MONASH UNIVERSITY

THESIS ACCEPTED IN SATISFACTION OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ON 7 September 2001

[redacted]

Sec. Research Graduate School Committee

2475/3743

ERRATA

p 17, Section (2.2.2), line 5 from top: "+" for "−".

p 21, Section (2.2.7), line 3 from bottom: "Pesaran" for "Pearson".

p 42, Section (2.4.3.1), line 6 from bottom: "$-1 < \theta < 1$" for "$0 < \theta < 1$".

p 47, Section (2.4.4.1), line 13 from top: "invertibility" for "inevitability".

p 51, Section (2.4.5.2), line 5 from bottom: "second and first" for "first and second actual" and "$b_0 = y_2 - y_1$" for "$b_0 = \frac{1}{2}(y_1 + y_2)$".

p 51, Section (2.4.5.2), line 4 from bottom: "$b_0 = \frac{1}{2}((y_2 - y_1) + (y_4 - y_3))$" for "$b_0 = \frac{1}{2}(y_2 - y_1) + (y_4 - y_3)$".

p 55, Section (2.5), line 6 from bottom: "study" for "studiy".

p 79, replace the second last sentence of second paragraph with "We hope that the ICL based information criteria will give better probabilities of correct selection."

p 83, Section (3.6.2), line 4 from bottom: "seasonality" for "seasonalitity".

p 129, lines 10 and 11, replace the description of design matrix $X2$ with "A constant dummy, the first n observations of Durbin and Watson's (1951, p.159) annual consumption of spirits example (two variables, namely, "INCOME" and "PRICE" were used)."

p 130, Section (4.6), line 9 from top: "performs uniformly better" for "performs better".

p 145, Section (5.2), line 12 from top: "outside" for "inside".

p 146, Section (5.2), line 2 from top: "$\mathcal{D}(\theta) = \Omega(\theta)^{-\frac{1}{2}}$" for "$\mathcal{D}(\theta) = \otimes(\theta)^{-\frac{\infty}{\epsilon}}$".

p 154, Section 5.5.2, line 9 from the bottom: "better" for "the best".

p 158, Section (5.7), line 11 from top: "Chapter" for "Cahpter".

p169, replace the last sentence of first paragraph with "With this in mind, we use IC for selecting forecasting models for the M3 competition data of Makridakis and Hibon (2000)."

p 171, Section (6.2): in equations (6.2.5) and (6.2.6) "T^{*j}" for "T^*" in the right hand sum.

p 179, Table 6.2, selection percentages for the SM2 model for AIC, MCp, GCV and FPE should be 0.56, 0.49, 0.56 and 0.56, respectively instead of 5.6, 4.9, 5.6 and 5.6, respectively.

p 187, lines 1 and 2 from bottom: delete "Difference (ISM minus BIC) of ..... data" and read "Difference of selection percentages for various IC from ISM".

p 188, Figure 6.4(c): "p_4" for "p_3".

P 198, line 2 from bottom: "1989" for "1996" and "4th" for "5th".

p 210, line 4 from top: include "3rd edition" between "Applications," and "John".

ADDENDUM

p 198, between lines 18 and 19 from top: add the reference "Burnham, K.P. and Anderson, D.R. (1998). *Model Selection and Inference: A Practical Inference-theoretic Approach*, Springer: New York."

p 211, line 11 from top: delete "Montgomery, D.C. and Johnson, L.A. (1976)" and read "Montgomery, D.C., Johnson, L.A. and Gardner, J.S. (1990)". Also include "3rd edition" between "Analysis," and "McGraw-Hill" in line 12 from top.

p 213, delete lines 7 to 9 (reference"Ray, B.K. and Crato, N. (1994))" from top.

p 215 & 216, delete the references "Sweet, A.L. (1983a) and Sweet, A.L. (1983b)."

p 215, between lines 23 and 24 from top: add the reference "Sweet, A.L. (1985). Computing the variance of the forecast error for the Holt-Winters seasonal models. *Journal of Forecasting,* 4, 235-243."

p 157: delete the last paragraph of section 5.6 and read "As was discussed earlier of this section that we require design matrices of order $n \times k$, where $n = 300$. However, the given row dimension, n^*, of the design matrices $X2$ to $X5$ are smaller than n. Therefore, we constructed modified design matrices of order $n \times k$ for the design matrices $X2$ to $X5$. The procedure is as follows. For a particular design matrix, draw n random integers between 1 to n^* by using the uniform distribution and for each number pick the corresponding row of the original design matrix. Then, use the selected row at the ith, $i = 1, 2, \cdots, n$, drawing as the ith row of the modified design matrix. This gives the required modified design matrix of order $n \times k$."

Add at the end of Chapter 4: " In this chapter, the optimal penalty function was estimated by optimizing the overall average probabilities of correct selection. The same set of generated data were used to estimate penalties and hence, for calculating overall average probabilities of correct selection. This may give biased probabilities of correct selection. However, from the results of Chapter 6 (where real data were used to estimate the optimal penalties) it is evident that the bias, if any, is minimal. Therefore, we are very optimistic that the results of Chapter 4 will carry through even if the penalty values are estimated from one set of simulated data and then these penalties are evaluated using another set of data, which is generated changing the seed number for random generation."

Add at the end of Chapter 5: " In this Chapter, the optimal penalties were estimated for selecting models for forecasting. Similar to Chapter 4, the same set of data were used for estimating penalties as well as to minimize overall average mean square error. The question of biased forecast error may arise. However, forecast error comparison between individual selection method (which is based on optimal penalty estimation) and existing information criteria in Chapter 6 shows that the individual selection method performs better than the existing information criteria for all forecasting horizons. Therefore, we are optimistic that the results of Chapter 5 will carry through even if the estimated penalties are evaluated using another set of simulated data."

MODEL SELECTION FOR TIME SERIES FORECASTING MODELS

A thesis submitted for the degree
of Doctor of Philosophy

Baki Billah
B.Sc. (Hon), M.Sc. (Dhaka), MAS (Memorial)

Department of Econometrics and Business Statistics
Faculty of Business and Economics
Monash University
Australia

March 2001

Contents

List of Tables

List of Figures

Abstract

The first contribution of this thesis is to investigate the use of information criteria (IC) based model selection procedures for choosing between different exponential smoothing models. A preliminary study showed that the probabilities of correct selection (PCS) depend on the values of exponential smoothing parameters. Therefore, we calculated average probabilities of correct selection (APCS) as well as overall APCS (OAPCS) in order to assess different selection strategies. We also proposed an improved conditional likelihood (ICL) method, based on marginal likelihood (MGL) methods, and compared the performances of conditional likelihood (CL) and ICL based existing IC procedures. The results of this study show that in terms of OAPCS, ICL based Bayesian Information Criterion (BIC) performs better than the other existing IC procedures considered in this study.

The second contribution is to investigate IC based model selection procedures for selecting autoregressive moving average (ARMA) errors in the context of the linear regression model. We observe that optimal penalties can be found by maximizing OAPCS. This requires the use of a robust optimization procedure capable of optimizing average probabilities obtained by simulation. A new penalty estimation method (PEM) can be applied for estimating optimal penalties. PEM, based on grid search (PEM-GS) is one possible method, but it does involve high level computational cost. As a result of our simulation experiment, we recommend the use of PEM, based on the simulated annealing (SA) algorithm (PEM-SA) to find optimal penalties for use with maximized MGL for this model selection problem.

In addition, we extend the PEM-GS and PEM-SA for selecting the order of autoregressive (AR) errors in the context of the linear regression model with forecasting accuracy as the selection criteria rather than OAPCS. The results of our simulation study show that for shorter forecast horizons no one existing IC pro-

cedure performs best. However, in general, BIC performs best for longer forecast horizons. Both the PEM-GS and PEM-SA perform better than the existing IC procedures including BIC. In general, PEM-GS performs better than PEM-SA, but with additional cost of high computational time. Hence, we recommend MGL based PEM-SA for selecting time series forecasting models.

Finally, based on PEM, we outlined an individual selection method (ISM) for selecting forecasting models for real life time series data. We applied the ISM to the M3 competition data of Makridakis and Hibon (2000) divided into three groups, namely, annual, quarterly and monthly data. The performance of ISM, the combined method (COM) of Makridakis et al. (1982) and various existing IC procedures are compared. The results of this study show that the ISM method performs better than the COM method and the existing IC procedures.

The main contributions of this thesis relate to the development of MGL based small sample model selection procedures. In general, the PEM performs better than those existing IC procedures used in our study for choosing small sample ARMA disturbance processes in the context of the linear regression model. Also, ISM is better than the existing IC procedures for selecting exponential smoothing models for the M3 competition data.

Acknowledgement

First of all, I wish to express my deep gratitude to almighty God, whose divine help enabled me to complete this thesis. In this regard, I wish to extend my sincere gratitude to my thesis supervisor, Professor Maxwell L. King, who first introduced me to the subject of model selection with his characteristic enthusiasm. His constructive criticisms, continuous encouragement, and above all friendly affection during the entire course work were invaluable. He read my work whenever I needed, and provided a lot of insightful comments with necessary English corrections which helped the thesis to a great extent. This work would not have been completed without his great help and support throughout the course of this thesis. I heartily thank Professor King very much for the help he provided me whenever I needed it.

I would like to thank Professor Anne B. Koehler, Associate Professor Ralph D. Snyder and Associate Professor Rob J. Hyndman for their valuable comments and discussions concerning aspects of this thesis.

A note of thanks must go to Mrs. Nevin Chowdhury, Faculty of Education and Mrs. Glenda Crosling, Language and Learning Services, Faculty of Business and Economics, Monash University for proof reading of my thesis.

