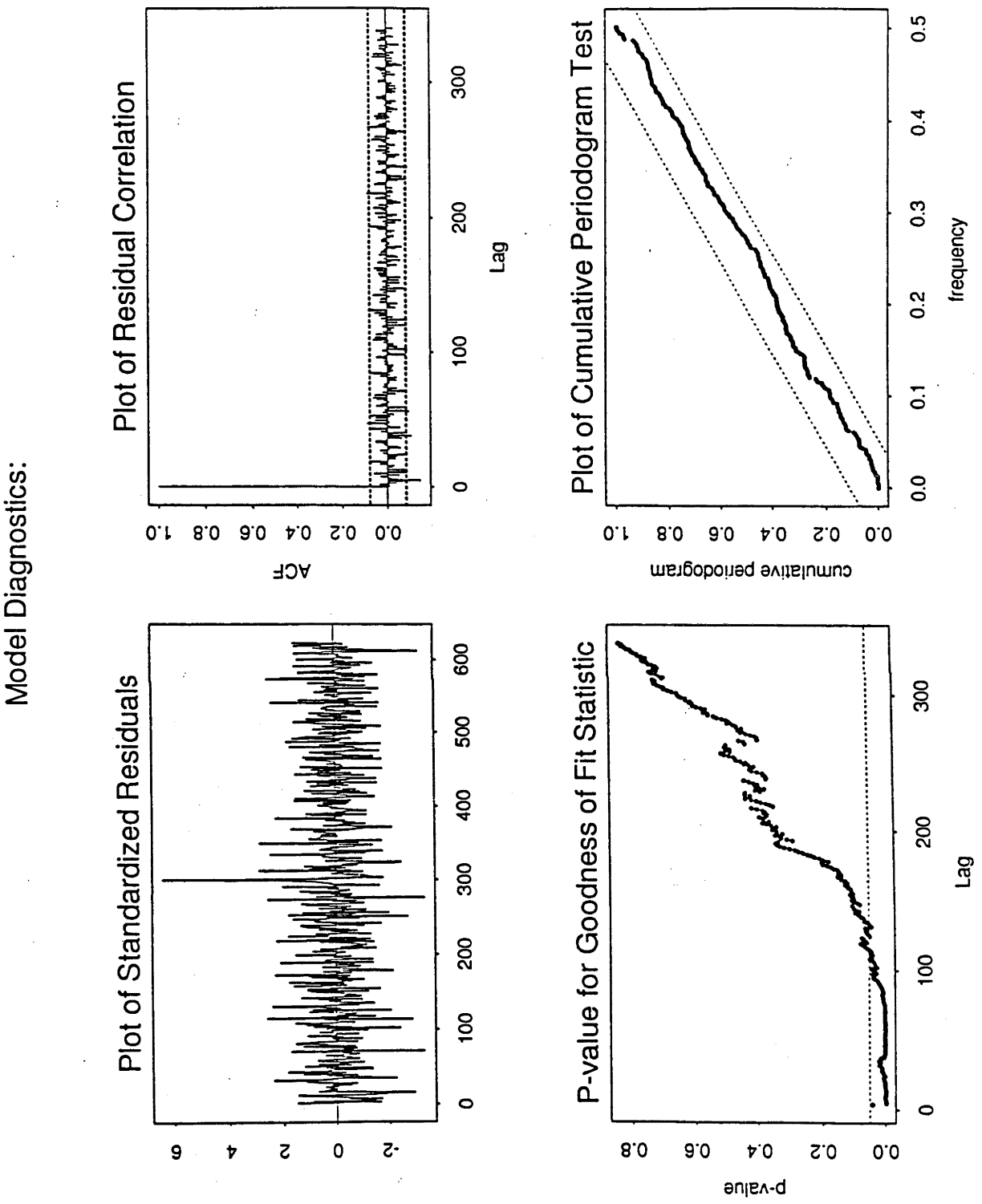


Figure B.5: Model 2: Model Fit Diagnostics for the Autumn Data Set

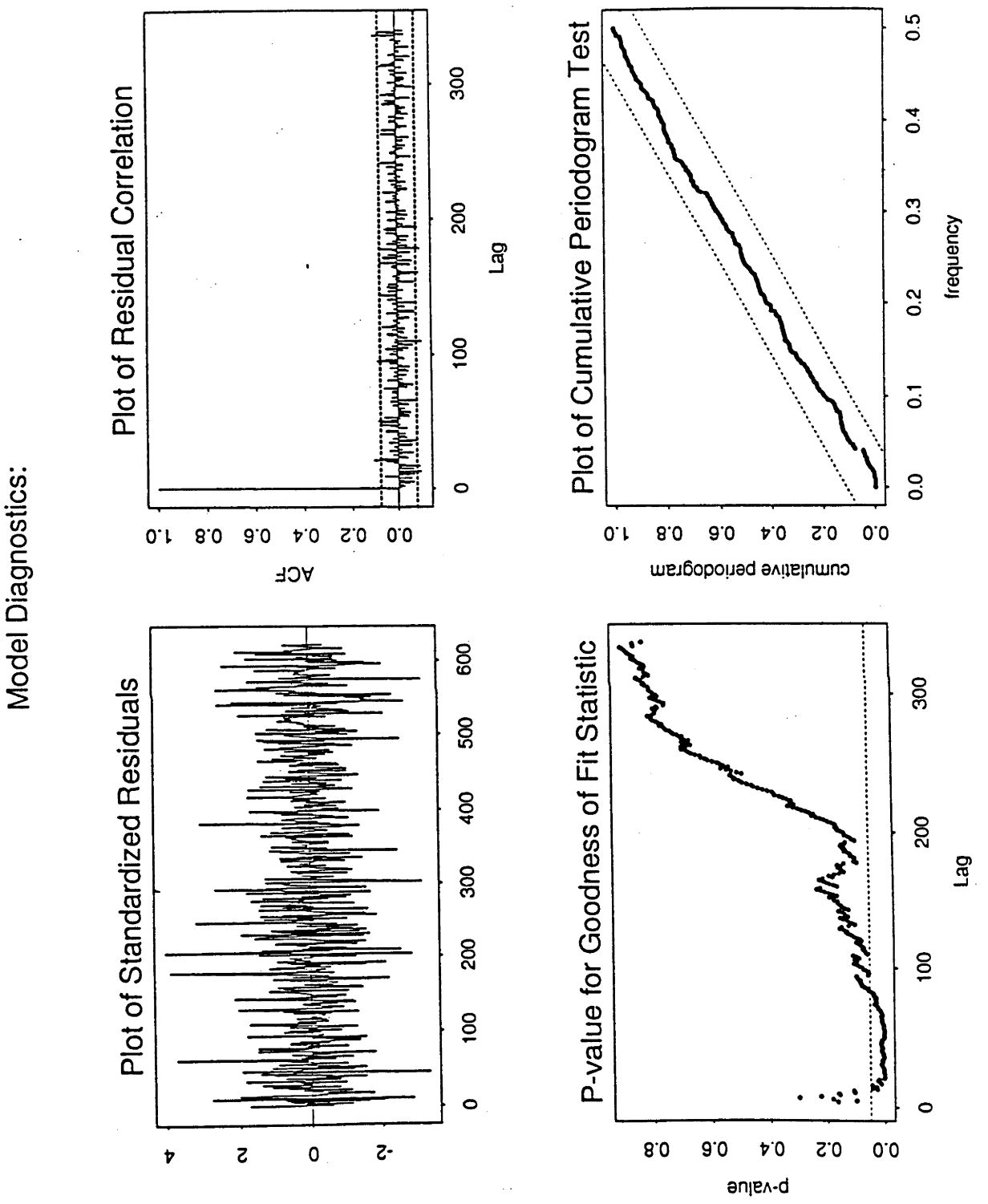


Figure B.6: Model 2: Model Fit Diagnostics for the Winter Data Set

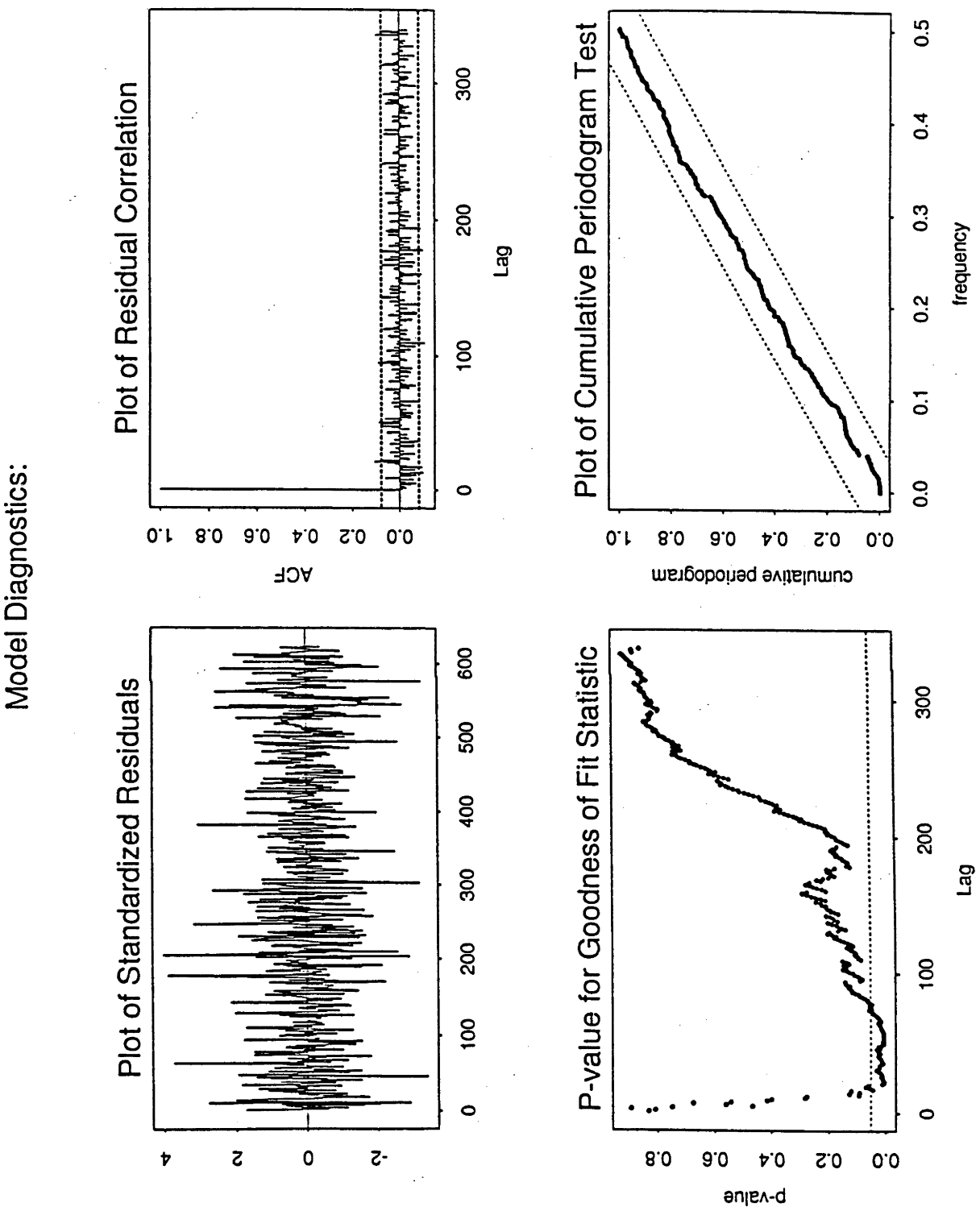


Figure B.7: Model 2: Model Fit Diagnostics for the Spring Data Set

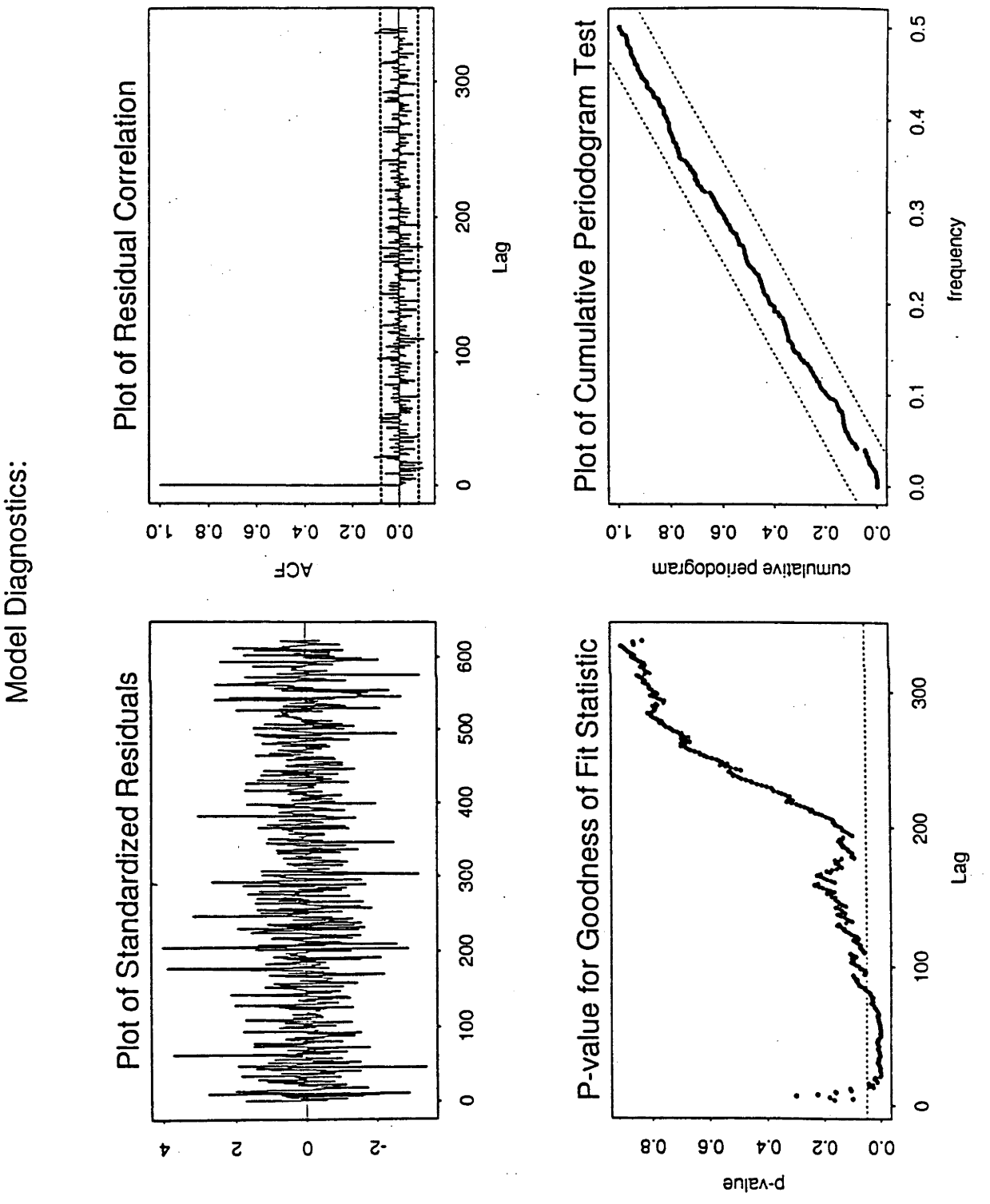


Figure B.8: Model 2: Model Fit Diagnostics for the Summer Data Set

B.3 Model 3

lag	coeff.	s.e.	t-ratio
	Long Memory Filter		
48	-0.31945175	0.51328354E-02	-62.236897
336	-0.66333479	0.51328354E-02	-129.23360
	Short Memory Filter		
1	-1.4157610	0.37876885E-01	-37.377968
2	0.41831249	0.51374629E-01	8.1423941
4	0.10509685	0.36171526E-01	2.9055135
48	0.18177316	0.49306516E-01	3.6865950
49	-0.10303809	0.37450057E-01	-2.7513466
size of 7, AIC = -9042.985, σ_l^2 = 0.1252618E-3			

Table B.1: Subset ARAR Model Fitting the Autumn Data Set

lag	coeff.	s.e.	t-ratio
	Long Memory Filter		
48	-0.18154544	0.13682197E-02	-132.68733
336	-0.82897669	0.13682197E-02	-605.87982
	Short Memory Filter		
1	-1.0273980	0.72239906E-01	-14.222029
2	0.13362987	0.72291248E-01	1.8484931
18	-0.30434862E-01	0.29222472E-01	-1.0414883
46	-0.40687922E-01	0.28778533E-01	-1.4138290
size of 6, AIC = -8956.618, σ_l^2 = 0.1367387E-3			

Table B.2: Subset ARAR Model Fitting the Winter Data Set

lag	coeff.	s.e.	t-ratio
	Long Memory Filter		
48	-0.20217681	0.25643010E-02	-78.842850
336	-0.80233228	0.25643010E-02	-312.88538
	Short Memory Filter		
1	-1.0132995	0.17830238E-01	-56.830395
5	0.77929199E-01	0.17064037E-01	4.5668678
45	-0.11167498	0.17064037E-01	-6.5444641
49	0.73816732E-01	0.17830238E-01	4.1399746
size of 6, AIC = -9120.78, σ_l^2 = 0.1161884E-3			

Table B.3: Subset ARAR Model Fitting the Spring Data Set

lag	coeff.	s.e.	t-ratio
	Long Memory Filter		
48	-0.19982231	0.24771870E-02	-80.665009
336	-0.80134100	0.24771870E-02	-323.48831
	Short Memory Filter		
1	-1.1507533	0.38743131E-01	-29.702124
2	0.18511222	0.38828932E-01	4.7673788
42	-0.64948291E-01	0.14507797E-01	-4.4767852
49	0.45982625E-01	0.14567110E-01	3.1566060
size of 6, AIC = -9384.542, σ_l^2 = 0.0894379E-3			

Table B.4: Subset ARAR Model Fitting the Summer Data Set

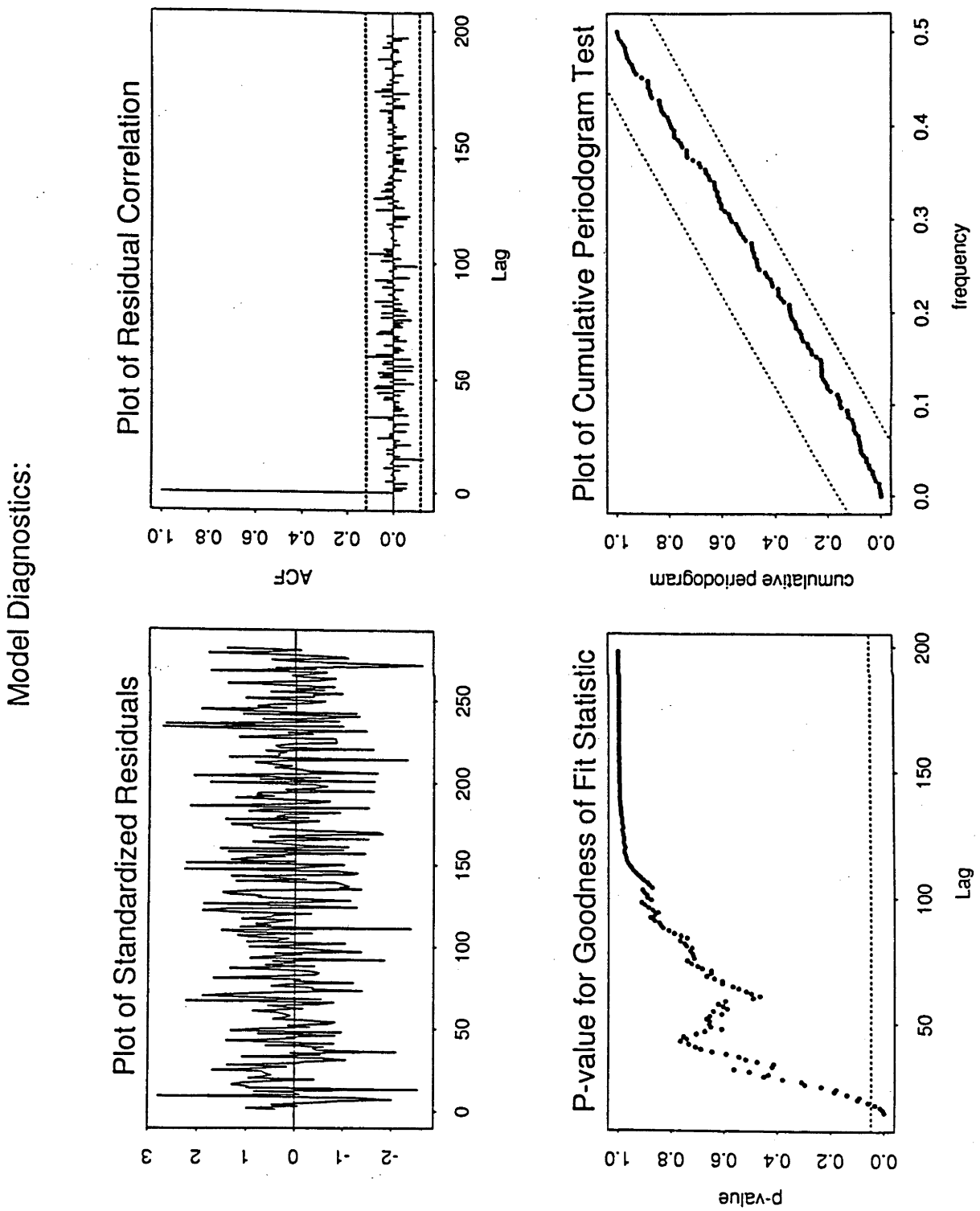


Figure B.9: Model 3: Model Fit Diagnostics for the Autumn Data Set

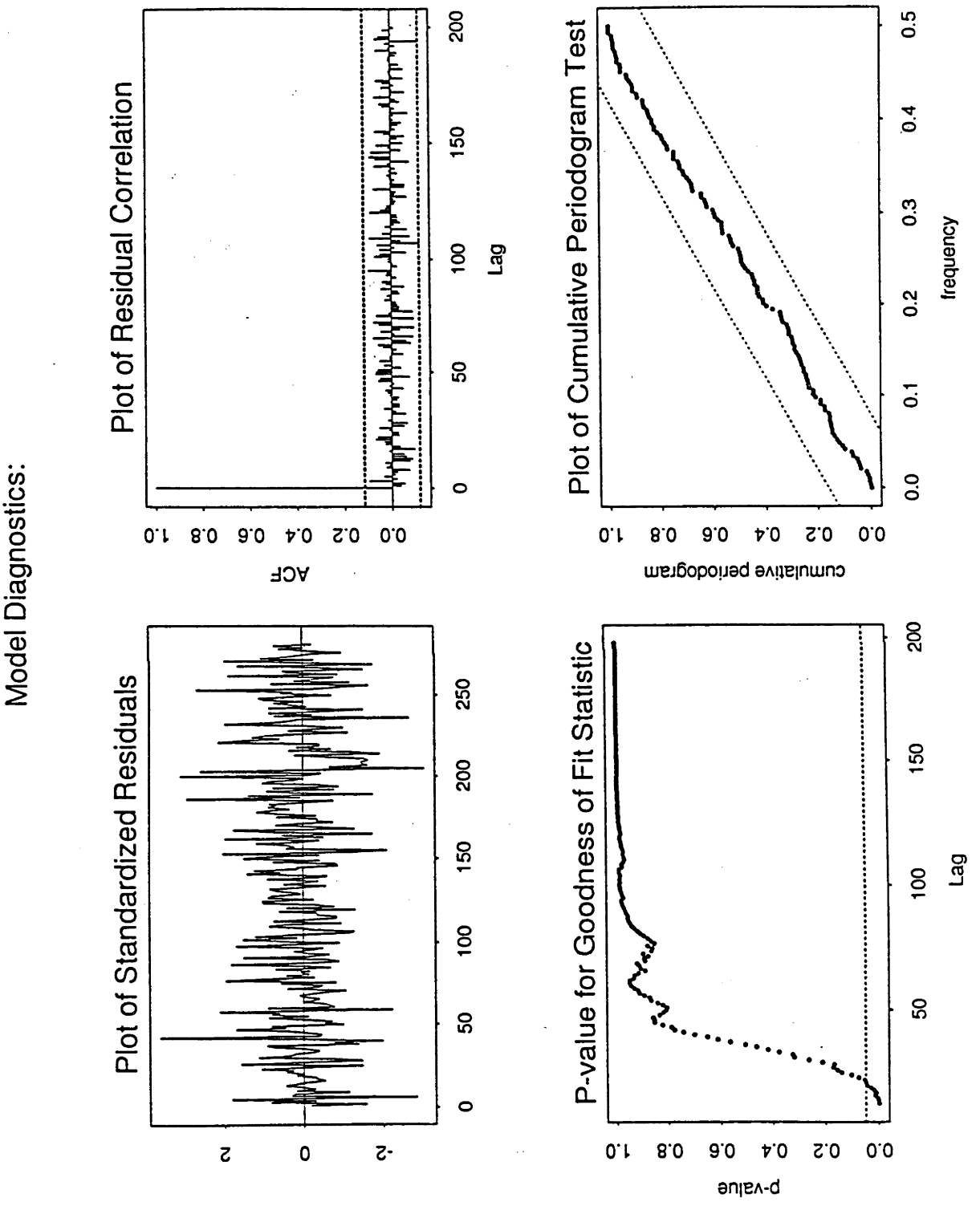


Figure B.10: Model 3: Model Fit Diagnostics for the Winter Data Set

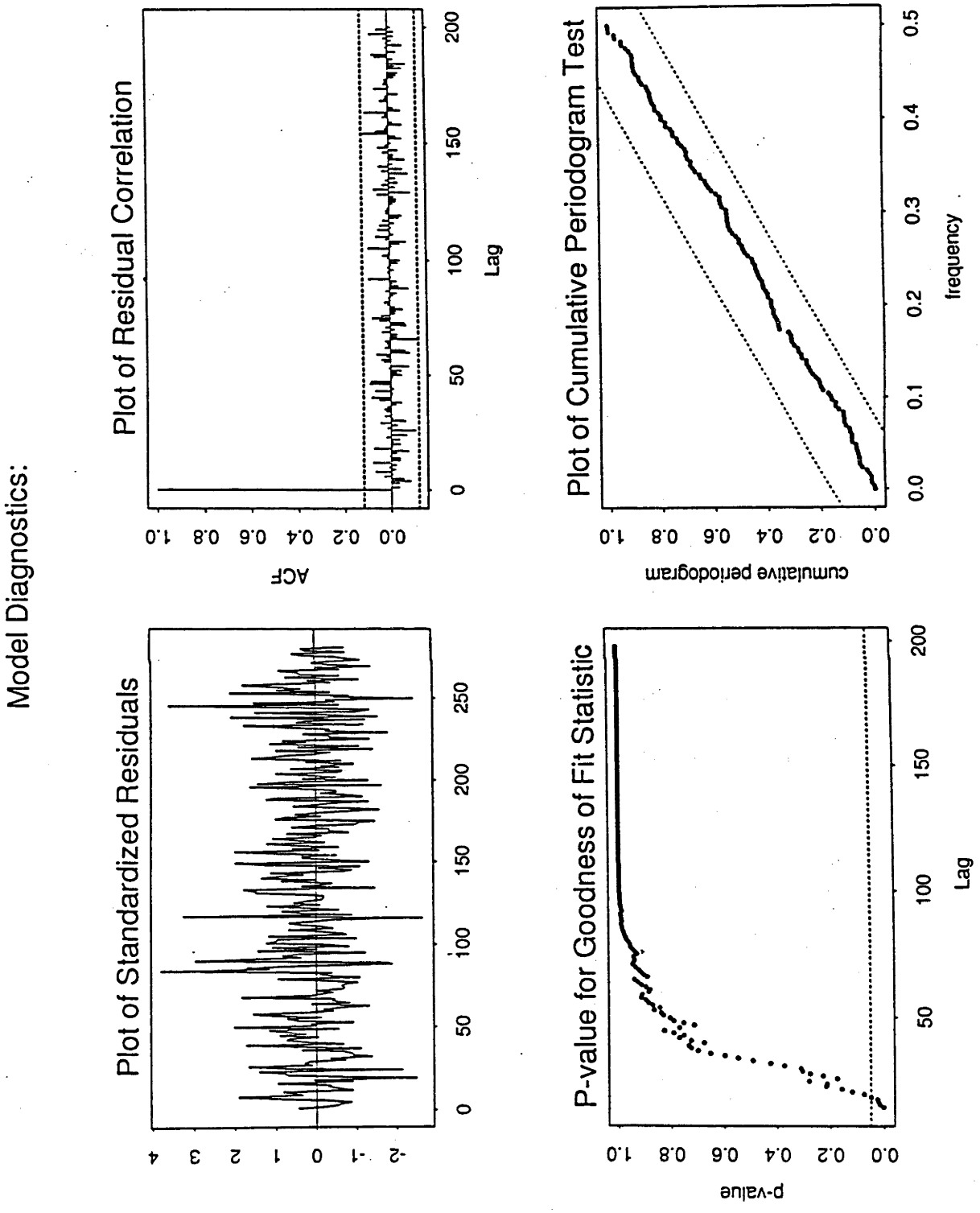


Figure B.11: Model 3: Model Fitting Diagnostics for the Spring Data Set

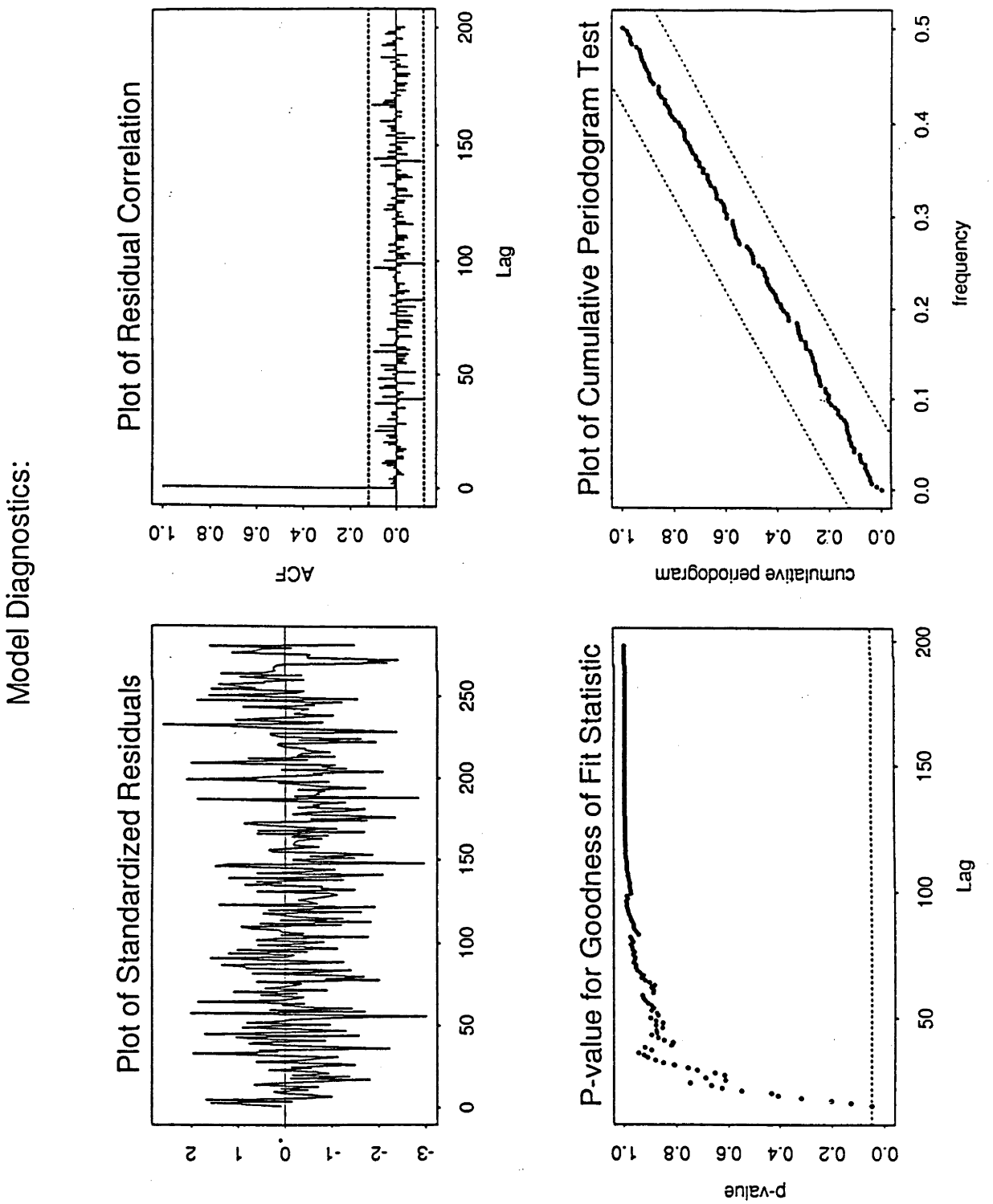


Figure B.12: Model 3: Model Fitting Diagnostics for the Summer Data Set

B.4 Model 4

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13092	47.99190	-0.212E+00	-0.195E+00	-0.741E+00	0.679E+04
2	0.26179	24.00104	-0.418E-01	-0.173E+00	-0.133E+01	0.259E+04
3	0.39310	15.98352	0.233E-01	0.304E-01	-0.917E+00	0.120E+03
4	0.52408	11.98904	-0.207E-01	0.265E-01	0.907E+00	0.925E+02
5	0.65486	9.59467	-0.239E-01	-0.301E-01	-0.898E+00	0.121E+03
6	0.78477	8.00638	0.869E-03	-0.860E-02	0.147E+01	0.612E+01
7	0.91710	6.85112	0.208E-03	0.126E-01	-0.155E+01	0.131E+02
residual variance = 0.527464E-03 signal-to-noise ratio = 0.486864E+05						

Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13074	48.05746	-0.138E+00	-0.215E+00	-0.998E+00	0.140E+04
2	0.26154	24.02351	-0.380E-02	-0.146E+00	-0.154E+01	0.457E+03
3	0.39325	15.97775	-0.300E-01	0.550E-01	0.107E+01	0.841E+02
4	0.65359	9.61330	-0.317E-02	-0.261E-01	-0.145E+01	0.148E+02
residual variance = 0.134516E-02 signal-to-noise ratio = 0.976302E+04						

Table B.5: Seasonal Components for Weekdays & Weekend Days from the Autumn Data Set

Subset AR Model for the Stochastic Component sample variance = 0.93083405E-03			
lag	coeff.	s.e.	t-ratio
1	-1.0809759	0.57830524E-01	-18.692135
2	0.26107142	0.67707618E-01	3.8558648
4	-0.86607029E-01	0.43140885E-01	-2.0075395
9	0.55338495E-01	0.32690020E-01	1.6928254
16	-0.50712684E-01	0.28717567E-01	-1.7659116
44	0.55157990E-01	0.31945978E-01	1.7266020
48	-0.28097338	0.60661377E-01	-4.6318331
size of 8, Schwarz criterion = -8421.6795 residual variance = 0.12607818E-03			

Table B.6: Subset AR Fit for the Stochastic Component from the Autumn Data Set

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13088	48.00767	-0.209E+00	-0.189E+00	-0.734E+00	0.742E+04
2	0.26183	23.99737	-0.254E-01	-0.199E+00	-0.144E+01	0.376E+04
3	0.39183	16.03540	0.285E-01	0.115E-01	-0.382E+00	0.885E+02
4	0.52374	11.99682	-0.290E-01	0.252E-01	0.716E+00	0.138E+03
5	0.65420	9.60433	-0.351E-01	-0.220E-01	-0.560E+00	0.160E+03
6	0.78534	8.00055	0.558E-02	-0.695E-02	0.894E+00	0.743E+01
7	0.91603	6.85915	0.670E-02	0.141E-01	-0.113E+01	0.229E+02
8	1.04666	6.00305	-0.102E-01	0.111E-02	0.108E+00	0.978E+01
residual variance = 0.461723E-03 signal-to-noise ratio = 0.580076E+05						
Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13005	48.31269	-0.175E+00	-0.233E+00	-0.926E+00	0.124E+04
2	0.26219	23.96415	0.237E-01	-0.124E+00	0.138E+01	0.234E+03
3	0.39235	16.01418	-0.219E-01	0.560E-01	0.120E+01	0.527E+02
4	0.65418	9.60461	-0.187E-02	-0.227E-01	-0.149E+01	0.760E+01
residual variance = 0.197237E-02 signal-to-noise ratio = 0.767306E+04						

Table B.7: Seasonal Components for Weekdays & Weekend Days from the Winter Data Set

Subset AR Model for the Stochastic Component sample variance =0.11242434E-02			
lag	coeff.	s.e.	t-ratio
1	-1.2483658	0.52642097E-01	-23.714212
2	0.50294812	0.77987221E-01	6.4491093
3	-0.13708463	0.52765354E-01	-2.5980046
46	0.49311855E-01	0.32203231E-01	1.5312704
48	-0.40916751	0.56937321E-01	-7.1862796
49	0.45759057	0.58513112E-01	7.8203082
51	-0.10173585	0.34597367E-01	-2.9405662
size of 7, Schwarz criterion = -8297.7979 residual variance = 0.14829228E-03			

Table B.8: Subset AR Fit for the Stochastic Component from the Winter Data Set

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13098	47.97228	-0.200E+00	-0.195E+00	-0.774E+00	0.784E+04
2	0.26170	24.00924	-0.534E-01	-0.189E+00	-0.130E+01	0.389E+04
3	0.39283	15.99478	0.231E-01	0.191E-01	-0.691E+00	0.904E+02
4	0.52358	12.00050	-0.246E-01	0.346E-01	0.952E+00	0.182E+03
5	0.65419	9.60457	-0.280E-01	-0.156E-01	-0.507E+00	0.104E+03
6	0.78462	8.00790	0.426E-02	-0.751E-02	0.105E+01	0.751E+01
7	0.91677	6.85365	0.330E-02	0.951E-02	-0.124E+01	0.102E+02
8	1.04613	6.00611	-0.755E-02	0.465E-02	0.552E+00	0.792E+01
residual variance = 0.428889E-03 signal-to-noise ratio = 0.606932E+05						

Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13067	48.08356	-0.139E+00	-0.201E+00	-0.966E+00	0.933E+03
2	0.26146	24.03098	-0.150E-01	-0.150E+00	-0.147E+01	0.354E+03
3	0.39189	16.03323	-0.131E-01	0.454E-01	0.129E+01	0.350E+02
residual variance = 0.183822E-02 signal-to-noise ratio = 0.661305E+04						

Table B.9: Seasonal Components for Weekdays & Weekend Days the Spring Data Set

Subset AR Model for the Stochastic Component Sample variance = 0.90805797E-03			
lag	coeff.	s.e.	t-ratio
1	-0.96165961	0.30865003E-01	-31.156958
5	0.13073248	0.38941287E-01	3.3571691
8	-0.13058645	0.43027698E-01	-3.0349393
11	0.85341818E-01	0.35726339E-01	2.3887647
17	-0.58554643E-01	0.30491837E-01	-1.9203383
22	0.42958111E-01	0.28896459E-01	1.4866220
48	-0.33046334	0.58291620E-01	-5.6691398
49	0.33913709	0.58426273E-01	5.8045306
size of 9, Schwarz criterion = -8507.3116 residual variance = 0.11718928E-03			

Table B.10: Subset AR Fit for the Stochastic Component from the Spring Data Set

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13127	47.86368	-0.179E+00	-0.164E+00	-0.744E+00	0.665E+04
2	0.26175	24.00466	0.686E-02	-0.129E+00	0.152E+01	0.190E+04
3	0.39340	15.97160	0.285E-01	0.172E-01	-0.541E+00	0.125E+03
4	0.52451	11.97914	-0.185E-01	-0.821E-02	-0.417E+00	0.463E+02
5	0.65478	9.59584	-0.219E-01	-0.255E-01	-0.862E+00	0.127E+03
6	0.91697	6.85213	-0.679E-03	0.112E-01	0.151E+01	0.142E+02
residual variance = 0.382932E-03 signal-to-noise ratio = 0.442760E+05						
Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13072	48.06763	-0.113E+00	-0.136E+00	-0.876E+00	0.665E+03
2	0.26142	24.03470	0.302E-01	-0.107E+00	0.130E+01	0.264E+03
3	0.39274	15.99832	-0.823E-02	0.262E-01	0.127E+01	0.160E+02
4	0.65681	9.56620	-0.725E-02	-0.173E-01	-0.117E+01	0.747E+01
residual variance = 0.135655E-02 signal-to-noise ratio = 0.476127E+04						

Table B.11: Seasonal Components for Weekdays & Weekend Days from the Summer Data Set

Subset AR Model for the Stochastic Component sample variance = 0.73936741E-03			
lag	coeff.	s.e.	t-ratio
1	-1.1343243	0.63352751E-01	-17.904894
2	0.27796562	0.64788312E-01	4.2903667
7	-0.81196143E-01	0.30688312E-01	-2.6458328
28	0.48197568E-01	0.28719812E-01	1.6781993
38	-0.52023094E-01	0.31363502E-01	-1.6587144
48	-0.24104069	0.64305567E-01	-3.7483643
49	0.27065095	0.63069347E-01	4.2913232
size of 7, Schwarz criterion = -8810.6824 residual variance = 0.85961787E-04			