Many thanks are also due to Mrs. Mary Englefield, Ms. Inge Meldgaard, Mrs. Philippa Geurens, Ms. Julie Larsen, Mrs. Pauline Froggatt, Mr. Myles Tooher and Ms. Deanna Orchard for their administrative support in various ways that have helped me successfully complete this thesis.

Also, Mr. M. N. Azam, Dr. Zakir Hossain, Dr. Munir Mahmood, Mrs. Mahbuba Yeasmin and Mr. Jahar Lal Bhowmik deserve my hearty thanks for their support and encouragement. Particularly, I am grateful to Mr. Azam for his help at the beginning of my Ph.D. program at Monash University.

I acknowledge with gratitude my debt to Monash University for awarding me a Monash Graduate Scholarship (MGS) and an Overseas Postgraduate Research Scholarship (OPRS), and to the Department of Econometrics and Business Statistics for offering me a Departmental Scholarship for the last four months and a fraction Assistant Lecturer position since March 1997 – all to assist my study for this thesis, financially.

My unceasing thanks must go to my beloved mother, Rabiya Khatoon, and father, Md. Mobarok Ali Bishwas. My achievements today would not have been possible without their love, sacrifices and moral support. I am also grateful to all of my brothers, sisters and relatives for their encouragement and love.

Finally, I would like to express my deep appreciation to my wife, Ripa, for her moral support, love, patience and understanding during the entire course of my study. I am grateful to my little daughter, Tabassum, who always thinks that I am a student like her.

Declaration

I hereby declare that, this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution, and that, to the best of my knowledge and belief, this thesis contains no material previously written by another person, except where due reference is made in the text of the thesis.

Md. Baki Billah

Chapter 1

Introduction

1.1 Background

In econometrics we are often forced to use data to make a choice between a number of competing alternative models. Thus, an obvious question arises as to which model provides the best characterization from the view point of the data. Econometric modeling is largely about the process of searching for a suitable specification or to put it another way, selecting the right model for the right job. In general, a number of models are typically considered reasonable and a selection is made based on how well each of them appears to fit the observed data. This is typically known as model selection in the econometrics literature.

In fact, in econometric applications, we expect a great deal of reliability from the selected model. In most practical situations, such a selection is made with a limited number of sample observations, which involve partial information regarding the underlying phenomena. When dealing with such cases of very little sample information, there may be many ways to make mistakes in deciding on the best possible model. In spite of this, we use our preferred model to perform a number of tasks such as forecasting future values, making inference about some or all of its parameters and conducting sensitivity analysis on the assumptions of the

model. Therefore, prudent care should be taken to choose the techniques for model selection.

The problem of selecting a true model, or the choice of an appropriate or best model for a given data set, is considered to be a fundamental problem in econometric modeling. For a given situation, the best model depends upon the loss function under consideration. If the aim is to use the model for forecasting, for example, then the best model will be that for which the forecast error, on average, is minimum in some sense. Therefore, the best model selection method could also depend on the characteristics of the data generating process (DGP) of both the true model and also of other models we are willing to entertain. However, in practice, we have no direct knowledge about the DGP. Typically, researchers use the information which is available at the time of modeling to develop rules that will lead us in the direction of the best model.

Econometric modeling which is used to help us understand complex economic relationships should be designed so that the model demonstrates the main characteristics of the economic system being modeled. A very complex model should be avoided so that the basic understanding is not jeopardized. However, while making the model simpler, care is needed to make sure that the important features of the data are not overlooked. This is because, in such cases, the estimated parameters of the model will suffer from an omitted variable bias which will result in invalid or at the very least, questionable inference. One must, therefore, compromise between the two and select models that are simple and fit the data well.

Monte Carlo experiments can be used to investigate the properties of various model selection procedures through the analysis of model selection probabilities. In such experiments, the data are generated from one of the models in the plausible group and the correct model is selected if the chosen model is the DGP. In

this thesis, we consider the model selection problem in the context of exponential smoothing methods (models) and regression models with autoregressive moving average (ARMA) error processes. However, the vast range of exponential smoothing methods as well as ARMA error processes make model selection more difficult because there are numerous DGPs from which to choose. Therefore, to make our study more manageable, we consider some simple models which are widely used, and which also perform well in applications.

1.2 Model Selection Procedures

A number of model selection methods have been suggested in the literature. These methods can mainly be classified into the following four categories:

(i) *Classical Hypothesis Testing Procedures:* This approach is one of the oldest methods developed for model selection problems. Classical test statistics such as t and F tests are used in the context of the linear regression model for small sample model selection problems, while for the models other than the linear model, Wald (W), likelihood ratio (LR) and Lagrange multiplier (LM) tests are often used.

(ii) *Minimum Residual Variance or Minimum Mean Prediction Error:* In this approach, models are selected by minimization of some function of the model's errors or forecast (prediction) errors, such as mean square error (MSE) or mean absolute percentage error (MAPE). The model with the smallest average forecast error is considered to be the best model.

(iii) *Bayesian Criteria*: Bayesian criteria uses posterior odds ratios for model comparison and is recognized as a well established concept in model selection. More details about Bayesian criteria can be found in DeGroot (1970).

(iv) *Information Criteria* (IC): IC is the most viable and popular model selection approach in the econometrics literature. Clayton et al. (1986) showed that IC based

model selection procedures can be regarded as a more substantial approach than any other procedures. Granger et al. (1995) pointed out that IC based procedures have a number of advantages over hypothesis test based approaches, and hence, have become more popular to practitioners. IC is defined as the maximized log-likelihood function minus a penalty function. The general form of IC is given by

$$\mathrm{IC}_i = \log L_i(\widehat{\theta}) - p_i(n, q), \tag{1.2.1}$$

where i denotes the ith model, $\log L_i(\widehat{\theta})$ is the maximized log-likelihood function for the ith model, θ is the $q \times 1$ vector of unknown parameters and $p_i(n, q)$ is the penalty function for the ith model, which is typically a function of sample size (n) and number of parameters (q) in the model. Among the models in the plausible group, the model with the largest IC is chosen as the best model.

The IC approach discussed above, which is based on maximizing the penalized maximized log-likelihood, is one of the three modes that IC can sometimes take. The other two are: choosing the model (a) that minimizes the penalized residual variance and (b) that maximizes the penalized posterior probability using Bayes theorem. In the context of linear regression model, as shown by Fox (1995), all the existing IC procedures can be expressed as Minimum Residual Variance criteria or equivalently, Theil's adjusted R^2, which is a function of the least squares estimates of the sum of squared errors. The Schwarz criterion (known as the Bayesian Information Criterion or BIC) is based on an approximation to the asymptotic expansion of the posterior probability that the model is true under a number of assumptions. In other words, a Bayesian procedure, which selects the model with the highest posterior probability is asymptotically equivalent to BIC. Thus, the IC approach covers both Minimum Residual Variance and asymptotically at least one particular Bayesian criterion.

Popular hypothesis testing procedures for model selection have some disadvantages. For example, different choices of the significance level for any pairwise hypothesis test may lead to a different solution. Also, because only one model is considered as the null hypothesis, the null hypothesis model is unfairly favored if the test lacks power, and when the power of the test is very high, the test could disadvantage the null hypothesis model. This classical approach also results in biases due to incorrect sizes of the test statistics caused by pre-test biases. Despite these limitations of hypothesis testing as a model selection procedure, it has a clear link with the IC approach. Potscher (1991) observed that the use of an IC approach to select a model is equivalent to testing each model against all the alternatives (by means of likelihood ratio tests) and selecting the model that is accepted against all alternative models. The penalty function of the IC approach determines the critical values, and hence, the significance level of the tests. Among the four model selection procedures outlined above, clearly, IC is the more general approach and in this thesis we focus only on IC based model selection procedures.

1.3 Model Selection Problems in ARMA and Exponential Smoothing Models

In the literature, IC based model selection procedures are the most widely used class of model selection procedures. Stone (1981) argued that IC procedures are simple and take into account parameter parsimony when choosing a model from a group of competing models to describe a given set of data. In general, for a given level of accuracy, simpler models are preferable to more complicated models. In the literature, many IC based model selection procedures have been proposed. For example, some prominent criteria are: Akaike's Information Criterion (AIC) of Akaike (1973), Bayesian Information Criterion (BIC) of Schwarz (1978) and Rissanen (1978) and

Hannan-Quinn Criterion (HQ) of Hannan and Quinn (1979). Among other criteria, Mallows' Criterion (MCp) of Mallows (1964), Generalized Cross Validation Criterion (GCV) of Schmidt (1971) and Finite Prediction Error Criterion (FPE) of Amemiya (1972, 1980) are also often used in applications in the literature. Most of these IC based procedures are derived on asymptotic arguments. Asymptotic properties of IC procedures, mainly for AIC, BIC and HQ are well documented, particularly in the context of ARMA models. However, much less is known about their small sample properties. In fact, comprehensive Monte Carlo studies to evaluate the relatively small sample performance of various IC procedures are few. The majority of research in this area has been related to asymptotic properties, and Monte Carlo studies have mainly been used to illustrate the asymptotic results. Further, the asymptotic as well as the finite sample properties of the IC procedures for selecting exponential smoothing methods (models) are still unexplored. In this thesis, we aim to investigate these properties of the IC procedures when selecting a model from a set of simple competing models.

In the literature, there is a long standing debate about the proper form of the penalty function in IC based model selection procedures. This is because, from the definition of IC, clearly, one can easily design a new criterion by slightly changing the value of the penalty function, which can also be justified asymptotically. As such, interest in introducing various IC based procedures for different types of models continues to grow and such vigorous growth in the literature may make the users confused as to which IC procedure to use for a particular problem in hand. Further, the small sample performance of these new IC procedures may not be satisfactory. Therefore, an IC based procedure that would work well for any kind of model selection problem is a current gap in the IC literature. In this thesis, we try to reduce this gap by introducing a new approach for model selection.