Table B.12: Subset AR Fit for the Stochastic Component from the Summer Data Set

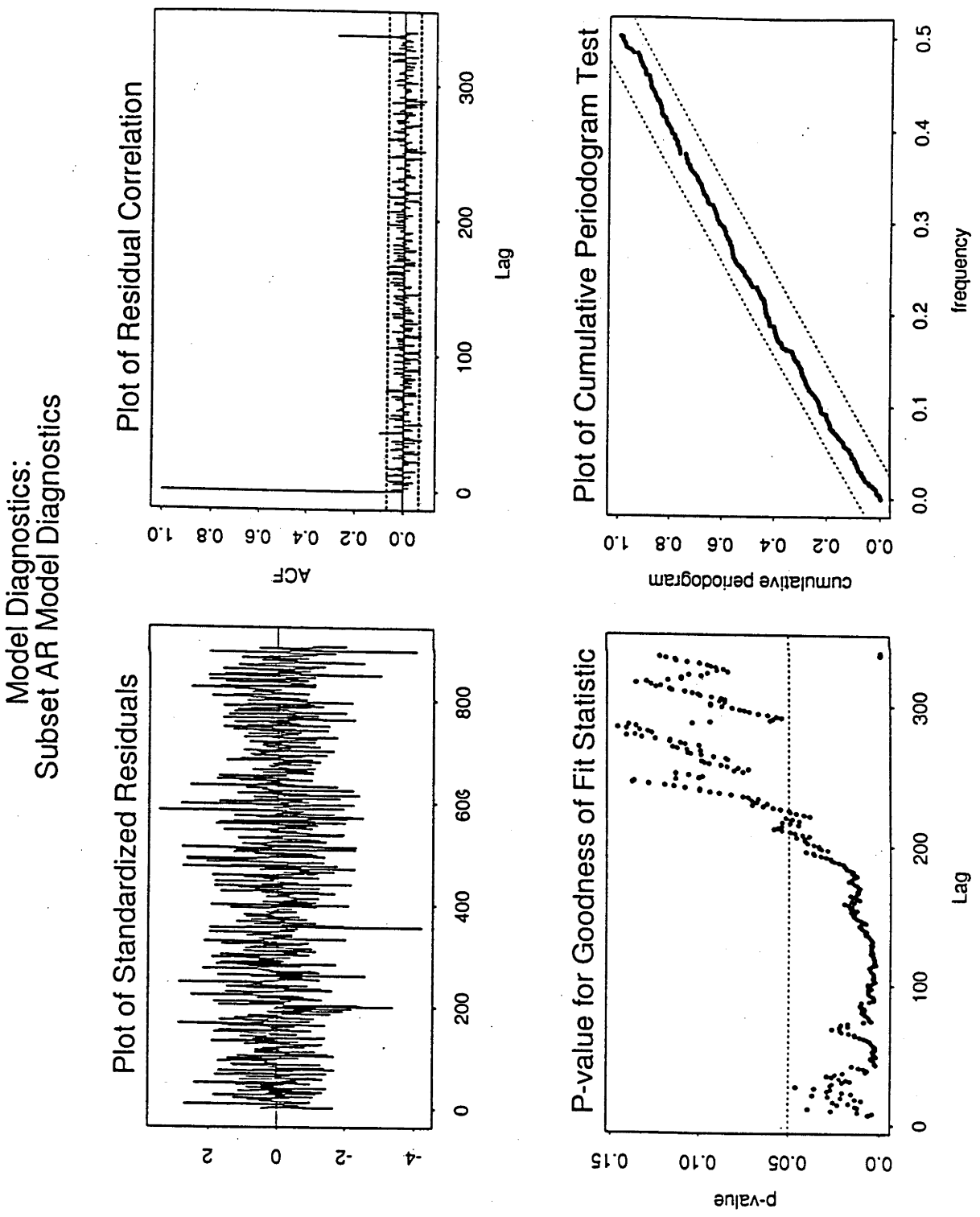


Figure B.13: Model 4: Model Fit Diagnostics for the Autumn Data Set

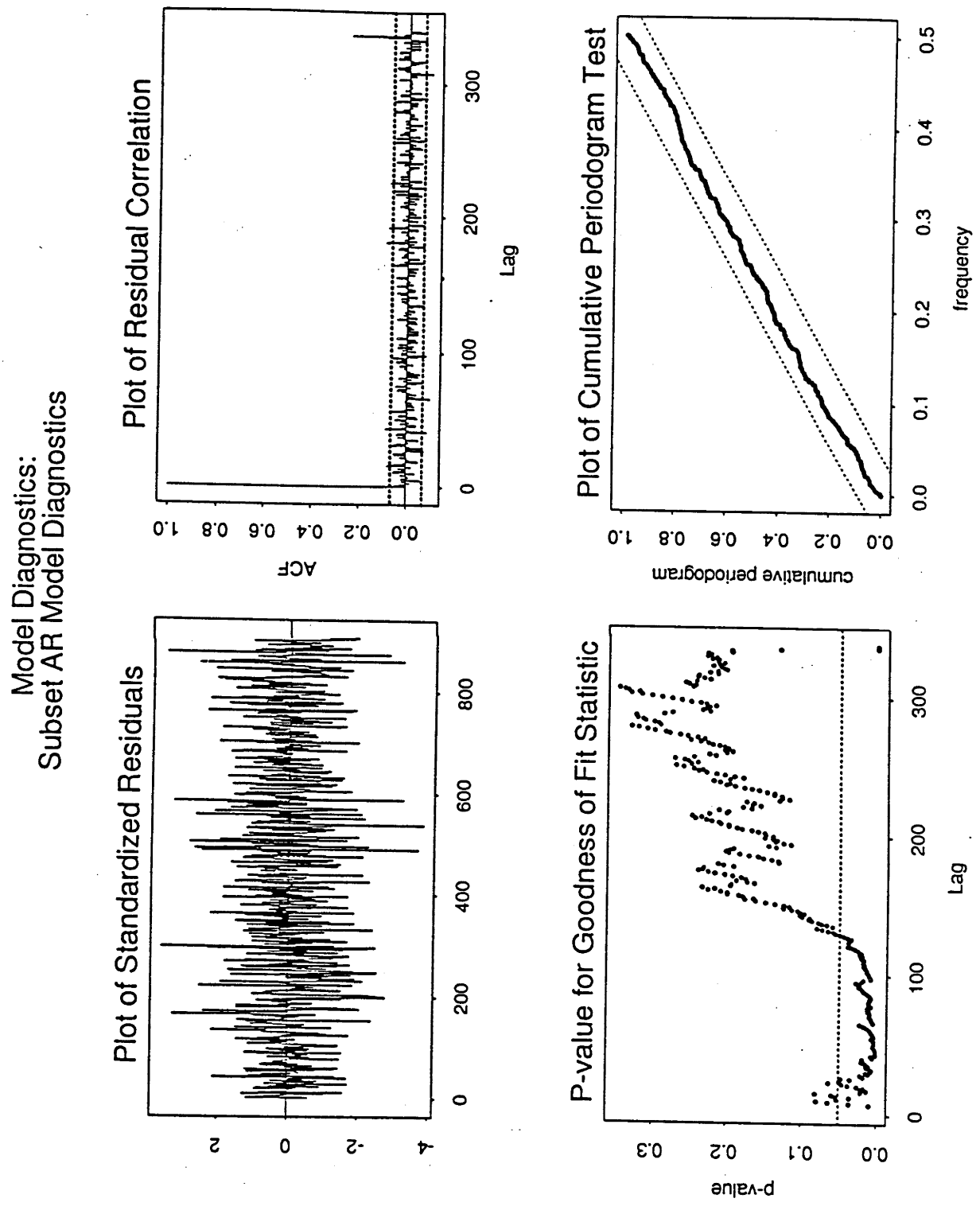


Figure B.14: Model 4: Model Fit Diagnostics for the Winter Data Set

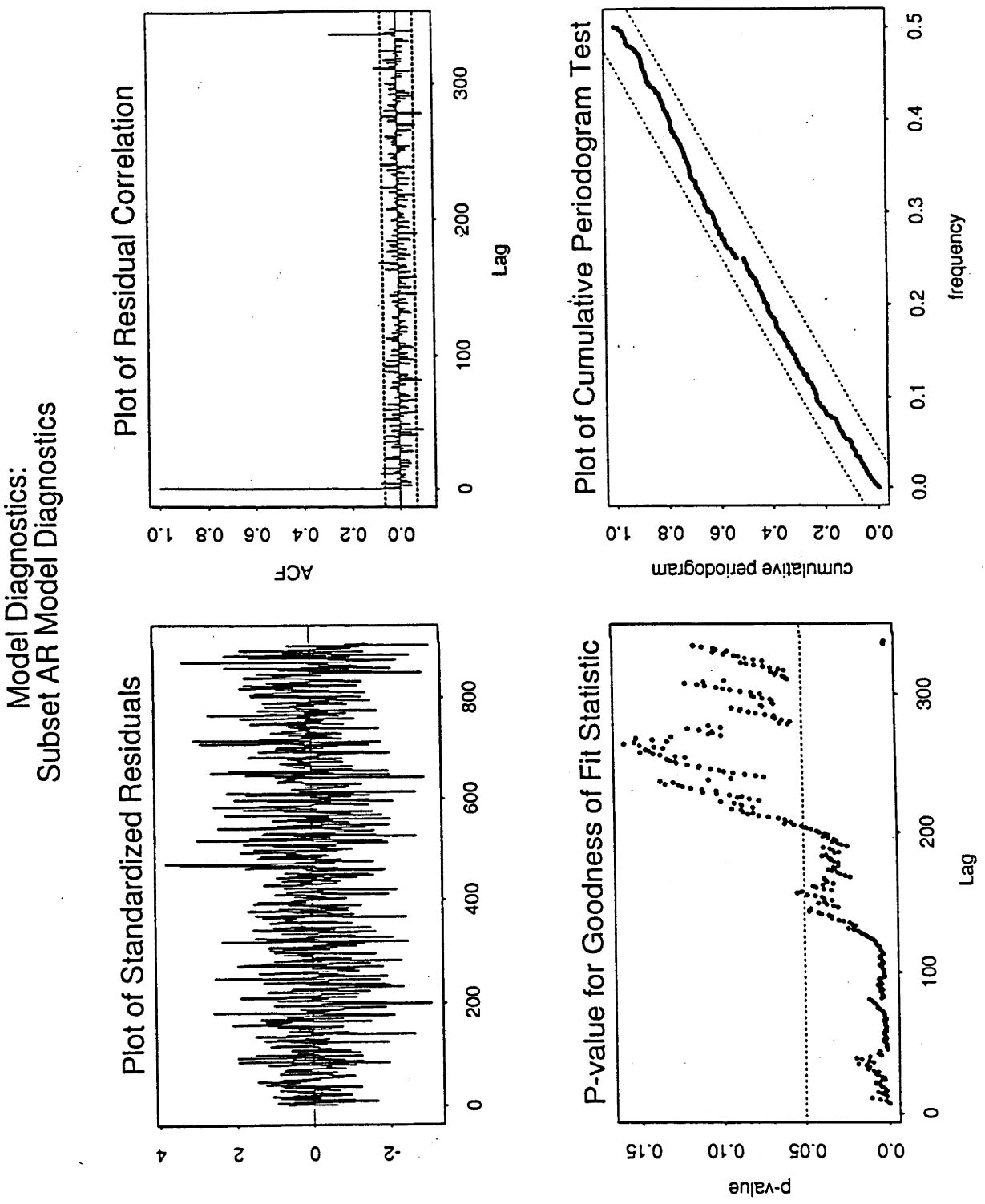


Figure B.15: Model 4: Model Fit Diagnostics for the Spring Data Set

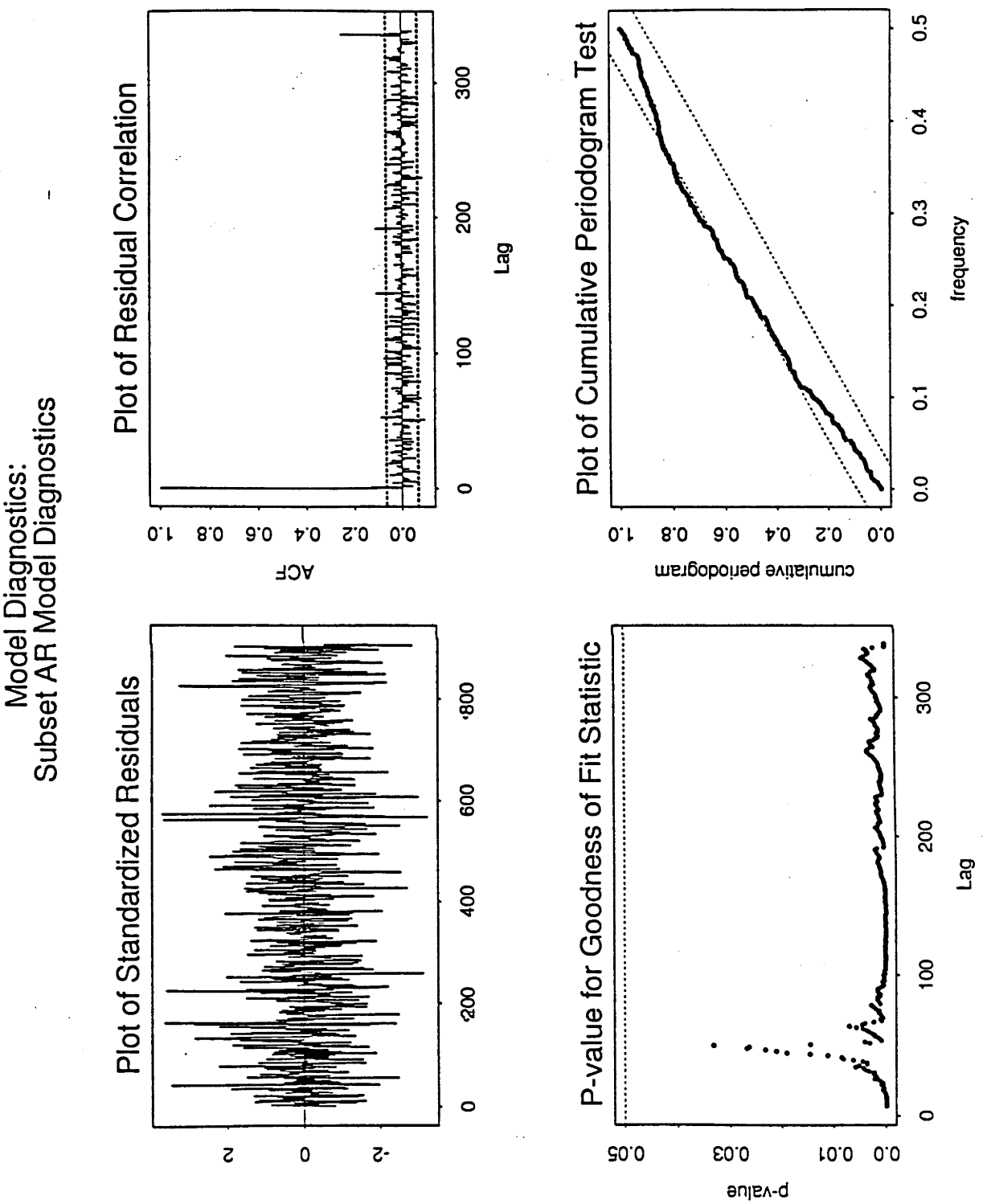


Figure B.16: Model 4: Model Fit Diagnostics for the Summer Data Set

B.5 Model 5

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13090	47.99965	-0.213E+00	-0.193E+00	-0.737E+00	0.768E+04
2	0.26189	23.99124	-0.382E-01	-0.173E+00	-0.135E+01	0.290E+04
3	0.65496	9.59330	-0.233E-01	-0.304E-01	-0.917E+00	0.136E+03
4	0.39317	15.98068	0.227E-01	0.310E-01	-0.939E+00	0.137E+03
5	0.52398	11.99137	-0.201E-01	0.271E-01	0.934E+00	0.106E+03
6	0.02871	218.85445	-0.107E-01	-0.437E-02	-0.386E+00	0.125E+02
7	0.07138	88.02461	-0.157E-02	-0.865E-02	-0.139E+01	0.717E+01
8	0.78579	7.99597	0.279E-02	-0.805E-02	0.124E+01	0.673E+01
9	0.22735	27.63614	-0.571E-02	-0.585E-02	-0.797E+00	0.620E+01
residual variance = 0.465984E-03 signal-to-noise ratio = 0.549363E+05						

Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13133	47.84218	-0.120E+00	-0.227E+00	-0.108E+01	0.475E+04
2	0.26200	23.98125	0.489E-02	-0.148E+00	0.154E+01	0.159E+04
3	0.39396	15.94892	-0.358E-01	0.510E-01	0.958E+00	0.280E+03
4	0.06394	98.27398	0.179E-01	0.336E-01	-0.108E+01	0.105E+03
5	0.65390	9.60877	-0.234E-02	-0.267E-01	-0.148E+01	0.521E+02
6	0.19880	31.60498	-0.114E-01	-0.433E-02	-0.364E+00	0.107E+02
7	0.45705	13.74712	0.113E-02	0.113E-01	-0.147E+01	0.933E+01
8	0.32468	19.35204	-0.370E-02	-0.989E-02	-0.121E+01	0.806E+01
residual variance = 0.398608E-03 signal-to-noise ratio = 0.340274E+05						

Table B.13: Seasonal Components for Weekdays & Weekend Days from the Autumn Data Set

Subset AR Model for the Stochastic Component sample variance = 0.66434490E-03			
lag	coeff.	s.e.	t-ratio
1	-1.0770758	0.52683531E-01	-20.444259
2	0.34413697	0.59371764E-01	5.7963070
4	-0.11760922	0.35764458E-01	-3.2884386
8	-0.92381091E-01	0.48928164E-01	-1.8880964
9	0.16349539	0.49333499E-01	3.3140847
12	-0.38875201E-01	0.30558075E-01	-1.2721744
43	-0.88763562E-01	0.48748310E-01	-1.8208542
44	0.13070448	0.49042352E-01	2.6651349
48	-0.35311464	0.51346747E-01	-6.8770596
49	0.40309174	0.75023946E-01	5.3728410
50	-0.11761845	0.54625101E-01	-2.1531942
size of 11, Schwarz criterion = -8522.9814 residual variance = 0.11405256E-03			

Table B.14: Subset AR Fit for the Stochastic Component from the Autumn Data Set

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13081	48.03111	-0.211E+00	-0.185E+00	-0.720E+00	0.754E+04
2	0.26195	23.98617	-0.199E-01	-0.199E+00	-0.147E+01	0.382E+04
3	0.65422	9.60402	-0.350E-01	-0.221E-01	-0.564E+00	0.163E+03
4	0.52356	12.00083	-0.280E-01	0.263E-01	0.755E+00	0.141E+03
5	0.39194	16.03109	0.282E-01	0.120E-01	-0.403E+00	0.895E+02
6	0.04031	155.86911	-0.129E-01	0.108E-01	0.700E+00	0.270E+02
7	0.08689	72.30933	-0.268E-02	-0.933E-02	-0.129E+01	0.897E+01
8	0.78614	7.99243	0.675E-02	-0.599E-02	0.725E+00	0.776E+01
residual variance = 0.453387E-03 signal-to-noise ratio = 0.589756E+05						

Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13064	48.09545	-0.154E+00	-0.244E+00	-0.101E+01	0.351E+04
2	0.26260	23.92657	0.297E-01	-0.125E+00	0.134E+01	0.703E+03
3	0.39339	15.97190	-0.307E-01	0.515E-01	0.103E+01	0.152E+03
4	0.06702	93.75058	0.247E-02	0.374E-01	-0.150E+01	0.595E+02
5	0.65344	9.61556	-0.480E-02	-0.231E-01	-0.137E+01	0.236E+02
6	0.45668	13.75843	0.311E-02	0.194E-01	-0.141E+01	0.163E+02
7	0.78550	7.99898	-0.904E-02	0.134E-01	0.976E+00	0.110E+02
8	0.20665	30.40467	0.354E-02	-0.151E-01	0.134E+01	0.101E+02
9	0.52622	11.94012	0.277E-02	-0.128E-01	0.136E+01	0.727E+01
10	0.32625	19.25894	-0.323E-02	-0.128E-01	-0.132E+01	0.736E+01
residual variance = 0.681142E-03 signal-to-noise ratio = 0.225080E+05						

Table B.15: Seasonal Components for Weekdays & Weekend Days from the Winter Data Set

Subset AR Model for the Stochastic Component sample variance = 0.53022244E-03			
lag	coeff.	s.e.	t-ratio
1	-1.1212477	0.50373564E-01	-22.258654
2	0.42874903	0.54687864E-01	7.8399301
4	-0.16030766	0.33761938E-01	-4.7481771
42	-0.48500348E-01	0.30528847E-01	-1.5886728
48	-0.37146518	0.51730294E-01	-7.1808055
49	0.35647923	0.50801664E-01	7.0170779
size of 6, Schwarz criterion = -8413.4173 residual variance = 0.13535962E-03			

Table B.16: Subset AR Fit for the Stochastic Component from the Winter Data Set

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13099	47.96778	-0.199E+00	-0.195E+00	-0.775E+00	0.725E+04
2	0.26183	23.99683	-0.480E-01	-0.191E+00	-0.132E+01	0.360E+04
3	0.52349	12.00250	-0.240E-01	0.350E-01	0.970E+00	0.168E+03
4	0.65422	9.60414	-0.280E-01	-0.158E-01	-0.513E+00	0.960E+02
5	0.39285	15.99374	0.229E-01	0.191E-01	-0.695E+00	0.828E+02
6	0.07537	83.35905	-0.981E-02	0.169E-02	0.171E+00	0.923E+01
7	0.78525	8.00155	0.530E-02	-0.682E-02	0.910E+00	0.696E+01
residual variance = 0.463305E-03 signal-to-noise ratio = 0.560678E+05						

Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13122	47.88428	-0.123E+00	-0.212E+00	-0.105E+01	0.341E+04
2	0.26199	23.98281	-0.447E-02	-0.153E+00	-0.154E+01	0.132E+04
3	0.39284	15.99416	-0.196E-01	0.422E-01	0.114E+01	0.122E+03
4	0.06353	98.90700	0.189E-01	0.393E-01	-0.112E+01	0.108E+03
5	0.65537	9.58728	-0.396E-02	-0.170E-01	-0.134E+01	0.173E+02
6	0.52049	12.07157	-0.224E-02	0.114E-01	0.138E+01	0.768E+01
7	0.45752	13.73314	0.425E-03	0.111E-01	-0.153E+01	0.699E+01
8	0.32414	19.38438	-0.777E-02	-0.899E-02	-0.859E+00	0.799E+01
residual variance = 0.509211E-03 signal-to-noise ratio = 0.250093E+05						

Table B.17: Seasonal Components for Weekdays & Weekend Days the Spring Data Set

Subset AR Model for the Stochastic Component Sample variance = 0.53022244E-03			
lag	coeff.	s.e.	t-ratio
1	-0.96963757	0.49756915E-01	-19.487494
2	0.20851450	0.53660504E-01	3.8858096
4	-0.11699036	0.36085650E-01	-3.2420191
9	0.10094524	0.30450812E-01	3.3150262
26	0.33976404E-01	0.27626083E-01	1.2298669
36	-0.84951985E-01	0.47668584E-01	-1.7821378
37	0.11422771	0.47771598E-01	2.3911218
48	-0.41098889	0.49466628E-01	-8.3084071
49	0.36467401	0.50472811E-01	7.2251576
size of 9, Schwarz criterion = -8644.3535 residual variance = 0.10238373E-03			

Table B.18: Subset AR Fit for the Stochastic Component from the Spring Data Set

Seasonal Component for Weekdays						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13171	47.70381	-0.156E+00	-0.177E+00	-0.847E+00	0.795E+04
2	0.26188	23.99294	0.104E-01	-0.129E+00	0.149E+01	0.240E+04
3	0.65492	9.59386	-0.212E-01	-0.262E-01	-0.891E+00	0.162E+03
4	0.39320	15.97980	0.292E-01	0.156E-01	-0.491E+00	0.156E+03
5	0.52466	11.97571	-0.183E-01	-0.900E-02	-0.457E+00	0.592E+02
6	0.04088	153.70238	-0.983E-02	-0.571E-02	-0.526E+00	0.184E+02
7	0.12079	52.01809	0.355E-02	0.192E-01	-0.139E+01	0.544E+02
residual variance = 0.303623E-03 signal-to-noise ratio = 0.539872E+05						

Seasonal Component for Weekend days						
i	freq. ω_i	period	A_i	B_i	phase	$\chi^2(2)$
1	0.13132	47.84805	-0.102E+00	-0.145E+00	-0.960E+00	0.280E+04
2	0.26171	24.00781	0.342E-01	-0.107E+00	0.126E+01	0.112E+04
3	0.06566	95.69768	0.125E-01	0.388E-01	-0.126E+01	0.149E+03
4	0.39229	16.01684	-0.675E-02	0.259E-01	0.132E+01	0.643E+02
5	0.65644	9.57167	-0.848E-02	-0.170E-01	-0.111E+01	0.324E+02
6	0.52328	12.00731	-0.160E-02	-0.153E-01	-0.147E+01	0.211E+02
7	0.78542	7.99982	-0.103E-01	0.732E-02	0.616E+00	0.144E+02
residual variance = 0.308913E-03 signal-to-noise ratio = 0.210352E+05						

Table B.19: Seasonal Components for Weekdays & Weekend Days from the Summer Data Set

Subset AR Model for the Stochastic Component sample variance = 0.40506964E-03			
lag	coeff.	s.e.	t-ratio
1	-1.0619051	0.52993717E-01	-20.038321
2	0.31123243	0.59746973E-01	5.2091748
4	-0.81574859E-01	0.37024595E-01	-2.2032614
10	0.49479057E-01	0.32583830E-01	1.5185157
17	0.75424073E-01	0.52012499E-01	1.4501144
18	-0.11800351	0.50910581E-01	-2.3178583
22	0.76201213E-01	0.30853458E-01	2.4697787
37	0.45674269E-01	0.32497490E-01	1.4054707
40	-0.66634203E-01	0.33799867E-01	-1.9714339
45	0.65440717E-01	0.33110285E-01	1.9764468
48	-0.28003313	0.55178349E-01	-5.0750545
49	0.25399478	0.53321285E-01	4.7634782
size of 12, Schwarz criterion = -8822.9621 residual variance = 0.85406862E-04			