Hurvich and Tsai (1989) provide an exception to the above picture by suggesting small sample criteria. With the aim of improving the small sample performance of IC based model selection, King, Forbes and Morgan (1995) and Forbes, King and Morgan (1995) proposed a new approach for estimating penalties through simulation. This method is known as the controlled probabilities approach. Hossain and King (1998) applied this approach to Box-Cox transformation models and found that it produces high selection rates in picking the true (data generating) model. Further, in the context of maximizing the overall average probabilities of correct selection (OAPCS), King and Bose (2000), Kwek and King (1997a, 1997b, 1998) and Azam and King (1998) considered model selection problems in linear regression models, conditional heteroscedastic models and structural break models, respectively. The OAPCS is calculated by averaging the average probabilities of correct selection (APCS) for all models in the plausible group. All of these applications produced, on average, a high probability of selecting the true model. This has motivated us to develop a new class of model selection approaches for the linear regression with ARMA error processes. This new approach can also be used for other model selection problems. For the sample size and plausible models under consideration, this new model selection approach will maximize the OAPCS through the estimation of penalty values numerically. The OAPCS is a step function, and hence, it may not be easy to maximize it using standard methods.

The above function can be maximized by estimating penalty values numerically, which we call the penalty estimation method (PEM). Grid search (GS) could be one of the many successful ways for estimating penalty values so that OAPCS is maximized, and in this thesis we denote this method by PEM-GS. This function can also be maximized by exploiting the use of a new global optimization algorithm called simulated annealing (SA). The SA algorithm works well, even for very

complicated functions such as functions with a large number of local maxima (see Corana et al., 1987; Kirkpatrick et al., 1983; Goffe et al., 1994). The SA algorithm can be modified to apply for functions like ours with ridges and plateaux. We call this modified algorithm PEM-SA. A contribution of this thesis is to investigate the use of the SA algorithm for this purpose.

Often the reason for selecting a model is for use in forecasting. All of the existing IC procedures select a model on the basis of its within-sample fit. However, a model's good within-sample fit may not necessarily provide better out-of-sample forecasts. Therefore, the PEM is extended to selecting the forecast model on the basis of the model's out-of-sample forecasting performance.

Briefly, the overall main aims of this thesis are as follows:

- To introduce IC methods for selecting the best model from a group of competing exponential smoothing models, and to compare the model selection performance of various existing IC procedures with respect to OAPCS.

- To apply the idea of OAPCS to time series model selection from regression models with ARMA error processes, and to exploring the use of PEM (PEM-GS and PEM-SA) for model selection through the estimation of optimal penalties that maximize OAPCS.

- To apply PEM to time series with forecasting accuracy as the model selection criteria rather than OAPCS.

- To introduce a new individual selection method (ISM), which is based on PEM, for selecting forecast models for real life time series such as the M3 competition data of Makridakis and Hibon (2000).

Further, we also aim:

- To compare the performance of profile likelihood (PL) and marginal likelihood (MGL) based IC procedures.
- To introduce improved conditional likelihood (ICL) based model selection in exponential smoothing via maximum MGL estimation methods.

1.4 Outline of the Thesis

In Chapter 2, various existing IC procedures are discussed and it is argued that, in small samples, no one IC procedure performs consistently better for all model selection problems. The SA algorithm is outlined, and it is expected that in small samples, the implementation of this algorithm will provide better model selection properties than the existing IC procedures. Further, this chapter includes a survey of exponential smoothing methods and their corresponding state space models, which have gained much popularity, particularly in inventory forecasting. An automatic model selection approach (for selecting exponential smoothing methods), which has been overlooked in the exponential smoothing literature, is also advocated in this chapter.

A comprehensive Monte Carlo study to compare the small sample performance of various IC procedures in selecting exponential smoothing models has never been considered in the literature. The majority of the research in this area has been related to the forecasting performance of various exponential smoothing methods. Chapter 3 compares the performance of different IC procedures for selecting exponential smoothing models discussed in Chapter 2. This chapter also outlines the theory of OAPCS and proposes ICL based model selection procedures via maximum MGL estimation methods.

As discussed in Section 1.3, this thesis attempts to utilize a recent algorithm (SA) that has been successfully implemented to optimize various types of complex

functions such as step functions. In Chapter 4, the theory of optimal penalties is outlined and the SA algorithm is modified to estimate the optimal penalties so that the OAPCS is maximized. Then, this chapter compares the small sample performance of some of the existing IC procedures and PEM (PEM-GS and PEM-SA) in selecting regression models with ARMA error processes, based on a large Monte Carlo study. Further, this chapter compares the MGL and PL based model selection procedures.

Chapter 5 is an extension of Chapter 4 to model selection for forecasting. In this chapter, PEM-GS and PEM-SA are extended so that, irrespective of the DGP, these methods can select the model with minimum out-of-sample forecast error on average. This chapter also extends the theory of optimal penalties outlined in Chapter 4 to model selection for forecasting. In this chapter, MGL based selection procedures are considered, because they gave better OAPCS in Chapters 3 and 4.

Chapter 6 discusses how PEM, which has been proposed in Chapter 5 in the context of model selection and forecasting, can be used to develop a new model selection procedure (called ISM) for forecasting observations for the M3 competition data of Makridakis and Hibon (2000). The combined forecasting (COM) method of Makridakis et al. (1982), which performs better than the individual forecasting methods is also discussed. The performance as measured by mean absolute percentage error (MAPE) of the new method, the COM method and the existing IC procedures are compared.

Chapter 7 concludes our study by summarizing our results and main findings. In summary, the theme of this thesis is that small sample properties are important in selecting a model selection procedure rather than asymptotic properties. The simulation studies have been conducted only in the context of exponential smoothing models and regression models with ARMA error processes. However, we are

optimistic that the proposed model selection procedures will also perform better than the existing IC procedures for model selection problems outside of ARMA error models and exponential smoothing models, and also for bigger sets of plausible models.

Chapter 2

Literature Review

2.1 Introduction

Model selection procedures play an important role in econometric as well as in statistical modeling. In the model selection literature, many methods have been suggested for selecting an appropriate model from a group of reasonable models. In fact, the area of model selection is quite vast in its scope, and therefore, a complete treatment is beyond the scope of this thesis. Among the various methods for model selection, IC based procedures are widely used and are very powerful methods for choosing among competing models (see Hampel et al., 1986). This chapter aims to review the literature on some popular IC procedures and their applications, particularly in small samples.

Almost all of the IC procedures proposed in the literature were actually designed for solving a particular type of model selection problem. For example, AIC grew out of Akaike's (1973, 1974) research on selecting the best order of an autoregressive process, while Mallows (1973) introduced a criterion (MCp) for selection of regressors in a model. All of the existing IC differ from each other only by their penalty function (see Fox, 1995). Therefore, by introducing a new penalty function, researchers can easily arrive at a new criterion. However, the performance of these

criteria may vary from one model selection problem to another (for example, see Billah and King, 1998a, 2000a, 2000b; Crato and Ray, 1996; Holmes and Hutton, 1989; Kwek, 2000).

Exponential smoothing methods are widely used for univariate time series forecasting procedures in many industrial applications, including production planning, inventory control and production scheduling (Makridakis and Wheelwright, 1989; Gardner, 1985; Brown, 1963; Winters, 1960). Many studies have shown that although these methods are extremely simple and easy to apply, they are as accurate as more complicated and statistically sophisticated alternatives for forecasting various types of time series (Makridakis et al., 1982; Makridakis et al., 1993; Armstrong and Collopy, 1992, 1993; Fildes et al., 1998; Makridakis and Hibon, 2000). Unfortunately, the development of IC based model selection procedures for exponential smoothing algorithms has not been popular because there has not been a well developed probability modeling framework for this approach in the literature. However, the work of Gardner (1985), Ord et al. (1997) and Hyndman et al. (2000) towards this framework provides an important foundation for developing IC based model selection procedures. This chapter aims to review the literature on exponential smoothing methods and their corresponding state space models, and also findings from some recent studies on model selection for exponential smoothing methods (models). Further, included in this chapter is the SA algorithm which has gained popularity in the recent years as the optimization method for very complex functions.

The chapter is organized as follows. Section 2.2 gives a review of the literature on some existing IC procedures. This section also discusses the consistency and small sample properties of some existing IC procedures. A brief survey of the SA algorithm is presented in Section 2.3. Exponential smoothing methods and their

corresponding state space models are outlined in Section 2.4. For some selected exponential smoothing methods (models), the literature on ARIMA equivalence and smoothing parameter spaces is also reviewed in this section. Further, this section includes a discussion of the various initialization methods for seed vectors. The accuracy of various forecasting procedures is discussed in Section 2.5, and Section 2.6 contains a survey of model selection in exponential smoothing methods (models). The final section presents concluding remarks.

2.2 A Review of Selected Information Criteria (IC)

In the last three decades, a number of model selection criteria have been proposed in the literature. However, we review only seven criteria, namely, AIC, BIC, HQ, MCp, GCV, FPE and Residual Variance Criterion (RVC) of Theil's (1961), which have been widely used in the literature. Recent reviews of these IC procedures can be found in Fox (1995), Hughes (1997), Hossain (1998) and Kwek (2000). Fox (1995) expressed these criteria in a penalized log-likelihood form (maximized log-likelihood minus a penalty function). The new format of these criteria allows easy comparison between different forms of model selection and also between the various marginal penalties. The above IC procedures are briefly outlined in the following subsections.

2.2.1 Akaike Information Criterion (AIC)

Akaike's information criterion (AIC) is a well known and most widely used procedure developed by Akaike (1973). This criterion gives a measure of distance between the estimated model and the true data generating process through examination of Kullback-Leibler's (KL) (1951) information or alternatively, the mean expected log-

likelihood of the model under consideration. It grew out of Akaike's (1973, 1974) research on selecting the best order of an autoregressive (AR) process. Although different forms by different authors are available, the penalized log-likelihood form given by Fox (1995), with the penalty term being q, the number of free parameters included in the model under consideration, is given by

$$\text{AIC} = \ln L(\widehat{\theta}) - q, \tag{2.2.1}$$

where $\widehat{\theta}$ is the estimated parameter vector in the model and $\log L(\widehat{\theta})$ is the maximized log-likelihood. The chosen model is the one that maximizes AIC within the set of models under consideration.

For a linear regression model, using AIC is equivalent to choosing the model that minimizes the following expression:

$$\text{AIC} = \ln(\widehat{\sigma}_1^2) + \frac{2q}{n}, \tag{2.2.2}$$

where $\widehat{\sigma}_1^2$ is the maximum likelihood estimate (MLE) of the error variance for the model with q parameters.

For estimating the order of ARMA(p^*,q^*) models, where p^* and q^* are the orders of the AR and moving average (MA) components, respectively, AIC can take the following form:

$$\text{AIC}(p^*, q^*) = \ln(\widehat{\sigma}_2^2) + \frac{2(p^* + q^*)}{n}, \tag{2.2.3}$$

where $\widehat{\sigma}_2^2$ is the corresponding MLE of the error variance.

Since the development of IC procedures, a significant amount of research has been undertaken concerning the properties of these criteria. Among others, the consistency property has been discussed by many researchers. Consistency refers to the ability of a model selection criterion to select a finite, fixed true model with certainty asymptotically, assuming that the true model is contained in the

competitive set. The meaning of a finite model is that the model contains a finite number of parameters. The AIC procedure is not consistent in this sense (see for example, Akaike, 1979; Shibata, 1986; Hannan, 1982; Bethel, 1984; Hampel et al., 1986; Koehler and Murphree, 1988; Atkinson, 1980).

AIC was developed to measure the goodness of fit and parsimony of a model. The goodness of fit is measured by mean expected log-likelihood, i.e., the larger the mean expected log-likelihood, the better the fit of the model. The parsimony of a model is related to its simplicity, meaning that the smaller the number of regressors, the more parsimonious the model is. AIC performs better at measuring the goodness of fit than the parsimony, because recent studies have shown that AIC has a tendency to overfit the data (see Hurvich and Tsai, 1989; Mills and Prasad, 1992). The study of Hurvich and Tsai (1989) showed that in small samples, AIC tends to provide a negatively biased estimate of the KL information. The authors argued that this under estimation is due to the fact that AIC tends to overfit when the sample size is small relative to the dimension of the model.

2.2.2 Schwarz Criterion (BIC)

Schwarz (1978) provides a criterion which is a large sample approximation for the posterior odds ratio of the models considered. This is a Bayesian solution to Akaike's criterion, and hence, it is typically referred to as the Bayesian information criterion (BIC). BIC is asymptotically equivalent to a Bayesian procedure that selects the model with the highest posterior probability. Another Bayesian information criterion that is based on KL divergence was proposed by Sawa (1978). This criterion is independent of the existence of a true model. Leamer (1978) proposed a similar criterion to BIC. However, the BIC proposed by Schwarz provides the most general asymptotic properties.

As for AIC, BIC is based on choosing the model with the largest value of

$$\text{BIC} = \ln L(\widehat{\theta}) - \frac{q}{2}\log(n), \tag{2.2.4}$$

where n is the sample size. For a wide-ranging discussion on this, see Akaike (1981).

For the linear regression model, BIC is equivalent to the maximization of

$$\text{BIC} = \ln L(\widehat{\sigma}_1^2) - \frac{q}{2}\log(n). \tag{2.2.5}$$

According to the results of Hannan and Quinn (1979), this criterion (BIC) has the property of strong consistency (see also Schwarz, 1978; Rissanen, 1978; Quinn, 1980; Hannan, 1980; Hannan and Kavalieris, 1984; Wei, 1992; Hurvich and Tsai, 1989, 1990). In a simulation study, Sneek (1984) found that BIC outperformed Akaike's information criterion in selecting the correct order of an ARMA process, especially if there was a large distance between competing models. BIC selects more parsimonious models than AIC if the number of observations is large. This is because, for increased n, BIC penalizes additional parameters much more than AIC.

Jeffreys (1967) proposed a criterion for model discrimination and Stone (1979) showed that Schwarz's (1978) BIC was a special case of Jeffrey's work. The study of Geweke and Meese (1981) showed that AIC tends to overfit when identifying an ARIMA model, however, asymptotically BIC correctly selects the true model. Kohn (1983) showed that when choosing from a large class of models, BIC consistently chooses a smaller (minimal dimension) model. Rissanen (1978) also proposed a Bayesian IC which is the same as Schwarz's BIC. For selecting ARMA models, Rissanen's BIC has been extended by both Hannan (1980) and Hannan and Rissanen (1982).

2.2.3 Hannan and Quinn's Criterion (HQ)

Using the theorem of the law of iterated logarithms given by Heyde and Scott (1973) and Heyde (1974), Hannan and Quinn (1979) derived a criterion for AR models which is known as the Hannan and Quinn (HQ) criterion. The penalized maximized log-likelihood form of this criterion (Fox, 1995) is based on choosing the model with the largest value of

$$\mathrm{HQ} = \log L(\hat{\theta}) - q\log(\log n). \tag{2.2.6}$$

Fox noted that HQ shares a property of both AIC and BIC in that its marginal penalty is constant as q increases, for a fixed sample size. The studies of Hannan and Quinn (1979), Nishii (1988) and Atkinson (1981) showed that this criterion is consistent.

For selecting between regression models, this criterion is equivalent to:

$$\min_{q}\left[\ln(\hat{\sigma}_1^2) + \frac{2q\ln(\ln n)}{n}\right]. \tag{2.2.7}$$

This criterion was also extended by Hannan (1980) for selecting between ARMA(p^*,q^*) models. According to this extended criterion, the order of the ARMA(p^*,q^*) model can be selected by minimizing:

$$\ln(\hat{\sigma}_2^2) + \frac{(p^*+q^*)c\ln(\ln n)}{n}, \tag{2.2.8}$$

where c is a constant to be specified. For $c > 2$, (2.2.8) is consistent for the order of the autoregression (see Hannan, 1980).

2.2.4 Mallows' Criterion (MCp)

This criterion was first suggested by Mallows (1964) and has enjoyed much popularity in many sciences. However, compared to AIC, MCp has lost its popularity as

there now exist a number of alternative criteria in the IC literature. Among others, Gorman and Toman (1966) and Mallows (1973) are common references for MCp. Indeed, Mallows proposed this criterion for the selection of regressors. The original statistic, MCp, is as follows:

$$\text{MCp} = \left(\frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\widehat{\sigma}_3^2} \right) - n + 2q, \tag{2.2.9}$$

where y_i is the ith value of the dependent variable, q is the total number of parameters in the model, $\widehat{\sigma}_3^2 = \sum (y_i - \widehat{y}_i)^2/(n - q + 1)$ is the unbiased estimator of the true residual variance and $\widehat{y}_i$ is the fitted y_i.

Fox (1995) expressed this criterion in the following penalized log-likelihood form:

$$\text{MCp} = \ln L(\widehat{\theta}) - \frac{n}{2} \ln \left(1 + \frac{2q}{n - q^{**}} \right), \tag{2.2.10}$$

where q^{**} is the number of free parameters in the smallest model which nests all models under consideration. The model with the highest MCp is chosen. Criteria proposed by Amemiya (1972, 1980) and Akaike (1973) are similar to that of Mallows, although they are derived from somewhat different considerations. Atkinson (1981) and Nishii (1988) showed that MCp is inconsistent.

2.2.5 Generalized Cross-Validation Criterion (GCV)

The Cross-Validation (CV) criterion was first advocated by Schmidt (1971). Other studies on CV can be found in Schmidt (1974, 1975), Allen (1971, 1974) and Stone (1974). Schmidt called CV, the sum of squared predictive errors (SSPE), while Allen called it prediction sum of squares (PRESS). The PRESS criterion was extensively investigated by Stone (1974). As an approximation to CV, Golub et al. (1979) derived the Generalized Cross-Validation (GCV) criterion. When applied to regression models, GCV chooses the model that gives the smallest value of

$$\left(1 - \frac{q}{n}\right)^{-2} \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n}. \tag{2.2.11}$$

In its logarithmic form, GCV is equivalent to:

$$\min_q \left(\ln(\widehat{\sigma}_1^2) - 2\ln(1 - \frac{q}{n}) \right). \tag{2.2.12}$$

One important problem is that, like AIC and MCp, GCV is inconsistent (see Nishii, 1988). More details on the GCV criterion can be found in Golub et al. (1979). The asymptotic equivalence of the CV criterion to AIC was discussed by Stone (1977) and Nishii (1986). The CV proposed by Stone (1974) and Geisser (1974, 1975) is a very general procedure, however, these criteria have not become as popular as AIC.

Fox (1995) derived the penalized log-likelihood form of GCV which is given by,

$$\text{GCV} = \ln L(\widehat{\theta}) + n\log\left(1 - \frac{q}{n}\right). \tag{2.2.13}$$

The model with maximum GCV is chosen as the best model.

2.2.6 Finite Prediction Error Criterion (FPE)

Akaike (1970) developed the following criterion for selecting the order of univariate AR models. The criterion uses prediction errors of an independent realization of the fitted model. The value:

$$\text{FPE} = \left(\frac{n+q}{n-q}\right)\widehat{\sigma}_1^2 \tag{2.2.14}$$

is computed for each order, where q is the order of the AR process. Then the order with the smallest value of FPE is selected as the best.