Table B.20: Subset AR Fit for the Stochastic Component from the Summer Data Set

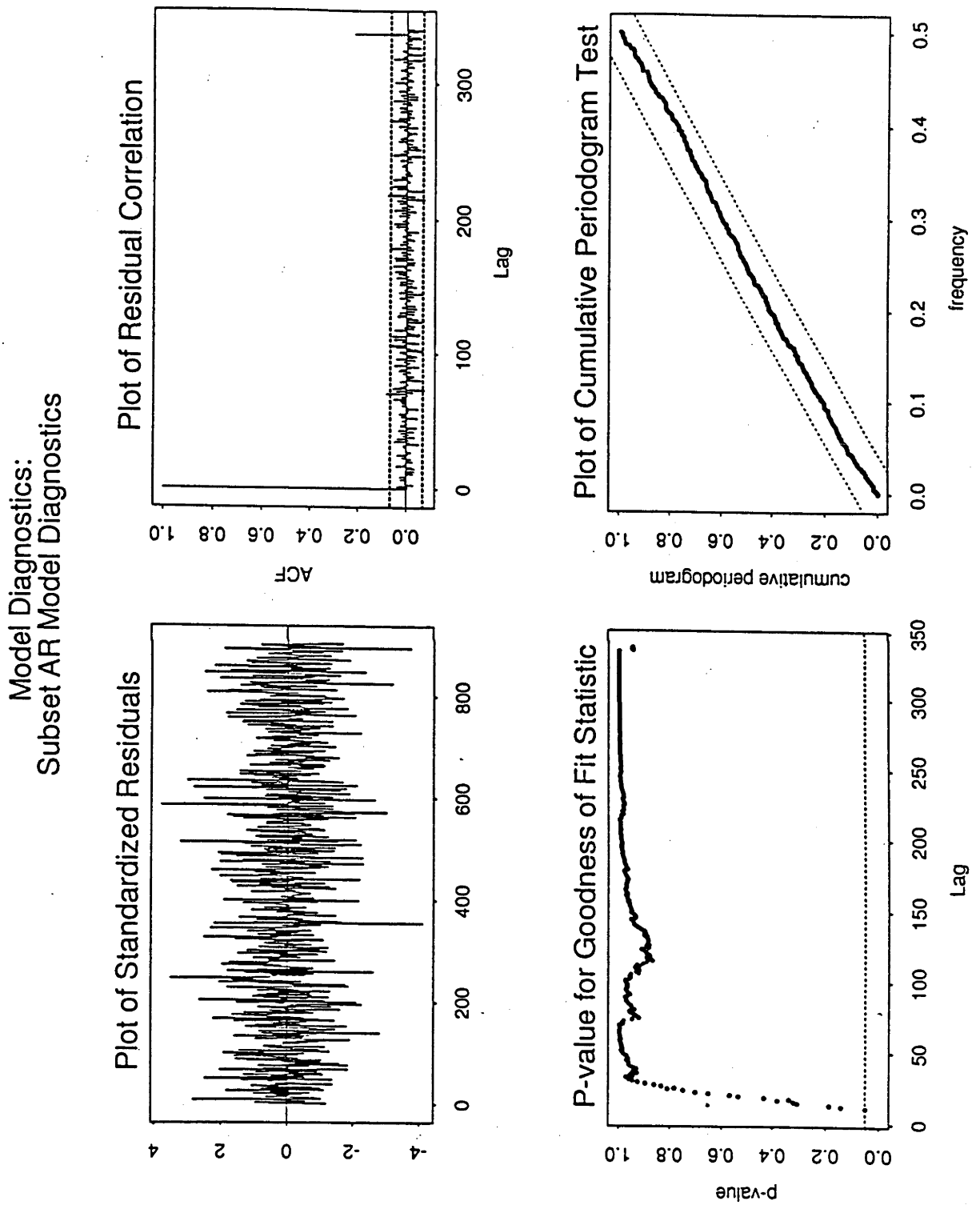


Figure B.17: Model 5: Model Fit Diagnostics for the Autumn Data Set

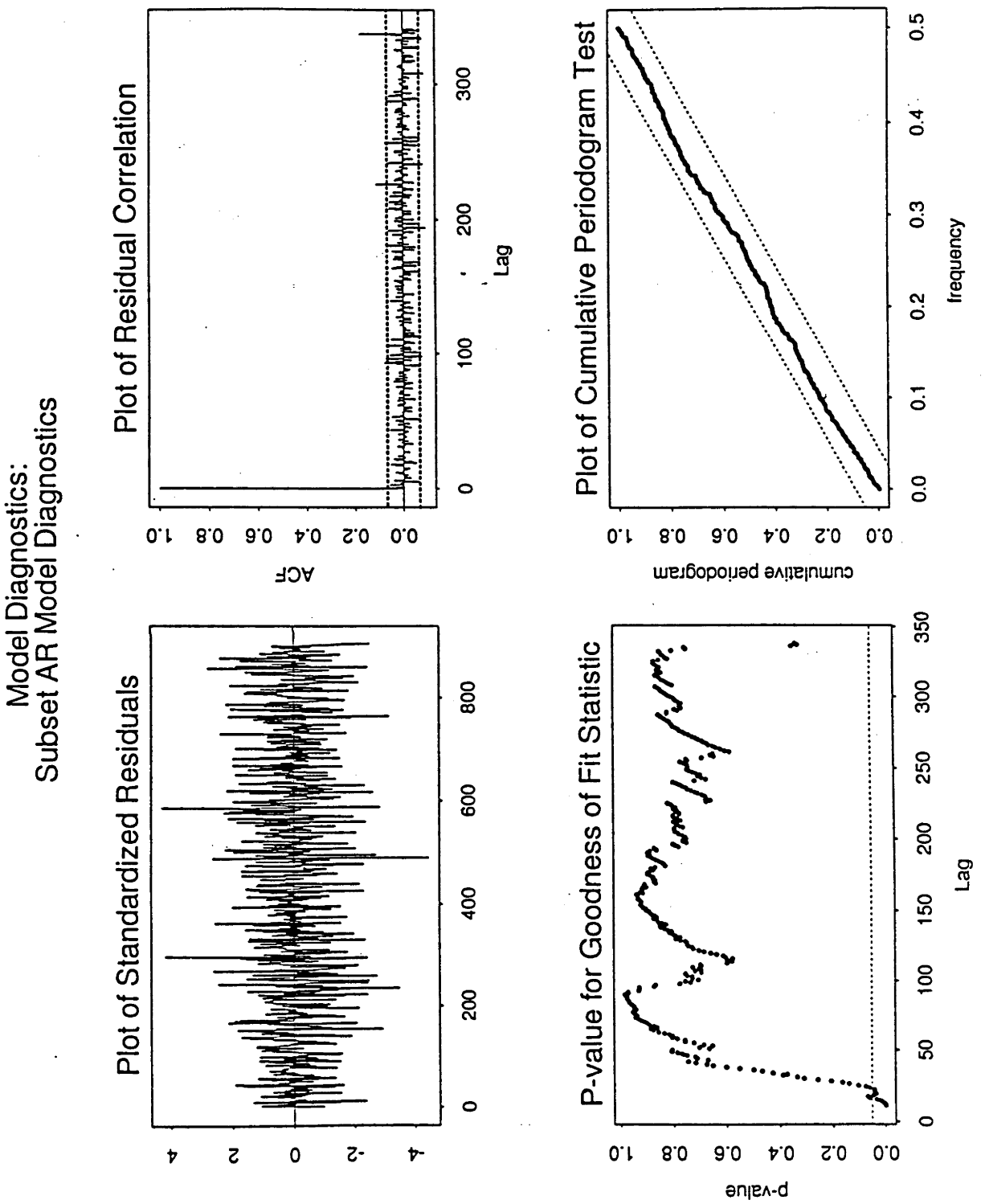


Figure B.18: Model 5: Model Fit Diagnostics for the Winter Data Set

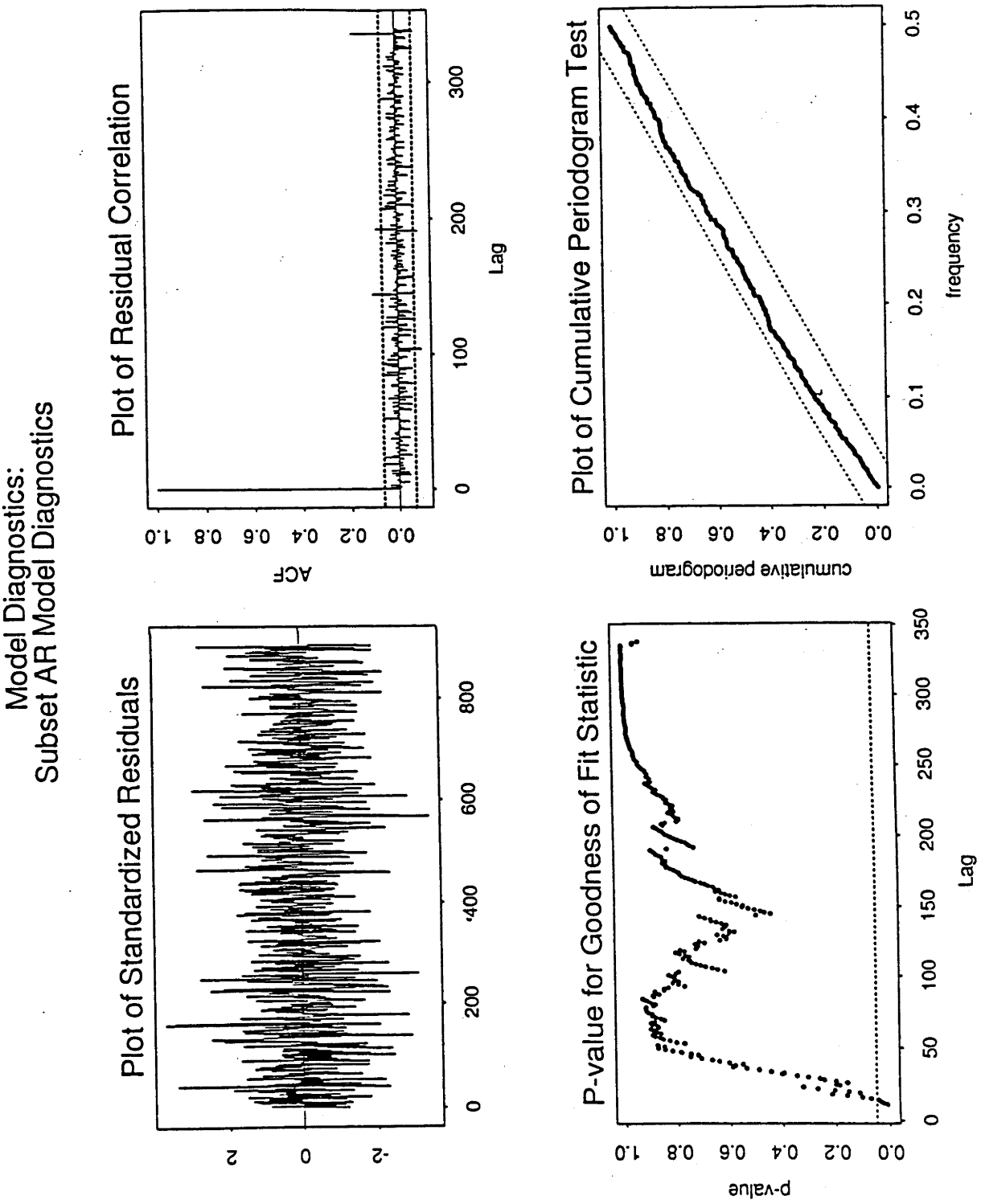


Figure B.19: Model 5: Model Fit Diagnostics for the Spring Data Set

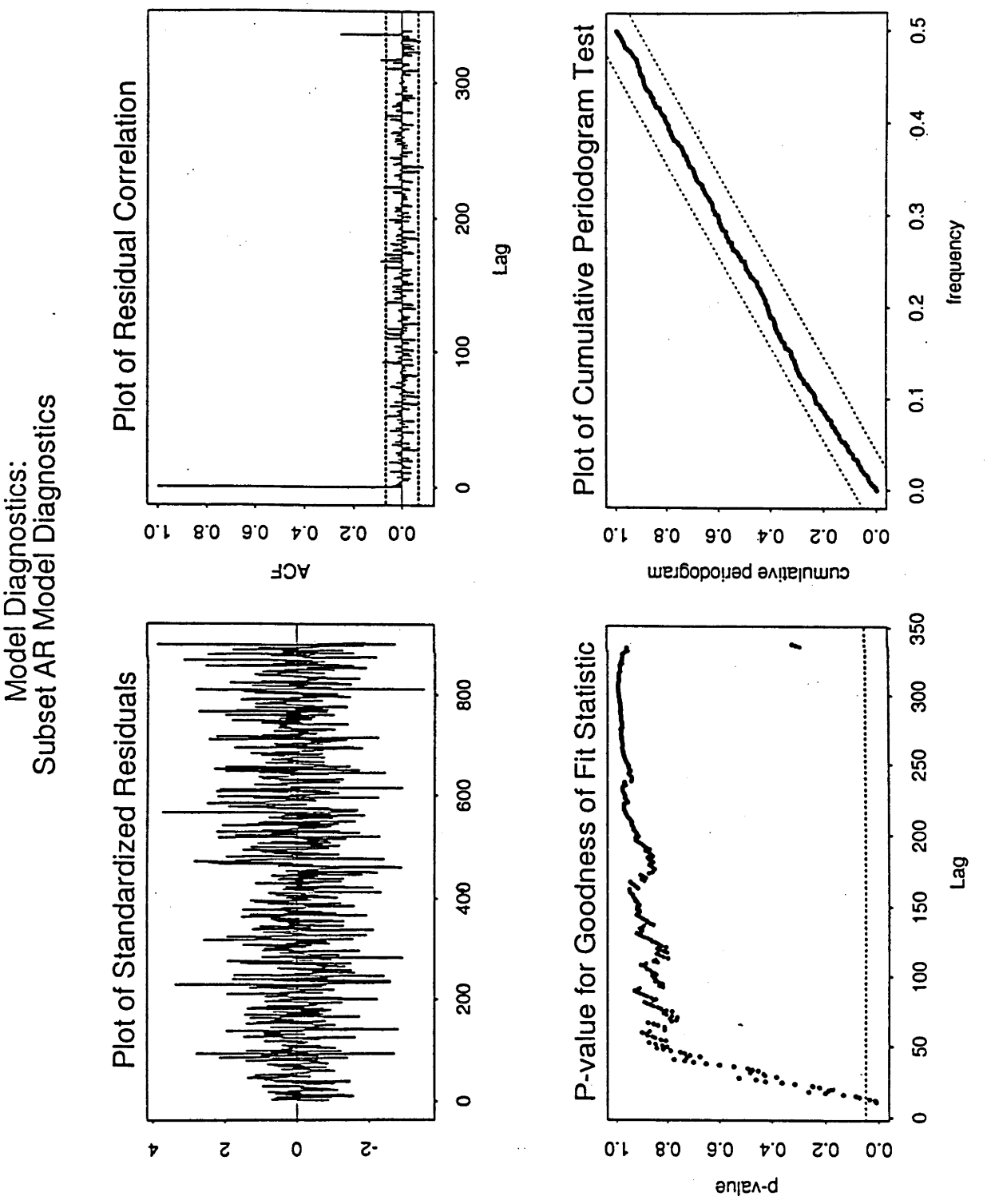


Figure B.20: Model 5: Model Fit Diagnostics for the Summer Data Set

Appendix C

Model Accuracy, Stability and Selection

Two main properties characterize the usefulness of a proposed model, f, in data analysis. The first is the model's fit to empirical data and is referred to as model accuracy. The second, in the models which contain unknown parameters to be estimated, is the dependence of the fitted model, or the estimated parameters on the particular observation data set. The second property is called model stability. The stability of a model may refer to either the variability of the estimated model function or, if the parameters are of primary interest, the variances of the estimated parameters.

C.1 Model Accuracy

We first consider the statistics for measuring the goodness of fit of a proposed model for a particular data set. The Residual Sum Square (RSS)

$$RSS(\hat{f}) = \sum_{t=1}^{n}(y(x_t) - \hat{f})^2$$

is a statistic to measure the goodness of fit of the estimated model $\hat{f}$, where n is the size of data set $\{y(x_t)\}$. Another statistic is R-Square(R^2) defined by

$$R^2(\hat{f}) = 1 - RSS(\hat{f})/RSS(\hat{y})$$

where $RSS(\bar{y}) = \sum_{t=1}^{n}(y(x_t) - \bar{y})^2$.

Here $R^2(\hat{f})$ takes values between zero and one. The former indicates no explanation of $y(x_t)$ and the latter implies that the responses lie exactly on the fitted model function.

For the different model functions, the model parameters and corresponding RSS and R^2 are estimated in Table C.1, C.2, C.3, C.4.