An alternative procedure was also derived by Amemiya (1972, 1980). This criterion is known as the Prediction Criterion (PC) and is the same as FPE. Fox (1995) expressed FPE as follows:

$$\text{FPE} = \ln L(\widehat{\theta}) - \frac{n}{2}\log(n+q) + \frac{n}{2}\log(n-q). \tag{2.2.15}$$

Nishii (1988) showed that this criterion is inconsistent.

2.2.7 Theil's Residual Variance Criterion (RVC)

Historically, the coefficient of multiple determination (R^2) probably was the first model selection procedure used in econometrics. R^2 measures the part of the error variance that is explained by the regression model under consideration. In the context of the linear regression model, R^2 is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \tag{2.2.16}$$

where $\bar{y}_i$ is the mean of y_is. This procedure involves the choice of a linear regression model that gives the best within sample predictions or equivalently, the largest R^2 value. However, the value of R^2 always increases with any addition of an explanatory variable. Therefore, R^2 cannot be used as a criterion for choosing between competing models. To handle this problem, Theil (1961) proposed the adjusted coefficient of determination denoted $\bar{R}^2$ that takes into account the number of estimated parameters. The $\bar{R}^2$ criterion is given by

$$\bar{R}^2 = 1 - R^2\left(\frac{n}{n-q+1}\right), \tag{2.2.17}$$

where n is the sample size and q is the total number of parameters included in the model. Theil's adjusted $\bar{R}^2$ is equivalent to:

$$\min_q \left[\widehat{\sigma}_4^2\left(\frac{n}{n-q+1}\right)\right], \tag{2.2.18}$$

where $\widehat{\sigma}_4^2 = \sum(y_i - \widehat{y}_i)^2/(n-q+1)$ is the ordinary least squares (OLS) error variance estimate of the respective model.

It has been shown by Theil (1971) that a decision rule which favors the model with the highest $\bar{R}^2$ will result on average in the correct choice of the model. Pearson (1974) showed that the use of $\bar{R}^2$ for choosing between nested models is misleading, because this criterion is equivalent to conducting a t test at the 25% significance

level, which is much larger than a normally used significance level such as 5%. However, the use of $\bar{R}^2$ for selecting between non-nested models has been justified by Pesaran (1974). Unfortunately, this criterion is inconsistent (Nishii, 1988).

Fox (1995) expressed Theil's (1961) $\bar{R}^2$ in the following penalized maximized log-likelihood form and called it the residual variance criterion.

$$\text{RVC} = \ln L(\widehat{\theta}) + \frac{1}{2} n \ln(n - q + 1). \tag{2.2.19}$$

2.2.8 Finite Sample Performance of IC

From the above discussion, clearly the large sample properties of IC procedures have been investigated by many researchers. The next question that arises is how these procedures perform in small samples. In investigating this, a small number of finite sample performance studies of the IC procedures have been conducted. The following is a short survey of results of some recent studies on the finite sample performance of IC procedures.

Using Monte Carlo methods, Lütkelpohl (1984) compared the forecasting performance of AIC and BIC for selecting models from AR and MA processes. The results of this study showed that AIC performs better than BIC in small samples for one-step ahead forecasts, but worse for five-step ahead forecasts. However, in large samples, BIC performs better than AIC. In another Monte Carlo experiment, Lütkelpohl (1985) compared the performance of AIC, BIC, HQ, FPE, CAT (Parzen, 1974), Shibata's criterion (Shibata, 1980) to the LR test (Hannan, 1970) in selecting the order of vector autoregressive (VAR) processes. The results of this study showed that BIC performs best, followed by HQ and LR tests, with CAT being the worst performer. Another finding of this paper was that BIC and HQ consistently estimate the order of VAR processes.

Meese and Geweke (1984) compared the forecasting performance of AIC, BIC,

Sawa's information criterion (SIC) (Sawa, 1978) and MCp for selecting the order of AR processes. The authors considered three forecasting accuracy measures, namely, mean error, mean absolute error and mean squared error. They found that AIC performs best for every measure of forecast error, except mean error. Engle and Brown (1986) empirically compared the forecast accuracy resulting from a variety of model selection procedures and found that selection criteria such as BIC, which most heavily penalizes overparameterized models, performs the best. In order to compare the performance of BIC, AIC and bias corrected AIC (AICc), Crato and Ray (1996) conducted a large-scale simulation study and concluded that for pure fractional noise, BIC performs better than AIC and AICc. However, AIC and AICc perform better than BIC for mixed fractionally integrated autoregressive moving average (ARFIMA) models.

Mills and Prasad (1992) investigated the performance of AIC, BIC, AICc, RVC or $\overline{R}^2$, BEC (Geweke and Meese, 1981), MDL (Rissanen, 1987, 1988) and PMDL (Rissanen, 1988) for selecting the order of AR processes and also for variable selection in the linear regression model. The evaluation was based on the number of times the correct model was chosen and out of sample forecast error. This study showed that AICc performs well for small sample sizes, but any advantages from using AICc diminishes as the sample size increases. In general, BIC and BEC were found to be more reliable and better than the other criteria and never considerably worse. Considering the theoretical justification and wide applicability of BIC and on the basis of the results of the study, Mills and Prasad (1992) recommended that BIC should be the first choice of applied researchers.

In the context of regression and AR time series models, Hurvich and Tsai (1991) compared the performance of AIC, BIC and AICc in terms of bias and concluded that for a very small sample size, AICc performs significantly better than BIC and

AIC, but it performs marginally better than BIC and AIC for moderately small sample sizes. In another study, Hurvich and Tsai (1990) investigated the coverage rate of confidence intervals in small samples for models selected by BIC and AIC and found that AIC performs better than BIC. The coverage rate is defined as the proportion of times the true parameter is contained in the confidence interval. Also, Hurvich and Tsai (1993) extended AICc to VAR(p^*) models, where p^* is the order of VAR model. Note that a VAR(p^*) model contains many more parameters than a univariate AR(p^*) model. The authors conducted a simulation study and found that AICc has superior bias properties and performs much better than AIC. The relationship between AIC and AICc was also investigated by Hurvich and Tsai (1993).

Using a Monte Carlo study, Holmes and Hutton (1989) compared the small sample performance of AIC, BIC, Hannan's criterion (HC) (Hannan, 1981), PC criterion (Amemiya, 1980) and Theil's (1961) $\overline{R}^2$ criterion in the context of the linear regression model. The results of the study showed that in general, the $\overline{R}^2$ criterion performs better than the other criteria when the true relationship between independent and dependent variables is weak. For a strong relationship, all of these criteria have a high probability of choosing the correct model and the probability of correct selection increases as the sample size becomes larger. However, $\overline{R}^2$ performs worst and generally, BIC is the best in this case. In the context of the linear regression model with AR(1) and MA(1) errors, a small sample Monte Carlo study by Grose and King (1994) showed that the estimated optimal penalties as well as the marginal log-likelihood based IC procedures result in improved probabilities of correct selection.

In order to investigate the performance of AICc, BIC and HQ in selecting stochastic models, Schmidt and Tschernig (1993) conducted a simulation study

and found that AICc performs best followed by BIC, and HQ is the worst. Wei (1992) compared the performance of AIC, BIC, MCp, FPE, the predictive least squares (PLS) principle and the Fischer information criterion (FIC) in a simulation study involving stochastic models and found that AIC and FPE are equivalent in small samples.

Hughes (1997) conducted a Monte Carlo study to investigate the performance of one sided AIC for selecting models for economic data and concluded that there is no reason for applied researchers to uniformly favor consistent criteria, for example, BIC. In a large simulation study involving autoregressive conditional heteroscedastic (ARCH) models and generalized ARCH (GARCH) models, Kwek (2000) compared the performance of penalized conditional log-likelihood based IC procedures (AIC, BIC, HQ, MCp, GCV, RVC and FPE) and found that in small samples, RVC is the best criterion and BIC is the worst. However, as the sample size becomes larger, RVC loses its efficacy to AIC and the latter becomes comparatively a better criterion.

Alternative methods of numerically estimating penalty value(s) using simulations can also be found in the literature. For model selection of an AR time series, Chen et al. (1993) proposed a resampling technique. For estimating penalties for a special case of a non-Gaussian time series with an AR random component, Grunwald and Hyndman (1998) proposed a parametric bootstrap method. King et al. (1995) introduced a procedure for numerically estimating penalty values by controlling probabilities of correct model selection. Following King et al. (1995), Hossain (1998) proposed an empirical based information criterion called CIC and compared the performance of AIC, BIC and CIC for selecting between Box-Cox transformation models and found that CIC performs better than the existing IC procedures considered in his study. By using Monte Carlo methods, Kwek (2000)

also proposed a small sample optimal model selection procedure, which performed better than all existing information criteria considered in her study. Rahman et al. (1998) and Rahman and King (1997) also proposed other forms of penalty functions where the functions consist of composite variables of n and q.

In the next section, we will introduce a global optimization algorithm called SA which works well even to optimize very complex functions such as functions with ridges and plateaux. More specifically, we wish to investigate whether this algorithm can be used to estimate penalty values to improve the small sample model selection properties.

2.3 Simulated Annealing (SA)

The basic algorithm for SA was introduced by Metropolis et al. (1953) who used it to simulate a collection of atoms at a given temperature. Kirkpatrick et al. (1983) were the first to show how Metropolis et al.'s model for simulating the annealing of solids could be used for optimization problems. The minimization of the objective function corresponds to the energy state of the solid. Therefore, the name of the algorithm is drawn from an analogy between solving an optimization problem and simulating the annealing of a solid.