Model f_1 : $y = A + C\,e^{-(x-65)^2/D}$					
Time(t)\Coef. (variance)	A	C	E	RSS	R^2
3AM	12.1133	-0.5682	743.8885	0.124	0.981
	(0.0001)	(0.0001)	(899.5280)		
6AM	12.3343	-0.6838	729.0518	0.201	0.979
	(0.0002)	(0.0001)	(818.1399)		
9AM	13.0273	-0.6348	280.3147	0.098	0.992
	(0.0000)	(0.0000)	(24.6231)		
12AM	12.8449	-0.5021	191.9326	0.241	0.954
	(0.0002)	(0.0001)	(97.1591)		
3PM	12.8697	-0.5778	270.6270	0.320	0.915
	(0.0022)	(0.0021)	(1164.4569)		
6PM	13.2772	-0.9590	308.8132	3.618	0.802
	(0.0071)	(0.0064)	(2349.9190)		
9PM	12.9624	-0.7520	260.4629	0.361	0.980
	(0.0001)	(0.000)	(65.215)		
12PM	12.3693	-0.5709	463.9415	0.124	0.985
	(0.0001)	(0.0001)	(198.0147)		

Table C.1: Gauss-Newton Nonlinear Least-Squares for Model f_1

From the goodness of fit point of view, RRS and R^2 can serve as discrimination functions, if we would like to choose an "optimal" model from the several proposed models. However, objections to R^2 as a discrimination function have been raised by Draper (1984), Healy (1984) and Hellend (1987) although R^2 genuinely represents the proportion of variation explained by a model. They argued that R^2 does not have any obvious meaning for a nonlinear regression model. It may mislead if R^2 is used as a discrimination function for nonlinear regression models. Therefore, we should ask how to decide whether a nonlinear regression model provides a good fit to a data set. Once having decided that there is no evidence of lack of independence, invariance,

Model f_2 : $y = A + Bx + Ce^{-(x-65)^2/D}$						
Time(t)\Coef. (variance)	A	B	C	E	RSS	R^2
3AM	11.9926 (0.0005)	0.0047 (0.0000)	-0.7382 (0.0008)	711.0174 (396.5578)	0.107	0.984
6AM	12.1689 (0.0009)	0.0073 (0.0000)	-0.9656 (0.0023)	731.9504 (366.0784)	0.176	0.982
9AM	13.1355 (0.0002)	-0.0033 (0.0000)	-0.5306 (0.0002)	262.4165 (28.7918)	0.077	0.994
12AM	13.0499 (0.0000)	-0.0056 (0.0000)	-0.3393 (0.0000)	142.7686 (2.9388)	0.007	0.997
3PM	12.9803 (0.000)	-0.0047 (0.0000)	-0.3840 (0.0000)	173.3181 (31.4440)	0.026	0.993
6PM	13.8276 (0.0000)	-0.0176 (0.0000)	-0.3509 (0.0000)	133.0054 (9.2097)	0.035	0.998
9PM	13.2261 (0.0002)	-0.0075 (0.0000)	-0.5253 (0.0002)	233.8078 (41.3625)	0.143	0.992
12PM	12.3090 (0.000)	0.0020 (0.0000)	-0.6386 (0.0004)	463.6913 (150.3425)	0.118	0.986

Table C.2: Gauss-Newton Nonlinear Least-Squares for Model f_2

and normality of the stochastic term (see equation (6.3)) for the data set/model combination in question, we can only look at the magnitude of the residual variance (equivalently RSS) and decide whether it is significant small.

For a nested model function set, $\Im$, in which $f_1 \subset f_2 \subset \cdots \subset f_k$, because of the nesting, the parameter sizes satisfy $m(1) < m(2) < \cdots < m(k)$ and $RSS(f_1) \geq RSS(f_2) \geq \cdots \geq RSS(f_k)$. In other words, as more complexity is added to include more parameters, the fit will be automatically improved. However, the improved fit may not be significant and obtained at the cost of more parameters to be estimated in the model function. A general procedure for testing nested models in the arbitrary model function setting is constructed as below.

We compare two models using the likelihood ratio test procedure. For $f_i \subset f_j$, we test

$$\mathbf{H_0}: \quad \mathbf{E}(y(x_t)) = f_i(t, \Theta_i) \tag{C.1}$$

Model f_3: $y = A + C\,e^{-(x-D)^2/E}$						
Time(t)\Coef. (variance)	A	C	D	E	RSS	R^2
3AM	12.0599	-0.5056	62.2667	492.2006	0.101	0.984
	(0.0001)	(0.0001)	(0.0674)	(594.9642)		
6AM	12.2683	-0.5970	61.3773	443.9121	0.161	0.983
	(0.0001)	(0.0001)	(0.0932)	(529.7476)		
9AM	13.0346	-0.6460	65.7653	312.8798	0.086	0.993
	(0.0000)	(0.0000)	(0.0178)	(62.4531)		
12AM	12.8505	-0.5082	66.8354	238.2069	0.027	0.995
	(0.0000)	(0.0000)	(0.0024)	(21.5077)		
3PM	12.8039	-0.5165	66.8485	252.3384	0.008	0.998
	(0.0000)	(0.0000)	(0.0005)	(16.6387)		
6PM	13.2086	-0.9299	70.4563	406.8597	0.053	0.997
	(0.0001)	(0.0001)	(0.0049)	(74.0542)		
9PM	12.9861	-0.7820	66.9658	342.2392	0.166	0.991
	(0.0001)	(0.0001)	(0.0178)	(109.1855)		
12PM	12.3517	-0.5507	63.9164	391.1739	0.113	0.987
	(0.0000)	(0.0000)	(0.0324)	(231.7551)		

Table C.3: Gauss-Newton Nonlinear Least-Squares for Model f_3

against

$$\mathbf{H_1}: \quad \mathbf{E}(y(x_t)) = f_j(t, \Theta_j) \tag{C.2}$$

Under the correct model $f_r(x_t, \Theta_r)$ with normality of the error we have

$$Y \sim \mathbf{N}(f_r(\Theta_r), \sigma^2 I) \quad \text{for } r = i \text{ or } j$$

For the likelihood ratio test we find $\sup_{\Theta_r, \sigma^2}(Y)$ or

$$\sup_{\Theta_r, \sigma^2} (2\pi\sigma^2)^{-n/2} \exp\{-\sum[y(x_t) - f_r(x_t, \Theta_r)]^2/(2\sigma^2)\} \tag{C.3}$$

Derivatives with respect to Θ_r and σ^2 yield the maximum likelihood estimator Θ_r and $\hat{\sigma}^2_{f_r} = \sum[y(x_t) - f_r(x_t, \Theta_r)]^2/n$ for $r = i$ or j. Under the specific models of hypothesis (C.1), (C.2) becomes, upon substitution of estimates $\hat{\Theta}_r$, $(2\hat{\sigma}^2_{f_r})^{-n/2} \exp(-n/2)$ where under $\mathbf{H_0}: r = i$ and under $\mathbf{H_1}: r = j$. The likelihood ratio statistic is $\lambda = \lambda_i/\lambda_j$, which simplifies to

$$\lambda = [RSS(\hat{f}_j)/RSS(\hat{f}_i)]^{n/2} \tag{C.4}$$

Model f_4: $y = A + B\,x + C\,e^{-(x-D)^2/E}$							
Time(t)\Coef. (variance)	A	B	C	D	E	RSS	R^2
3AM	12.3467	-0.0117	-0.1232	54.6684	43.0952	0.018	0.997
	(0.0000)	(0.0000)	(0.0000)	(0.0047)	(2.9341)		
6AM	12.6316	-0.0144	-0.1453	54.0883	38.8437	0.029	0.997
	(0.0000)	(0.0000)	(0.0000)	(0.0054)	(3.3465)		
9AM	13.4503	-0.0127	-0.2522	61.4037	93.9856	0.046	0.996
	(0.0002)	(0.0000)	(0.0001)	(0.0175)	(31.8116)		
12AM	13.091	-0.0067	-0.3130	64.5802	126.1005	0.006	0.999
	(0.0000)	(0.0000)	(0.0000)	(0.0026)	(5.3749)		
3PM	12.8239	-0.0007	-0.4928	66.6211	238.1101	0.008	0.998
	(0.0001)	(0.0000)	(0.0001)	(0.0098)	(48.9132)		
6PM	13.6913	-0.0143	-0.4343	66.3372	178.1201	0.017	0.999
	(0.0001)	(0.0000)	(0.0001)	(0.0105)	(25.9664)		
9PM	13.5612	-0.0168	-0.275	61.5364	93.3020	0.092	0.995
	(0.0002)	(0.0000)	(0.0001)	(0.0181)	(28.4945)		
12PM	12.7170	-0.0124	-0.1661	57.3215	66.9038	0.036	0.996
	(0.0000)	(0.0000)	(0.0000)	(0.0093)	(9.9861)		

Table C.4: Gauss-Newton Nonlinear Least-Squares for Model f_4

We reject $\mathbf{H_0}$ if λ is small. Under general regularity conditions, $-2\ln(\lambda) \sim \chi^2$ with degrees of freedom $m(j) - m(i)$. Hence, an α significance level critical region is $\tilde{\lambda}_{i:j} = n\ln[RSS(\hat{f}_i)/RSS(\hat{f}_j)] > \chi^2_{(1-\alpha)}(m(j) - m(i))$.

The $-2ln(\lambda)$ values for the different pairs of model functions at different times of a day are shown in Table C.5.

Noting $\chi^2_{(.99)}(1) = 6.63$, we reject $\mathbf{H_0}$, given by (C.1) with sequentially paired models for a set time of day in the two nested models at significance level $\alpha = 0.01$. Significant gains in fit to the response, y, are obtained by the complexity added in going from f_1 to f_2 then to f_4 or from f_1 to f_3 then to f_4. Therefore, we conclude that the model function f_4 is the optimal model among the candidate models in the sense of statistically best fitting. The stability of the model is taken into account in next section.

	3AM	6AM	9AM	12AM	3PM	6PM	9PM	12PM
				f_1 against f_2				
$\tilde{\lambda}_{1:2}$	38.83	34.69	61.73	83.89	613.90	1158.12	251.92	13.88
				f_2 against f_4				
$\tilde{\lambda}_{2:4}$	1063.14	461.39	130.22	51.10	276.65	176.14	120.07	310.88
				f_1 against f_3				
$\tilde{\lambda}_{1:3}$	53.58	57.22	34.33	527.63	887.60	1056.28	210.84	25.50
				f_3 against f_4				
$\tilde{\lambda}_{3:4}$	456.38	438.87	157.62	362.33	2.9545	277.98	161.16	299.26

Table C.5: Likelihood Ratio Statistics

C.2 Model Stability

Based on the asymptotic theory resulting from LS estimation, a discrimination function, which measures a model function's stability at each location can be employed as statistical evidence which helps separate effective from non-effective models. The effective model, we refer to here, should fit the responses to a desired accuracy while possessing a degree of stability over the range of applications. Therefore, a stable model with a moderate RSS may be preferable to an unstable model with a smaller RSS. The stability function SF of an estimated model $\hat{f}$ at location x is defined as

$$SF(\hat{f}(\Theta, x_t)) = \hat{\alpha}(x_t) + \hat{\sigma}^2\hat{\delta}(x_t) \tag{C.5}$$

where

$$\hat{\alpha}(x_t) = \hat{F}'(\Theta, x_t)\hat{M}\hat{F}(\Theta, x_t)$$

and

$$\hat{\delta}(x_t) = Tr[\hat{M}\hat{H}(x_t)\hat{M}\hat{H}(x_t)]/2 + \{Tr[\hat{M}\hat{H}(x_t)]\}^2/4$$

$$\hat{F}(\Theta, x_t) = (\hat{f}_1(\Theta, x_t), \hat{f}_2(\Theta, x_t), \cdots, \hat{f}_m(\Theta, x_t))'$$

$$\hat{f}_i(\Theta, x_t) = \frac{\partial f(\Theta, x_t)}{\partial \Theta_i}, \quad i = 1, 2, \cdots, m$$

$$\hat{H}(\Theta, x_t) = \begin{pmatrix} \hat{f}_{1,1}(\Theta, x_t) & \hat{f}_{1,2}(\Theta, x_t) & \cdots & \hat{f}_{1,m}(\Theta, x_t) \\ \hat{f}_{2,1}(\Theta, x_t) & \hat{f}_{2,2}(\Theta, x_t) & \cdots & \hat{f}_{2,m}(\Theta, x_t) \\ \cdots & \cdots & \cdots & \cdots \\ \hat{f}_{m,1}(\Theta, x_t) & \hat{f}_{m,2}(\Theta, x_t) & \cdots & \hat{f}_{m,m}(\Theta, x_t) \end{pmatrix}$$

$$\hat{f}_{i,j}(\Theta, x_t) = \frac{\partial^2 f(\Theta, x_t)}{\partial\Theta_i \partial\Theta_j}, \quad i, j = 1, 2, \cdots, m$$

$$\hat{M} = (\hat{F}(\Theta)\hat{F}'(\Theta))^{-1}$$

$$\hat{F}'(\Theta) = (\hat{F}(\Theta, x_1), \cdots, \hat{F}(\Theta, x_t), \cdots)$$

$\hat{\sigma}^2$ is a consistent estimation of the residual variance of the true model.

The stability function $SF(\hat{f}(\Theta, x_t))$ is the sum of two parts. The first term, $\alpha(x_t)$ is a stability measure for the linear part of the model function, while the second term, $\hat{\sigma}^2\delta(x_t)$ is a stability measure for the non-linear component of the model function. It can be proved that, ignoring a small bias, an asymptotic confidence interval for the assumed true model value $f(x_t)$ is

$$\hat{f}(x_t) \pm [SF(\hat{f}(x_t)\sigma^2]^{1/2} z_{(1-\alpha/2)} \tag{C.6}$$

where z is the standard normal distribution and α is the significance level. The overall stability can be measured by the sum of the values of the stability function at each location. It is noted that $\sum \alpha(x_t) = Tr[F'MF] = m$ (number of parameters to be estimated in the model). The overall stability of the model f is equal to the number of parameters in f plus $\sum_{t=1}^{n} \delta(x_t)$ weighted by $\hat{\sigma}^2$. $\hat{\sigma}^2$, however, is unknown unless a desired accuracy is given, since the true model is unknown. In table C.6, $\hat{\Delta}_i = \sum_{t-1}^{n} \hat{\delta}_i(x_t)$ is listed for all candidate models at different times of a day.

It is obvious from the Table C.6 that the more parameters in a model, the more instability there is in the model. Comparing the RSS of the four models f_i $i = 1, 2, 3, 4$) in Table C.2 and C.3, and $\hat{\Delta}_i$ in Table C.6, it is obvious that the model f_1 is the most inaccurate but the most stable model; the model f_4 is the most accurate but the most unstable model. The model f_2 is more accurate and stable than model f_3 at 12AM, 6PM, 9PM and f_3 is more accurate and stable than model f_2 at 3AM, 6AM,

$\hat{\Delta}_i$\ Time	3AM	6AM	9AM	12AM	3PM	6PM	9PM	12PM
$\hat{\Delta}_1$	16.44	10.36	2.32	8.43	60.83	7.56	2.22	7.36
$\hat{\Delta}_2$	85.68	60.40	10.52	15.17	44.94	12.45	7.19	24.02
$\hat{\Delta}_3$	51.38	33.61	10.42	16.05	39.80	15.29	9.21	22.74
$\hat{\Delta}_4$	85.67	71.00	104.13	88.16	312.98	107.06	49.61	67.72

Table C.6: $\hat{\Delta}$ for the Different Models

9AM, 3PM, 12PM. This indicate that the parameter B and E do play significant roles. However, we are not sure if it is worth the cost of the inaccuracy of model f_1 to achieve stability or the cost of the instability of model f_4 to achieve accuracy. The model criterion CF can trade-off these two contradict measures for a model function. See details in section 6.3.3 of chapter 6 and the following section.

C.3 Model Criterion for the Model Family

Table C.5 shows that the model f_4 achieves significant gains in the sense of accuracy over models f_1, f_2 and f_3. If assuming the model f_4 is the true model, and using $\hat{\sigma}^2$ from f_4 as a consistent estimate of σ^2, we obtain the model criterion function, CF, (see section 6.3.3 of chapter 6) values for the all candidate models at different time of a day in Table C.7.

On the other hand, $\hat{\sigma}^2$ from f_4 should not replace σ^2 in the CF function (6.14) if there is not enough evidence to establish that f_4 is the true model. The average of $\hat{\sigma}^2$ for all candidate models at different time or the overall average of $\hat{\sigma}^2$ can be used as an estimate of σ^2 . Table C.8 and Table C.9 list the criterion values when the σ^2 are estimated by the average of $\hat{\sigma}^2$ for all candidate models at different times and the overall average of $\hat{\sigma}^2$, respectively.