Many econometric methods, for example, the maximum likelihood method, the generalized method of moments and nonlinear least squares, depend upon optimization to estimate parameters in the model. However, almost all conventional algorithms sometimes fail to estimate the optimum value of parameters. Conventional algorithms, such as Newton-Raphson, attempt to move up hill in an iterative manner. More specifically, starting from a point, these algorithms determine the best direction and step length to head up hill. Popular statistical packages such as SAS, RATS and TSP use these algorithms. Reviews on these packages can be found

in Judge et al. (1985) and Press et al. (1986). Many conventional algorithms assume that the function to be optimized is approximately quadratic. Unfortunately, some functions violate this assumption. Another assumption that is very common to classical algorithms is that the function has one optimum, and hence, any local optimum is also the global optimum. Also, the conventional algorithms may have difficulties with ridges and plateaux. When such problems occur, researchers often attempt to solve them by trying different starting values (see Cramer, 1986; p.72 and Finch et al., 1989). Even when these algorithms do converge, they may converge to a local rather than the global optimum. Interestingly, the SA algorithm, which assumes very little about the function, can tackle the optimization problem very efficiently (see Corana et al, 1987; Goffe et al., 1994). The advantage of this algorithm is that it is explicitly designed for functions with multiple optima and also works well for functions with ridges and plateaux. SA explores the function's entire surface and tries to optimize the function while moving both up hill and down hill. Therefore, SA is much more robust than classical algorithms.

An early SA algorithm was introduced in combinatorial optimization and is known as combinatorial SA. This SA algorithm has been used successfully in computer and circuit design (Kirkpatrick et al., 1983 and Wong et al., 1988), neural networks (Wasserman and Schwartz, 1988), pollution control (Derwent, 1988), reconstruction of pollycrystalline structures (Telly et al., 1987) and image processing (Carnevali et al., 1985). Corana et al. (1987) derived a new SA algorithm for optimization of functions of continuous variables from the SA algorithm introduced in combinatorial optimization. This new SA algorithm has been found to be more reliable, being nearly always able to find the optimum, or at least a point very close to it. Other SA algorithms proposed in the literature are as follows: adaptive random search (Pronzato et al., 1984), fast SA (Szu and Hartly, 1987), down hill

simplex with annealing (Vetterling et al., 1994) and direct search SA (Ali et al., 1997). However, the algorithm introduced by Corana et al. (1987) appears to be the best with respect to the combination of ease of use and robustness. Goffe et al. (1994) compared the Corana et al. (1987) implementation of SA to conventional algorithms on four econometric models. Compared to the three conventional algorithms, SA was found to have several advantages. The most important advantage is that it can optimize functions with which conventional algorithms have extreme difficulty or simply cannot optimize at all. This algorithm can also be used as a diagnostic tool to understand how conventional algorithms fail. Further, it can step around regions in the parameter space for which the function does not exist.

The SA algorithm is a stochastic optimization algorithm, which borrows ideas from statistical physics. Although it was first introduced by Metropolis et al. (1953) in a pioneering paper, it was made operational by Kirkpatrick et al. (1983). Since then, the SA algorithm has become a popular method for a wide class of optimization problems. Let us assume that $\vartheta = (\vartheta_1, \vartheta_2, \cdots, \vartheta_N)'$ is a vector of parameters to be estimated and $f(\vartheta)$ is a bounded function to be maximized. Let V be the $N \times 1$ step length vector for ϑ and T be the temperature. The algorithm needs starting values for ϑ, V and T which are assumed to be ϑ_0, V_0 and T_0, respectively. The algorithm computes the value of $f(\vartheta)$ at ϑ_0 and also sets $\vartheta_{opt} = \vartheta_0$ and $f_{opt} = f(\vartheta_0)$, where *opt* stands for "optimum". The explicit description of the algorithm (for maximizing a function) is as follows.

The required steps are:

1. Using the following equation, the algorithm generates a random point ϑ' for ϑ by changing the ith element of ϑ as follows:

$$\vartheta_i' = \vartheta_i + u^* v_i, \tag{2.3.1}$$

where u^* is a random number generated in the range [-1,1] by a pseudo random number generator and v_i is the ith component of the step vector V.

2. The function value f' is computed at this new point ϑ'_i. If this new function value is greater than the function value f_{opt}, ϑ' is accepted and ϑ is replaced by ϑ' i.e., $\vartheta = \vartheta'$. At this stage f' and ϑ' are recorded as *opt* values.

3. If f' is less than or equal to f_{opt}, the acceptance of the new point is decided by the Metropolis criterion. This works as follows: the value

$$p_r = e^{(f'-f_{opt})/T} \tag{2.3.2}$$

is calculated and then compared with a value p_u, which is randomly generated from a uniform distribution ranging from 0 to 1. If p_r is greater than p_u, the new point is accepted and ϑ is updated by ϑ' (in this case the algorithm moves down hill). Otherwise, the new point ϑ' is rejected. Two factors, large differences in function values and lower temperature, decrease the probability of a down hill move.

4. In order to accept 50% of all moves, the step length vector is adjusted after N_s steps through all elements of the vector ϑ_0. The aim of doing this is to sample the function widely. If more than 60 percent of points are accepted for ϑ_i, then the relevant component of V is enlarged by the factor $1+2.5c_i(m_i/N_s-0.6)$, where m_i is the number of points accepted and c_i is the ith element of the vector that controls step length adjustment. If less than 40 percent of points are accepted, the component is declined by $1+2.5c_i(0.4-m_i/N_s)$. Otherwise, the component remains unchanged.

5. After N_T times through the steps 1 to 4, the temperature T is reduced and

the new temperature is given by

$$T' = r_T T, \tag{2.3.3}$$

where r_T lies between 0 to 1.

6. From the end of the last four temperature reductions, the largest function values are recorded and are compared. The algorithm terminates if all of these differences are less than ϵ, a very small quantity.

The SA algorithm has a number of potential advantages over classical optimization methods. First, the SA algorithm can escape from local maxima by moving both up hill and down hill. Also, the function to be optimized does not need to be approximately quadratic and it does not even need to be differentiable (see Corana et al., 1987). Second, the SA algorithm can snuggle up to a corner for functions that do not exit in some region, and hence, the algorithm can identify corner solutions. Another advantage of this algorithm is that it provides valuable information about the function through the step length vector. A large element of V indicates that the function is flat in that parameter. The most important advantage of SA is that it can properly optimize functions that are very complex and impossible to optimize (see Goffe et al., 1994). The only drawback of SA is that the required computer power can be high. However, this problem has already disappeared with the availability of high levels of computer power. Thus, SA is an attractive optimization algorithm for difficult functions. In this study, we implement SA to estimate penalty functions (which is a step function) for small sample model selection and forecasting.

2.4 Exponential Smoothing Methods/Models

Exponential smoothing methods have been used extensively in industry. For example, typical applications include production planning, production scheduling and inventory control. The related literature can be found in Brown (1959, 1963, 1967), Holt et al. (1960), Winters (1960), Gardner (1985), Makridakis and Wheelwright (1989), among others. Exponential smoothing methods are very popular in short-range forecasting because of several practical considerations. Model formulations for exponential smoothing methods are relatively simple. These methods have been found by many studies to be as accurate as more complex and statistically sophisticated alternatives (Groff (1973), Chatfield (1978), Makridakis and Hibon (1979), Makridakis et al. (1982), Makridakis et al. (1993), Fildes et al. (1998), Makridakis and Hibon (2000)). Furthermore, exponential smoothing methods are easy to program, robust, require a minimum of historical data, and the cost of running them on the computer is the smallest of all available alternatives. Exponential smoothing methods, their corresponding state space models along with various properties and initialization methods, are discussed in the following subsections.

2.4.1 Exponential Smoothing Methods

There are many versions of exponential smoothing methods in the literature. Pegels (1969) has given a simple but very useful classification which has been extended by Gardner (1985). Following Pegels (1969) and Gardner (1985), exponential smoothing methods can be summarized in a two-way classification. Each exponential smoothing method has or does not have a trend component and/or seasonal component, as shown in Table 2.1.

Cell NN describes the simple exponential smoothing (SES) method, cell AN describes Holt's exponential smoothing method and the additive and multiplicative

Table 2.1: Major categories of exponential smoothing methods.

Trend Component	Seasonal Component		
	None (N)	Additive (A)	Multiplicative (M)
None (N)	NN	NA	NM
Additive (A)	AN	AA	AM
Multiplicative (M)	MN	MA	MM
Damped (D)	DN	DA	DM

Holt-Winters' exponential smoothing methods are given by cells AA and AM, respectively. Formulae for calculations and forecasting using the classification in the above table are as follows.

Following Makridakis, Wheelwright and Hyndman (1998), each of the above exponential smoothing methods can be written as:

$$l_t = \alpha_1 \mathcal{P}_t + (1-\alpha_1)\mathcal{Q}_t, \tag{2.4.1}$$

$$b_t = \alpha_2 \mathcal{R}_t + (\phi-\alpha_2)b_{t-1}, \tag{2.4.2}$$

$$c_t = \alpha_3 \mathcal{T}_t + (1-\alpha_3)c_{t-s}, \tag{2.4.3}$$

where α_1, α_2 and α_3 are smoothing parameters for local level l_t, local trend b_t and seasonal factors c_t respectively, ϕ is the damping parameter, s is the number of seasons in a year and $\mathcal{P}_t$, $\mathcal{Q}_t$, $\mathcal{R}_t$ and $\mathcal{T}_t$ vary according to which of the cells the method belongs. Following Hyndman et al. (2000), the values of $\mathcal{P}_t$, $\mathcal{Q}_t$, $\mathcal{R}_t$ and $\mathcal{T}_t$ for different cells of Table 2.1 and the formulae for computing h-step ahead forecasts $y_t(h)$ at time t are as follows.