It is noted that the criterion function values of the two nested systems $f_1 \subset f_2 \subset f_4$ and $f_1 \subset f_3 \subset f_4$ are decreasing for each time column except $CF(\hat{f}_3) < CF(\hat{f}_4)$ at 3PM in Table C.7, Table C.8 and Table C.9 which will be explained later. This indicates that the model f_4 is overwhelmingly supported as the true model by the

$CF(\hat{f}_i)\backslash$ Time	3AM	6AM	9AM	12AM
$\hat{\sigma}^2$	6.8302E-05	1.1395E-4	1.8086E-4	2.4583E-5
$CF(\hat{f}_1)$	1.0611E-1	1.7204E-1	5.2242E-2	2.3477E-1
$CF(\hat{f}_2)$	8.9274E-2	1.4686E-1	3.1424E-2	1.49836E-3
$CF(\hat{f}_3)$	8.3474E-2	1.3216E-1	4.0124E-2	2.090E-2
$CF(\hat{f}_4)$	3.4191E-4	5.7069E-4	9.0770E-4	1.2297E-4
$CF(\hat{f}_i)\backslash$ Time	3PM	6PM	9PM	12PM
$\hat{\sigma}^2$	3.3745E-05	6.9600E-5	3.3787E-4	1.3702E-4
$CF(\hat{f}_1)$	3.1210E-1	3.6003	2.6991E-1	8.8511E-2
$CF(\hat{f}_2)$	1.7535E-2	1.8079E-2	5.2354-2	8.2249E-2
$CF(\hat{f}_3)$	2.3511E-4	3.5779E-2	7.5653E-2	7.7149E-2
$CF(\hat{f}_4)$	1.6908E-4	3.4852E-4	1.6950E-3	6.8639E-4

Table C.7: The *CF* Values When σ^2 Is Estimated from Model f_4

$CF(\hat{f}_i)\backslash$ Time	3AM	6AM	9AM	12AM
$\hat{\sigma}^2$	3.3066-e4	5.4981E-4	2.9980E-4	2.9219E-4
$CF(\hat{f}_1)$	3.7372E-2	6.0909E-2	2.2150E-2	1.7135E-1
$CF(\hat{f}_2)$	2.0826E-2	3.6204E-2	1.4521E-3	-6.1652E-2
$CF(\hat{f}_3)$	1.5014E-2	2.1480E-2	1.0152E-2	-4.2252E-2
$CF(\hat{f}_4)$	-6.7844E-2	-1.0964	-2.8923E-2	-6.274148E-2
$CF(\hat{f}_i)\backslash$ Time	3PM	6PM	9PM	12PM
$\hat{\sigma}^2$	3.7274E-4	3.723E-3	7.0018E-4	3.7214E-4
$CF(\hat{f}_1)$	2.3077E-1	2.6982	1.7245	2.7619E-2
$CF(\hat{f}_2)$	-6.3465E-2	-8.8014E-1	5.2353E-2	2.1598E-2
$CF(\hat{f}_3)$	-8.0767E-2	-8.623221E-1	-2.1436E-2	1.6498E-2
$CF(\hat{f}_4)$	-8.0381E-2	-8.9028E-1	-9.4976E-2	-5.9711E-2

Table C.8: The *CF* Values for When σ^2 Is Estimated from the Average of All Models at Different Times

$CF(\hat{f}_i)$\ Time	3AM	6AM	9AM	12AM
$\hat{\sigma}^2$ = 4.0E-4 (overall mean)				
$CF(\hat{f}_1)$	1.9208E-2	9.9105	-3.1989-3	1.4580E-1
$CF(\hat{f}_2)$	2.7411E-3	7.4229E-2	-2.3795E-2	-8.7093E-2
$CF(\hat{f}_3)$	-3.0753E-2	5.9516E-2	-1.5095E-2	-6.7692E-2
$CF(\hat{f}_4)$	-8.5859E-2	-7.1766E-2	-5.4050E-2	-8.8058
$CF(\hat{f}_i)$\ Time	3PM	6PM	9PM	12PM
$CF(\hat{f}_1)$	2.2423E-2	3.5187	2.5320E-1	2.0404E-2
$CF(\hat{f}_2)$	-6.9978	-6.3194	5.2354E-2	3.0438E-2
$CF(\hat{f}_3)$	-8.7281E-2	-4.5493E-2	5.9004E-2	9.3109E-3
$CF(\hat{f}_4)$	-8.6850E-2	-8.0549E-2	-1.4876E-2	-6.6868E-2

Table C.9: The *CF* Values When σ^2 Is Estimated by the Average over All Models

criterion function. The consistent negative values of the criterion function of model f_4 in Table C.8 and Table C.9 imply that the average $\hat{\sigma}^2$ over all candidate models at different times or the overall average of $\hat{\sigma}^2$ are both over estimates of σ^2, and then, indicate that the $\hat{\sigma}^2$ from model f_4 is most likely to be a consistent estimate of σ^2. If we assume the σ^2 is time invariant, the average of $\hat{\sigma}^2$ over different times is around 1.0E-4 and is used to calculate the criterion function values for the model candidates in Table C.10.

$CF(\hat{f}_i)$\ Time	3AM	6AM	9AM	12AM
$\hat{\sigma}^2$ = 1.0E-4 (overall mean)				
$CF(\hat{f}_1)$	9.7800E-2	1.7560E-1	7.2700E-2	2.1690E-2
$CF(\hat{f}_2)$	8.1003E-2	1.5040E-2	5.1800E-2	-1.6230E-2
$CF(\hat{f}_3)$	7.5202E-2	1.3570E-2	6.0500E-2	3.1005E-2
$CF(\hat{f}_4)$	-7.8974E-3	4.1021E-3	2.1203E-2	-1.7597E-2
$CF(\hat{f}_i)$\ Time	3PM	6PM	9PM	12PM
$CF(\hat{f}_1)$	2.9620E-1	3.5928	3.3390E-1	9.8100E-2
$CF(\hat{f}_2)$	1.7013E-3	1.0600E-2	1.1610E-1	9.1801E-2
$CF(\hat{f}_3)$	-1.5599E-2	2.8300E-2	1.3940E-1	8.6701E-2
$CF(\hat{f}_4)$	-1.5591E-2	-7.0968E-3	6.5201E-2	1.0202E-2

Table C.10: The *CF* Values When $\hat{\sigma}^2$ Is Desired Accuracy 1.0E-4

From the criterion values for all candidate models in Table C.10, we are convinced that the model f_4 is the true model and the residual variance is around 1.0E-4 except that $CF(\hat{f}_3) < CF(\hat{f}_4)$ at 3PM in Table C.10 which conflicts with the declared conclusion. Now, we examine the model f_3 and f_4 at 3PM from Table C.3 and C.4. The estimated parameters and model statistics are listed in Table C.11

Model\Coef. (variance)	A	B	C	D	E	RSS	R^2
f_3	12.8039	0.0(fixed)	-0.5165	66.8485	252.3384	8.3E-3	0.998
	(0.0000)		(0.0000)	(0.0005)	(16.6387)		
f_4	12.8239	-0.0007	-0.4928	66.6211	238.1101	8.2E-3	0.998
	(0.0001)	(0.0000)	(0.0001)	(0.0098)	(48.9132)		

Table C.11: Comparison between the Estimated Model f_3 and f_4 at 3PM

It is obvious that the corresponding estimated parameters are very close in both models and the overall fits of the two models are not significantly different since the value of the maximum likelihood ratio test f_3 against f_4, $\tilde{\lambda}_{3:4} = 2.9545$, is not significant (see Table C.5). Model f_4 is superior to model f_3 at all three hour intervals of the day except for 3PM. The only reason for this is that the parameters in model f_4 are time variant, and parameter B may be very close to zero at 3PM, therefore, model f_3 is chosen by model criterion CF at this particular time. However, since we seek one model function form for the load/weather relation at all times of a day, we still choose model f_4 as the optimum model function.

C.4 Selection of the Variance Function

Assuming the smoothed $log(|y_t - f(x_t, \Theta)|)$ is the logarithms of a realization from the "true" variance function, and regressing the smoothed $log(|y_t - f(x_t, \Theta)|)$ on the smoothed x_t, we obtain an estimate of Ψ by nonlinear least squares as listed in Table C.12 and Table C.13 for the proposed model h_1 and h_2 respectively.

The test of significance as to whether b differs from zero can be based on the likelihood ratio test for the two nested models to assess if the model h_2 improves

Model $log(\|y-f\|) = a + c\exp\{-(x-d)^2/e\}$					
Time(t)\Coef. (variance)	a	c	d	e	RSS
3AM	-4.1752	1.7824	45.0456	199.8261	0.4426
	(0.0005)	(0.0004)	(0.0013)	(22.9696)	
6AM	-4.1431	2.1032	45.5000	295.0790	0.2467
	(0.0009)	(0.0008)	(0.0007)	(48.8095)	
9AM	-4.3801	1.8046	50.9185	271.2555	0.6715
	(0.0017)	(0.0015)	(0.0026)	(124.2456)	
12AM	-4.6797	1.5074	56.2607	238.1021	0.5573
	(0.0018)	(0.0016)	(0.0038)	(139.5430)	
3PM	-4.8171	1.5492	56.5947	479.6495	0.0930
	(0.0019)	(0.0018)	(0.0025)	(396.2297)	
6PM	-4.8268	2.2506	56.0366	294.1145	0.115
	(0.0004)	(0.0003)	(0.0004)	(18.1126)	
9PM	-4.8775	2.2744	50.6802	477.8534	0.2409
	(0.0040)	(0.0038)	(0.0017)	(377.4077	
12PM	-3.8203	1.3969	47.6306	195.8720	0.2409
	(0.0002)	(0.0002)	(0.0012)	(17.0640)	

Table C.12: Gauss-Newton Nonlinear Least-Squares for Smoothed Residuals

fitting significantly over model h_1. We, therefore, test

$$\mathbf{H_0}: \quad h(x_t, \Psi) = h_1 \tag{C.7}$$

against

$$\mathbf{H_1}: \quad h(x_t, \Psi) = h_2 \tag{C.8}$$

The likelihood ratio test is

$$\lambda = [RSS(\hat{h}_2)/RSS(\hat{h}_1)]^{n/2} \tag{C.9}$$

We reject $\mathbf{H_0}$ if λ is small. Under general regularity conditions, $-2\ln(\lambda) \sim \chi^2$ with degrees of freedom equal to 1. Hence, an α significance level critical region is $\tilde{\lambda} = n\ln[RSS(\hat{h}_1)/RSS(\hat{f}_2)] > \chi^2_{(1-\alpha)}(1)$. The $\tilde{\lambda}$ values at different times of a day are shown in Table C.14.

Noting $\chi^2_{(.99)}(1) = 6.63$, we reject $\mathbf{H_0}$, given by (C.7) at significance level $\alpha = 0.01$. Significant gains in fitting logarithms of absolute residuals are obtained by the model

Model $log(\|y-f\|) = a + bx + c\exp\{-(x-d)^2/e\}$						
Time(t)\Coef. (variance)	a	b	c	d	e	RSS
3AM	-4.4165	0.0037	1.8523	44.6284	213.4366	0.3733
	(0.0020)	(0.0000)	(0.0005)	(0.0055)	(27.9056)	
6AM	-4.0770	-0.0022	2.1364	45.7639	301.5693	0.2316
	(0.0011)	(0.0000)	(0.0010)	(0.0057)	(53.8842)	
9AM	-3.8752	-0.0108	1.8686	52.3393	272.3462	0.1277
	(0.0004)	(0.000)	(0.0003)	(0.0028)	(23.1516)	
12AM	-3.8277	-0.0101	1.2450	57.5560	173.9382	0.2753
	(0.0020)	(0.0000)	(0.0004)	(0.0069)	(32.3565)	
3PM	-3.5361	-0.0120	0.9657	58.8643	278.8338	0.0724
	(0.0097)	(0.0000)	(0.0019)	(0.0388)	(277.0157)	
6PM	-4.5821	-0.0026	2.1532	56.2826	277.0813	0.1085
	(0.0043)	(0.0000)	(0.0009)	(0.0046)	(32.6751)	
9PM	-3.7062	-0.0113	1.7001	52.3494	317.5352	0.1722
	(0.0029)	(0.0000)	(0.0010)	(0.0079)	(92.5598)	
12PM	-3.5392	-0.0041	1.3179	48.1855	174.6949	0.1424
	(0.0005)	(0.0000)	(0.0001)	(0.0024)	(9.0047)	

Table C.13: Gauss-Newton Nonlinear Least-Squares for Smoothed Residuals

	3AM	6AM	9AM	12AM	3PM	6PM	9PM	12PM
	h_0 against h_1							
$\tilde{\lambda}$	38.82	14.53	381.76	151.63	53.08	12.92	455.72	120.39

Table C.14: Likelihood Ratio Statistics

h_2, i.e. the term b in model h_2 cannot be neglected. Therefore, we conclude that the model function h_2 is the better model of the two candidate models.

On the aspect of model stability, we have discussed the model stability function (see equation (C.5) in section C.2. To compare the average stability of the two nested models, we list $\hat{\Delta}_i = \sum_{t=1}^{n} \hat{\delta}_i(x_t)$ in Table C.15 since $\sum_{t=1}^{n} \hat{\alpha}_i(x_t) = m$, the number of unknown parameters in model h_i where $i = 1, 2$ and the error variance σ is unknown. It is obvious that the model h_1 is more stable than model h_2. We have seen from the likelihood ratio test that model h_2 is a significantly better fit than model h_1. The model selection criterion *CF* introduced in section 6.3.3 of chapter 6 can trade-off the model accuracy and stability, and then select a better model.

$\hat{\Delta}_i$\ Time	3AM	6AM	9AM	12AM	3PM	6PM	9PM	12PM
$\hat{\Delta}_1$	1.2710	2.4032	2.9453	4.5246	25.3314	2.0586	6.1769	1.7088
$\hat{\Delta}_2$	2.2613	3.9783	3.7830	4.8422	130.7058	7.8059	9.5455	2.5246

Table C.15: $\hat{\Delta}$ for the Different Models

Assuming the disturbance variance is 1.0E-3, the model selection criterion *CF* for the two nested models are listed in Table C.16. The negative values of *CF* indicate that the assumed disturbance variance is over estimated. The disturbance variance is adjusted to 1.0E-4 and the model criteria are calculated and listed in Table C.17. It can be seen that the *CF* values of model h_1 are consistently larger than *CF* value of model h_2 at corresponding but different times of a day in both Table C.16 and Table C.17. This fact tells us the variance function model h_2 is better than h_1 and it is clear that *CF* weights more heavily the better fitting properties of h_2 than those of the more stability of h_1.

$CF(\hat{f}_i)$\ Time	3AM	6AM	9AM	12AM
$\hat{\sigma}^2$ = 1.0E-3 (overall mean)				
$CF(\hat{h}_0)$	0.2186	0.0207	0.4455	0.3463
$CF(\hat{h}_1)$	0.1503	0.0066	-0.0973	0.0653
$CF(\hat{f}_i)$\ Time	3PM	6PM	9PM	12PM
$CF(\hat{h}_0)$	-0.1149	-0.1030	0.1078	0.0159
$CF(\hat{h}_1)$	-0.1342	-0.1085	-0.0558	-0.0816

Table C.16: The *CF* Values When σ^2 Is Assumed 1.0E-3

$CF(\hat{f}_i)$\ Time	3AM	6AM	9AM	12AM
$\hat{\sigma}^2$ = 1.0E-4 (overall mean)				
$CF(\hat{h}_0)$	0.4202	0.2241	0.6489	0.5362
$CF(\hat{h}_1)$	0.3510	0.2091	0.1052	0.2543
$CF(\hat{f}_i)$\ Time	3PM	6PM	9PM	12PM
$CF(\hat{h}_0)$	0.0722	0.0932	0.3139	0.2184
$CF(\hat{h}_1)$	0.0517	0.0868	0.1494	0.1200

Table C.17: The *CF* Values When σ^2 Is Assumed 1.0E-4

Bibliography

ABOU-HUSSIEN, M. S., KANDIL, M. S., TANTAWY, M. A., and FARGHAL, S. A. (1981): An Accurate Model for Short-Term Load Forecasting, *IEEE Trans. on Power Apparatus and System*, PAS-100(9), pp. 4158 – 4165.

ABRAHAM, B. and BOX, G. E. P. (1978): Deterministic and Forecast-adaptive Time-dependent Models, *Applied Statistics*, 27(2), pp. 120 – 130.

ABU-EL-MAGD, M. A. and SINHA, N. K. (1982): Short-term Load Demand Modelling and Forecasting: A Review, *IEEE Trans. on Syst., Man, Cybern.*, SMC-12(3), pp. 370 – 382.

AKAIKE, H. (1970): Statistical Predictor Identification, *Annal of Institute of Statistics and Mathematics*, 22, pp. 203 – 217.

AKAIKE, H. (1974): A New Look at Statistical Model Identification, *IEEE Trans. Automatic Control*, AC - 19, pp. 716 – 723.

AKAIKE, H. (1975): Markovian Representation of Stochastic Processes by Canonical Variables, *SIAM Journal on Control*, 13, pp. 162 – 173.

ANDERSON, B. D. O. and MOORE, J. B. (1979): *Optimal Filtering*, Prentice-Hall, Inc.

AOKI, M. (1987): *State Space Modelling of Time Series*, Springer-Verlay.

BLOOMFIELD, P. (1976): *Fourier Analysis of Time Series: An Introduction*, John Wiley & Sons Inc.

BODGER, P. S., BROOKS, D. R. D., and MOUTER, S. P. (1987): Spectral Decomposition of Variation in Electricity Loading Using Mixed Radix Fast Fourier Transform, *IEE Proceedings*, 134(3), pp. 197 – 202.

BOROWIAK, D. S. (1989): *Model Discrimination for Nonlinear Regression Models*, Marcel Dekker Inc., New Yark.

BOX, G. E. P. and JENKINS, M. (1976): *Time Series Analysis: forecasting and control*, Holden-Day, 2nd edition.

BOX, G. E. P., PIERCE, D. A., and NEWBOLD, P. (1987): Estimating Trend and Growth Rates in Seasonal Time Series, *Journal of the American Statistical Association*, 82(397), pp. 276 – 282.

BOX, G. E. P. and TIAO, G. C. (1976): Comparison of Forecast with Actuality, *Applied Statistics*, 25, pp. 195 – 200.

BROWN, R. G. (1965): *Smoothing, Forecasting and Prediction of Discrete Time Series*, Engleword Cliffs, NJ: Prentice-Hall.

CAINES, P. E. (1988): *Linear Stochastic System*, John Wiley & Sons.

CAINES, P. E. and MEYNE, D. Q. (1970): On the Discrete-time Matrix Riccati Equation of Optimal Control, *Int. J. Control*, 12(5), pp. 785 – 796.

CAMPO, R. and RUIZ, P. (1987): Adaptive Weather Sensitive Short Term Load Forecast, *IEEE Trans. on Power System*, PWRS-2(3), pp. 592 – 600.

CARROLL, R. J. and RUPPERT, D. (1988): *Transformation and Weighting in Regression*, Chapman and Hall.

CHAMBERS, J. (1973): Fitting Nonlinear Models: Numerical Techniques, *Biometria*, 60, pp. 1 – 13.

CHAMBERS, J. M. (1983): *Graphical Methods for Data Analysis*, Wadsworthm, Belmont, California.

CHAN, S. W., GOODWIN, G. C., and SIN, K. S. (1984): Convergence Properties of the Riccati Difference Equation in Optimal Filtering of Nonstabilizable Systems, *IEEE Transactions on Automatic Control*, AC-29, pp. 10 – 18.

CHRISTIAANES, W. R. (1971): Short-term Load Forecasting, using general exponential smoothing, *IEEE Trans. Power App. Syst.*, PAS-90(2), pp. 900 – 910.

CLEVELAND, J. S. (1979): Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 74(368), pp. 821 – 836.

COHEN, M. L. (1984): *Robust Smoothly Heteroeneous Variance Regression*, Preprint.

DHRYMES, P. J. (1978): *Introductory Econometrics*, Spriger-Verlag New Yark Inc.

DRAPER, N. R. (1984): The Box-Wetz Criterion Versus R^2, *Journal of Royal Statistics Society*, A 147, pp. 100 – 103.

ENGLE, R. F. (1978): Estimating Structural Models of Seasonality, In Zellner, A., editor, *Seasonal Analysis of Economic Times Series*, pp. 281 – 308. Washington D.C: Bureau of the Census.

ENGLE, R. F. and GRANGER, C. W. J. (1987): Cointegration and Error Correction: Representation, Estimation, and Testing, *Econometrica*, 55, pp. 251 – 276.

ENGLE, R. F., GRANGER, C. W. J., and HALLMAN, J. J. (1989): Merging Short- and Long-run Forecasts, *Journal of Econometrics*, 40, pp. 45 – 62.

ENGLE, R. F., GRANGER, C. W. J., RICE, J., and WEISS, A. (1986): Semiparametric Estimates of the Relation Between Weather and Electricity Sales, *Journal of the American Statistical Association*, 81(394), pp. 310 – 320.

FARMER, E. D. and POTTON, M. J. (1968): Development of On-line Load Prediction Techniques with Results from Trials in the South-West Region of the CEGB, *Proc. IEEE*, 115(10), pp. 1549 – 1588.

FRANZINI, L. and HARVEY, A. C. (1983): Testing for Deterministic Components in Time Series Models, *Biometrika*, 70, pp. 673 – 682.

FURNIVAL, G. M. (1971): All Possible Regressions with Less Computation, *Technometrics*, 13, pp. 403 – 408.

FURNIVAL, G. M. and WILSON, R. W. (1974): Regressions by Leaps and Bounds, *Technometrics*, 8, pp. 27 – 51.

GALIANA, F. D., HANSCHIN, E., and FIECHTER, A. R. (1974): Identification of Stochastic Electric Load Models from Physical Data, *IEEE Trans. Automatic Control*, AC-19(6), pp. 887 – 893.

GERSCH, W. and KITAGAWA, G. (1983): The Prediction of Time Series with Trend and Seasonalities, *Journal of Business & Economic Statistics*, 1, pp. 253 – 264.

GOURIEROUX, C., HOLLY, A., and MOUTFORT, A. (1982): Likelihood Ratio Test, Wald Test and Kuhu-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters, *Econometric*, 50, pp. 63 – 80.

GRANGER, C. W. J. (1986): Development in the Study of Co-integrated Economic Variables, *Oxford Bulletin of Economics and Statistics*, 68, pp. 213 – 228.

GROSS, A. M. (1977): Confidence Intervals for Bisquare Regression Estimates, *Journal of the American Statistical Association*, 72, pp. 341 – 354.

GROSS, G. and GALIANA, F. D. (1987): Short-term Load Forecasting, *Proceedings of the IEEE*, 75(12), pp. 1558 – 1573.

GUPTA, P. C. (1971): A Stochastic Approach to Peak Power Demand Forecasting in Electrical Utility Systems, *IEEE Trans. Power App. Syst.*, PAS-90, pp. 824 – 832.

GUPTA, P. C. and YAMADA, K. (1972): Adaptive Short-term Forecasting of Hourly Loads. using weather information, *IEEE Trans. Power App. Syst.*, PAS-91, pp. 2085 – 2094.

HAGAN, M. T. and BEHR, S. M. (1987): The Time Series Approach to Short Term Load Forecasting, *IEEE Transaction on Power System*, PWRS-2(3), pp. 785 – 791.

HAGAN, M. T. and KLEIN, R. (1977): Off-line and Adaptive Box and Jenkins, Applied to the Problem of Short-term Load Forecasting, *Proc. Lawrence Symp. Syst. and Decision Sci.*, Berkeley, CA.

HAGAN, M. T. and KLEIN, R. (1978): On-line Maximum Likelihood Estimation for Load Forecasting, *IEEE Trans, Syst., Man, Cybern*, SMC-8(9), pp. 711 – 715.

HAGGAN, V. and OYETUNJI, O. B. (1984): On the Selection of Subset Autoregressive Time Series Models, *Journal of Time Series Analysis*, 5(2), pp. 103 – 113.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., and STAHEL, W. A. (1986): *Robust Statistics: The Approach Based on Influence Functions*, Wiley: New Yark.

HANNAN, E. J. (1970): *Multiple Time Series*, Wiley: New Yark.

HANNAN, E. J. and DEISTLER, M. (1988): *Linear System*, John Wiley & Sons.

HANNAN, E. J. and QUINN, B. G. (1979): The Determination of the Order of An Autoregression, *Journal of Statistics Society. Series B*, 41, pp. 190 – 915.

HANNAN, E. J., TERRELL, R. D., and TUCKWELL, N. (1970): The Seasonal Adjustment of Economic Time Series, *International Economic Review*, 11, pp. 24 – 52.

HARVEY, A. C. (1976): Estimating Regression Models with Multiplicative Heteroscedasticity, *Econometrics*, 44, pp. 461 – 465.

HARVEY, A. C. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

HARVEY, A. C. and PHILLIPS, G. D. A. (1979): The Estimation of Regression Models with Autoregressive-moving Average Disturbances, *Biometrika*, 66, pp. 49 – 58.

HARVEY, A. C. and TODD, P. H. J. (1983): Forecasting Economic Time Series with Structural and Box-Jenkins Models, *Journal of Business and Economic Statistics*, 1, pp. 299 – 315.

HEALY, M. J. R. (1984): The Use of R^2 as A Measure of Goodness of Fit, *Journal of Royal Statistics Society*, A 147, pp. 605 – 909.

HELLEND, I. S. (1987): On the Interpretation and Use of R^2 in Regression Analysis, *Biometrics*, 43, pp. 61 – 69.

HOAGLIN, D. C., MOSTELLER, F., and TUKEY, J. W. (1983): *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

HOCKING, R. R. and LESLIE, R. N. (1967): Selection of the Best Subset in Regression Analysis, *Technometrics*, 9(4), pp. 531 – 540.

HOLST, J. and JONSSON, G. (1984): Adaptive Short Term Prediction of Power Load, *IFAC: 9th Triennial World Congress*, pp. 2013 – 2018. Budapest, Hungary.

HUBER, P. J. (1981): *Robust Statistics*, Wiley, New York.

IEEE COMMITTEE REPORT (1980): Load Forecast Bibliography: Phase I, *IEEE Trans. Power App. Syst.*, PAS-99(1), pp. 53 – 58.

IEEE COMMITTEE REPORT (1981): Load Forecast Bibliography: Phase II, *IEEE Trans. Power App. Syst.*, PAS-100(7), pp. 3217 – 3220.

IMHOF, J. P. (1961): Computing the Distribution of Quadratic Forms in Normal Variables, *Biometrics*, 48, pp. 419 – 462.

IRISARRI, G. D., WIDERGREN, S. E., and YEHSAKUL, P. D. (1982): On-line Load Forecasting for Energy Control Center Application, *Applied Statistics*, PAS-101(1), pp. 71 - 78.

JUDGE, G. G. (1985): *The Theory and Practice of Econometrics*, Wiley, New Yark, 2 edition.

KAILATH, T. (1980): *Linear Systems*, Engleword Cliffs, N.J. Prentice-Hall.

KALMAN, R. E. (1960): A New Approach to Linear Filtering and Prediction Problems, *Trans. ASME, J. Basic Engineering*, 82, pp. 34 - 45.

KALMAN, R. E. and BUCY, R. S. (1961): New Results in Linear Filtering and Prediction Theory, *Trans. ASME, J. Basic Engineering*, 83, pp. 95 - 107.

LEHMANN, E. L. (1959): *Testing Statistical Hypotheses*, New York: Wiley.

LIJESEN, D. P. and ROSING, J. (1971): Adaptive Forecasing for Hourly Loads Based on Load Measurements and Weather Information, *IEEE Trans. Power App. Syst.*, PAS-90(4), pp. 1757 - 1767.

LU, Q. C., GRADY, W. M., CRAWFORD, M. M., and ANDERSON, G. M. (1989): An Adaptive Nonlinear Predictor with Orthogonal Escalator Structure for Short-Term Load Forecasting, *IEEE Transaction on Power Systems*, 4(1), pp. 158 - 164.

MANN, H. B. and WALD, W. (1943): On the Stochastical Treatment of Linear Stochastic Difference Equations, *Econometrica*, 11, pp. 173 - 220.

McCLAVE, J. (1975): Subset Autoregression, *Technometrics*, 17(2), pp. 213 - 219.

METTEREN, H. P. V. and SON, P. J. M. V. (1979): Short-term Load Prediction with A Combination of Different Models, *Proc. PICA Conf.*, pp. 192 - 197.

MOUTTER, S. P., BODGER, P. S., and GOUGH, P. T. (1986a): Spectral Decomposition and Extrapolation of Variations in Electricity Loading, *IEE Proceedings*, 133(5), pp. 247 – 255.

MOUTTER, S. P., BODGER, P. S., and GOUGH, P. T. (1986b): A Super-resolution Algorithm for Spectra Estimation and Time Series Extrapolation, *Journal of Forecasting*, 5, pp. 169 – 187.

NYBLOM, N. (1986): Test for Deterministic Linear Trend in Time Series, *Journal of the American statistical Association*, 81(394), pp. 545 – 549

PAGANO, M. (1972): An Algorithm for Fitting Autoregressive Schemes, *Journal of Royal Statistics Society. Series C*, 21, pp. 274 – 281.

PAGANO, M. (1978): On Periodic and Multiple Autoregressions, *The Annals of Statistics*, 6(6), pp. 1310 – 1317.

PANDIT, S. M. and WU, S. M. (1983): *Time Series and System Analysis with Application*, New York: John Wiley.

PANUSKA, V. and KOUTCHOUK, J. P. (1975): Electrical Power System Load Modelling by Two-stage Stochastic Approximation Procedure, *Proc. 6th Triennial World IFAC Congress, Part IIA.*

PAPOULIS, A. (1975): A New Algorithm in Spectral Analysis and Bandlimited Extrapolation, *IEEE Trans. Circuit and System*, CAS-22(9), pp. 735 – 742.

PAPOULIS, A. and CHAMZAS, C. (1979): Detection of Hidden Periodicities by Adaptive Extrapolation, *IEEE Trans. on Acoustics, Speech & Signal Processing*, ASSP-27(5), pp. 492 – 500.

PARZEN, E. (1974): Some Recent Advances in Time Series Modelling, *IEEE Trans. Automatic Control*, AC - 19, pp. 723 – 730.

PARZEN, E. (1982): ARARMA Models for Time Series Analysis and Forecasting, *Journal of Forecasting*, 1, pp. 67 – 82.

PENM, J. H. W. and TERRELL, R. D. (1982): On the Recursive Fitting of Subset Autoregressions, *Journal of Time Series Analysis*, 3(1), pp. 43 – 59.

PRIESTLEY, M. B. (1982): *Spectral Analysis and Time Series*, volume 1, Academic Press Inc.

RAJURKAR, K. P. and NISSEN, J. L. (1985): Data-Dependent Systems Approach to Short-Term Load Forecasting, *IEEE Trans. on Syst., Man, Cybern.*, SMC-15(4), pp. 532 – 536.

RATKOWSKY, D. A. (1983): *Nonlinear Regression Modeling*, Marcel Dekker Inc. New York.

ROGERS, A. J. (1986): Modified Lagrange Multiplier Test for Problems with One-side Alternatives, *Journal of Econometrics*, 31, pp. 341 – 361.

ROSENBERG, B. (1973): Random Coefficient Models: the analysis of a cross-section of time series by stochastically convergent parameter regression, *Annals of Economic and Social Measurement*, 2, pp. 399 – 428.

SACHDEV, M. S., BILLINTON, R., and PETERSON, C. A. (1977): Representative Bibliography on Load Forecasting, *IEEE on Power Apparatus and System*, PAS-96(2), pp. 697 – 700.

SCHWARZ, G. (1978): Estimating the Dimension of A Model, *Annal of Statistics*, 6, pp. 461 – 464.

SEBER, G. A. F. (1984): *Multivariate Observations*, John Wily & Sons, Inc.

SEBER, G. A. F. and WILD, C. J. (1989): *Nonlinear Regression*, John Wily & Sons, Inc.

SCHWEPPE, F. C. (1965): Evaluation of likelihood functions for Gaussian signals, *IEEE Trans. Inform. Theory*, IT-11 61-70.

SHARMA, K. L. S. and MAHALANABIS, A. K. (1974): Recursive Short-term Load Forecasting Algorithm, *Proc. Inst. Elec. Eng.*, volume 121, pp. 59 – 62.

SHEPHARD, N. G. and HARVEY, A. C. (1990): On the Probability of Estimating A Deterministic Component in the Local Level Model, *Journal of Time Series Analysis*, 11(4), pp. 339 – 347.

SINGH, G., BISWAS, K. K., and MAHALANABIS, A. K. (1977): Power System Load Forecasting, using smoothing techniques, *IEEE on Power Apparatus and System*, PAS-96(2), pp. 697 – 700.

STANTON, K. N., GUPTA, P. C., and EL-ABIAD, A. H. (1969): Long Range Demand Forecasting for Electric Utility Industry, *Proc. PICA Conf.*

TONG, H. (1977): Some Comments on the Canadian Lynx Data, *Journal of Royal Statistics Society. Series A*, 140, pp. 432 – 436.

TOYODA, J., CHEN, M.-S., and INOUE, Y. (1970a): An Application of State Estimation to Short-Term Load Forecasting, Part I: Froecasting Modelling, *IEEE Trans. on Power App. and Syst.*, PAS-89(7), pp. 1678 – 1682.

TOYODA, J., CHEN, M.-S., and INOUE, Y. (1970b): An Application of State Estimation to Short-Term Load Forecasting, Part II: Implementation, *IEEE Trans. on Power App. and Syst.*, PAS-89(7), pp. 1683 – 1688.

VEMURI, S., HILL, E. F., and BALASUBRAMANIAN, R. (1973): Load Forecasting Using Stochastic Models, *Proc. PICA Conf.*, volume TP1-A, pp. 31 – 37.

VEMURI, S., HOVEIDA, B., and MOHEBBI, S. (1986): Short Term Load Forecasting Based on Weather Load Models, *IFAC Power Systems and Power Plant Control*, Beijing, China.

WATTERS, R. L. (1987): Error Modeling and Confidence Interval Estimation for Inductively Coupled Plasma Calibration Curves, *Anal. Chem.*, 59, pp. 1639 – 1643.