2.4.1.1 Methods with no Trend

These are the simplest of the exponential smoothing methods. These methods are appropriate if there is no increasing or decreasing pattern in the data. However, there might be seasonality in the data and the seasonal component can be modeled

either as an additive or multiplicative factor. Additive seasonality is appropriate when the seasonal variation is constant around the mean and the multiplicative component is appropriate when the magnitude of the seasonal variation is proportional to the local mean. The exponential smoothing methods with no trend are as follows:

NN: *Simple exponential smoothing:*

$$\begin{aligned} \mathcal{P}_t &= y_t, \quad \mathcal{Q}_t = l_{t-1}, \\ \phi &= 1, \quad y_t(h) = l_t. \end{aligned} \tag{2.4.4}$$

NA: *Simple exponential smoothing with additive seasonality:*

$$\begin{aligned} \mathcal{P}_t &= y_t - c_{t-s}, \quad \mathcal{Q}_t = l_{t-1}, \quad \mathcal{T}_t = y_t - \mathcal{Q}_t, \\ \phi &= 1, \quad y_t(h) = l_t + c_{t+h-s}. \end{aligned} \tag{2.4.5}$$

NM: *Simple exponential smoothing with multiplicative seasonality:*

$$\begin{aligned} \mathcal{P}_t &= y_t/c_{t-s}, \quad \mathcal{Q}_t = l_{t-1}, \quad \mathcal{T}_t = y_t/\mathcal{Q}_t, \\ \phi &= 1, \quad y_t(h) = l_t c_{t+h-s}. \end{aligned} \tag{2.4.6}$$

2.4.1.2 Methods with Linear Trend

Holt (1957) extended the simple exponential smoothing method to time series that have a discernible trend. The algorithm considers different smoothing parameters for level and trend and also assumes that the trend is evolving and locally linear. Winters (1960) extended the Holt's method by including an additional equation to smooth the seasonal factors in the data, explicitly using a third smoothing parameter. This algorithm is known as the Holt-Winters' method. The seasonal component can be included in the algorithm either as an additive or a multiplicative factor. The equations of these algorithms are given below.

AN: *Holt's linear exponential smoothing:*

$$\begin{aligned} \mathcal{P}_t &= y_t, \quad \mathcal{Q}_t = l_{t-1} + b_{t-1}, \quad \mathcal{R}_t = l_t - l_{t-1}, \\ \phi &= 1, \quad y_t(h) = l_t + hb_t. \end{aligned} \tag{2.4.7}$$

AA: *Additive Holt-Winters' method:*

$$\begin{aligned} \mathcal{P}_t &= y_t - c_{t-s}, \quad \mathcal{Q}_t = l_{t-1} + b_{t-1}, \quad \mathcal{R}_t = l_t - l_{t-1}, \\ \mathcal{T}_t &= y_t - \mathcal{Q}_t, \quad \phi = 1, \quad y_t(h) = l_t + hb_t + c_{t+h-s}. \end{aligned} \tag{2.4.8}$$

AM: *Multiplicative Holt-Winters' method:*

$$\begin{aligned} \mathcal{P}_t &= y_t/c_{t-s}, \quad \mathcal{Q}_t = l_{t-1} + b_{t-1}, \quad \mathcal{R}_t = l_t - l_{t-1}, \\ \mathcal{T}_t &= y_t/\mathcal{Q}_t, \quad \phi = 1, \quad y_t(h) = (l_t + hb_t)c_{t+h-s}. \end{aligned} \tag{2.4.9}$$

2.4.1.3 Methods with Multiplicative Trend

Holt's algorithm as well as additive and multiplicative versions of Holt-Winters' methods assume that the underlying trend is linear. However, this may not be true for all real time series data (Gardner, 1985). The exponential smoothing algorithms with multiplicative (nonlinear) trend were developed to incorporate exponential trend in the series into the smoothing equations (Pegels, 1969; Gardner, 1985). Different versions of exponential smoothing methods with multiplicative trend are outlined as follows:

MN: *Multiplicative trend with no seasonality:*

$$\begin{aligned} \mathcal{P}_t &= y_t, \quad \mathcal{Q}_t = l_{t-1} + b_{t-1}, \quad \mathcal{R}_t = l_t/l_{t-1}, \\ \phi &= 1, \quad y_t(h) = l_t b_t^h. \end{aligned} \tag{2.4.10}$$

MA: *Multiplicative trend with additive seasonality:*

$$\begin{aligned} \mathcal{P}_t &= y_t - c_{t-s}, \quad \mathcal{Q}_t = l_{t-1} + b_{t-1}, \quad \mathcal{R}_t = l_t/l_{t-1}, \\ \mathcal{T}_t &= y_t - \mathcal{Q}_t, \quad \phi = 1, \quad y_t(h) = l_t b_t^h + c_{t+h-s}. \end{aligned} \tag{2.4.11}$$

MM: *Multiplicative trend with multiplicative seasonality:*

$$\begin{aligned} \mathcal{P}_t &= y_t/c_{t-s}, \quad \mathcal{Q}_t = l_{t-1} + b_{t-1}, \quad \mathcal{R}_t = l_t/l_{t-1}, \\ \mathcal{T}_t &= y_t/\mathcal{Q}_t, \quad \phi = 1, \quad y_t(h) = l_t b_t^h c_{t+h-s}. \end{aligned} \tag{2.4.12}$$

2.4.1.4 Methods with Damped Trend

Damped trend exponential smoothing methods are appropriate when there is trend in the time series data. In additive (linear) trend methods, the growth rate estimated from the end of the time series data are used for forecasting at all forecast

horizons. However, this may give unrealistic forecasts because the estimated trend is local, and therefore, should not be used without any modification. A number of researchers have argued that a generalization of the linear trend algorithm was possible by introducing an extra parameter to control the trend (see Gilchrist, 1976; Roberts, 1982 and Gardner 1985). Gardner and McKenzie (1985, 1988 and 1989) used an autoregressive damping parameter to control the rate of trend extrapolation by incorporating a new parameter ϕ into the additive trend methods, which are known as damped trend methods. The damped trend methods are given as follows:

DN: *Holt's damped trend method:*

$$\begin{array}{lll} \mathcal{P}_t = y_t, & \mathcal{Q}_t = l_{t-1} + b_{t-1}, & \mathcal{R}_t = l_t - l_{t-1}, \\ \alpha_2 < \phi < 1, & y_t(h) = l_t + b_t \sum_{i=0}^{h-1} \phi^i. & \end{array} \tag{2.4.13}$$

DA: *Additive Holt-Winters' damped trend method:*

$$\begin{array}{ll} \mathcal{P}_t = y_t - c_{t-s}, & \mathcal{Q}_t = l_{t-1} + b_{t-1}, \\ \mathcal{R}_t = l_t - l_{t-1}, & \mathcal{T}_t = y_t - \mathcal{Q}_t, \\ \alpha_2 < \phi < 1, & y_t(h) = l_t + b_t \sum_{i=0}^{h-1} \phi^i + c_{t+h-s}. \end{array} \tag{2.4.14}$$

DM: *Multiplicative Holt-Winters' damped trend method:*

$$\begin{array}{ll} \mathcal{P}_t = y_t / c_{t-s}, & \mathcal{Q}_t = l_{t-1} + b_{t-1}, \\ \mathcal{R}_t = l_t - l_{t-1}, & \mathcal{T}_t = y_t / \mathcal{Q}_t, \\ \alpha_2 < \phi < 1, & y_t(h) = (l_t + b_t \sum_{i=0}^{h-1} \phi^i) c_{t+h-s}. \end{array} \tag{2.4.15}$$

When $\phi = 1$, the damped trend methods and the linear trend methods are the same. When $0 < \phi < 1$, the trend is damped and for $\phi > 1$, the trend is exponential and may result in an explosive and unstable situation. Therefore, the value of ϕ is restricted to the interval [0,1]. The damped trend equation dampens the trend as the length of the forecast horizon increases. From equation (2.4.13), one can see that the h-step ahead forecast is $y_t(h) = l_t + b_t \sum_{i=0}^{h-1} \phi^i$. For each additional future time period, the trend is dampened by a factor of ϕ. The formula for damped trend

proposed by Gardner (1985) ($y_t(h) = l_t + b_t \sum_{i=1}^{h} \phi^i$) differs from that of Hyndman et al. (2000) by a factor ϕ. Hyndman et al. begin dampening the trend from the two-step ahead forecast as shown in equation (2.4.13).

The difference between the smoothing methods presented in equations (2.4.4) to (2.4.15) and those given by Makridakis et al. (1998) is that these authors did not consider the damped trend methods. Further, in the above equations, l_t is replaced by $\mathcal{Q}_t$ in the $\mathcal{T}_t$ equations. The effect of this is that when the seasonal component is updated, the level l_{t-1} and the growth rate b_{t-1} are used from the previous time period rather than the newly revised level l_t from the current time period. The alternative equations allow the exponential smoothing methods to be expressed in state space form, which will be discussed in the next section. Also, the equations used in cell AM are not those of the usual Holt-Winters' method. They are equivalent to those used by Ord et al. (1997). The additive seasonal method is not affected by this change, but it does affect the forecasts slightly for the multiplicative seasonal method.

The error correction form (see, Gardner 1985) of the above exponential smoothing methods can also be obtained by writing equations (2.4.1) to (2.4.3) in the following form.

$$l_t = \alpha_1(\mathcal{P}_t - \mathcal{Q}_t) + \mathcal{Q}_t, \tag{2.4.16}$$

$$b_t = \alpha_2(\mathcal{R}_t - b_{t-1}) + \phi b_{t-1}, \tag{2.4.17}$$

$$c_t = \alpha_3(\mathcal{T}_t - c_{t-s}) + c_{t-s}. \tag{2.4.18}$$

The exponential smoothing method with fixed level, fixed trend and fixed seasonal pattern can be obtained by setting $\alpha_1 = 0$, $\alpha_2 = 0$ and $\alpha_3 = 0$, in equations (2.4.16), (2.4.17) and (2.4.18), respectively. Also, $\phi = 1$ in the damped trend method gives the linear trend method.

Each of the exponential smoothing methods discussed above can be expressed in a state space modeling framework. We discuss this in the following subsections.

2.4.2 State Space Models

Since the development of exponential smoothing methods in the 1950s, an appropriate modeling framework incorporating stochastic models, likelihood calculation, prediction intervals and IC based model selection procedures has been needed. Some earlier work towards this framework can be found in Gardner (1985) and Ord et al. (1997). Also Chatfield and Yar (1991), Ord et al. (1997) and Koehler et al. (1999) worked on developing prediction intervals for exponential smoothing methods. Snyder (1985a) and Ord et al. (1997) developed state space models for the linear methods NN, AN and AA, and non-linear method AM, respectively. Following the general approach of Ord et al. (1997), for the remaining methods, Hyndman et al. (2000) derived an equivalent state space formulation with a single source of error. This assists easy calculation of the likelihood and prediction intervals for all models. A single source of error model has some advantages over a multiple source of error model. Firstly, it allows the state space formulation of linear as well as nonlinear cases. Secondly, it helps to express the state equations in a form that coincides with the error correction form of the usual exponential smoothing equations.

For each exponential smoothing method, two models, namely a multiplicative error model and an additive error model, can be obtained. The point forecast for the two models are the same, but their prediction intervals are different. The general framework for state space equations given by Ord et al. (1997) is as follows:

$$y_t = \zeta(\beta_{t-1}) + \kappa(\beta_{t-1})e_t, \tag{2.4.19}$$

$$\beta_t = f(\beta_{t-1}) + g(\beta_{t-1})e_t, \tag{2.4.20}$$

where $\beta_t = (l_t, b_t, c_t, c_{t-1}, \cdots, c_{t-(s-1)})'$ is the state vector and e_t is from an iid$(0,\sigma^2)$ series of disturbances. Equations (2.4.19) and (2.4.20) are known as the observation equation and state equation, respectively.

Assuming $\epsilon_t = \kappa(\beta_{t-1})e_t$ and $\mu_t = \zeta(\beta_{t-1})$, equation (2.4.19) can be written as $y_t = \mu_t + \epsilon_t$. Also, assuming μ_t is the one-step ahead forecast made at time $t-1$ and $\kappa(\beta_{t-1}) = 1$, the additive error model is given by

$$y_t = \mu_t + e_t. \tag{2.4.21}$$

The model with multiplicative errors is given by

$$y_t = \mu_t(1 + e_t). \tag{2.4.22}$$

Comparing equation (2.4.19) and (2.4.22), we have $\kappa(\beta_{t-1}) = \mu_t$ and thus $e_t = \epsilon_t/\mu_t = (y_t - \mu_t)/\mu_t$. Hence, e_t is the relative error for the multiplicative model. All exponential smoothing methods can be written in the state space form (2.4.19) and (2.4.20). The state space equations for each additive error model in the classification are outlined in this section. Note that these equations are not unique. Changing the value of $\kappa(\beta_{t-1})$, one can have many different models which will give identical point forecasts for y_t. By changing the value of $\kappa(\beta_{t-1})$, Archibald (1994), Koehler et al. (1999) have given a number of models for the multiplicative Holt-Winters' model.

2.4.2.1 Models with no Trend

NN: *Simple exponential smoothing (local level) model:*

$$y_t = l_{t-1} + e_t, \tag{2.4.23}$$

$$l_t = l_{t-1} + \alpha_1 e_t, \tag{2.4.24}$$

where e_t is independently and identically distributed with mean 0 and variance σ^2.

NA: *Simple exponential smoothing model with additive seasonality:*

$$y_t = l_{t-1} + c_{t-s} + e_t, \tag{2.4.25}$$

$$l_t = l_{t-1} + \alpha_1 e_t, \tag{2.4.26}$$

$$c_t = c_{t-s} + \alpha_2 e_t. \tag{2.4.27}$$

NM: *Simple exponential smoothing model with multiplicative seasonality:*

$$y_t = l_{t-1} c_{t-s} + e_t, \tag{2.4.28}$$

$$l_t = l_{t-1} + \alpha_1 e_t / c_{t-s}, \tag{2.4.29}$$

$$c_t = c_{t-s} + \alpha_2 e_t / l_{t-1}. \tag{2.4.30}$$

2.4.2.2 Models with Linear Trend

AN: *Holt's linear (local trend) model:*

$$y_t = l_{t-1} + b_{t-1} + e_t, \tag{2.4.31}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t, \tag{2.4.32}$$

$$b_t = b_{t-1} + \alpha_1 \alpha_2 e_t. \tag{2.4.33}$$

AA: *Additive Holt-Winters' model:*

$$y_t = l_{t-1} + b_{t-1} + c_{t-s} + e_t, \tag{2.4.34}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t, \tag{2.4.35}$$

$$b_t = b_{t-1} + \alpha_1 \alpha_2 e_t, \tag{2.4.36}$$

$$c_t = c_{t-s} + \alpha_3 e_t. \tag{2.4.37}$$

AM: *Multiplicative Holt-Winters' model:*

$$y_t = (l_{t-1} + b_{t-1}) c_{t-s} + e_t, \tag{2.4.38}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t / c_{t-s}, \tag{2.4.39}$$

$$b_t = b_{t-1} + \alpha_1\alpha_2 e_t/c_{t-s}, \tag{2.4.40}$$

$$c_t = c_{t-s} + \alpha_3 e_t/(l_{t-1} + b_{t-1}). \tag{2.4.41}$$

2.4.2.3 Models with Multiplicative Trend

MN : *Multiplicative trend model with no seasonality:*

$$y_t = l_{t-1}b_{t-1} + e_t, \tag{2.4.42}$$

$$l_t = l_{t-1}b_{t-1} + \alpha_1 e_t, \tag{2.4.43}$$

$$b_t = b_{t-1} + \alpha_1\alpha_2 e_t/l_{t-1}. \tag{2.4.44}$$

MA : *Multiplicative trend model with additive seasonality:*

$$y_t = l_{t-1}b_{t-1} + c_{t-s} + e_t, \tag{2.4.45}$$

$$l_t = l_{t-1}b_{t-1} + \alpha_1 e_t, \tag{2.4.46}$$

$$b_t = b_{t-1} + \alpha_1\alpha_2 e_t/l_{t-1}, \tag{2.4.47}$$

$$c_t = c_{t-s} + \alpha_3 e_t. \tag{2.4.48}$$

MM: *Multiplicative trend model with multiplicative seasonality:*

$$y_t = l_{t-1}b_{t-1}c_{t-s} + e_t, \tag{2.4.49}$$

$$l_t = l_{t-1}b_{t-1} + \alpha_1 e_t/c_{t-s}, \tag{2.4.50}$$

$$b_t = b_{t-1} + \alpha_1\alpha_2 e_t/c_{t-s}l_{t-1}, \tag{2.4.51}$$

$$c_t = c_{t-s} + \alpha_3 e_t/l_{t-1}b_{t-1}. \tag{2.4.52}$$

2.4.2.4 Damped Trend Models

DN: *Holt's damped trend model:*

$$y_t = l_{t-1} + b_{t-1} + e_t, \tag{2.4.53}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t, \tag{2.4.54}$$

$$b_t = \phi b_{t-1} + \alpha_1\alpha_2 e_t. \tag{2.4.55}$$

DA: *Additive Holt-Winters' damped trend model:*

$$y_t = l_{t-1} + b_{t-1} + c_{t-s} + e_t, \tag{2.4.56}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t, \tag{2.4.57}$$

$$b_t = \phi b_{t-1} + \alpha_1 \alpha_2 e_t, \tag{2.4.58}$$

$$c_t = c_{t-s} + \alpha_3 e_t. \tag{2.4.59}$$

DM: *Multiplicative Holt-Winters' damped trend model:*

$$y_t = (l_{t-1} + b_{t-1})c_{t-s} + e_t, \tag{2.4.60}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t / c_{t-s}, \tag{2.4.61}$$

$$b_t = \phi b_{t-1} + \alpha_1 \alpha_2 e_t / c_{t-s}, \tag{2.4.62}$$

$$c_t = c_{t-s} + \alpha_3 e_t / (l_{t-1} + b_{t-1}). \tag{2.4.63}$$

The difference between the additive error model and the multiplicative error model is only in the observation equation. Multiplicative error models can be obtained by replacing e_t by $\mu_t e_t$ in the above equations. For example, for the local level model, $\mu_t = l_{t-1}$ and hence the equations for the multiplicative version of this, are given by $y_t = l_{t-1}(1 + e_t)$ and $l_t = l_{t-1}(1 + \alpha_1 e_t)$.

The simple exponential smoothing (SES) method, Holt's linear exponential smoothing method, the additive Holt-Winters' method and the additive damped trend method as well as their corresponding state space models discussed above have ARIMA equivalence. The remainder of the smoothing methods (models) have no known ARIMA equivalence. A brief survey of this issue is given in the following subsection.

2.4.3 ARIMA Equivalence of Exponential Smoothing

It can be shown that the forecasts obtained from some exponential smoothing methods (models) are equivalent to those from particular ARIMA processes (see McKen-

zie, 1984, 1986; Gardner and McKenzie, 1985, 1989; Chatfield and Yar, 1991; Yar and Chatfield, 1990). In this subsection, we discuss ARIMA equivalence of exponential smoothing methods and their equivalent state space models. It would make a great deal of sense to choose a forecasting procedure, based on the underlying model. However, such a choice is rarely made in applications. Many exponential smoothing methods as well as their underlying state space models are equivalent to convenient representations of ARIMA processes. The best procedure based on exponential smoothing methods or state space models will provide the minimum forecast error for the equivalent ARIMA process. The ARIMA equivalence of the local level model, Holt's local trend model, Holt-Winters' additive seasonal model and Gardner's damped trend model are discussed below.

2.4.3.1 ARIMA Equivalence of the SES Method

Muth (1960) first proved that the SES method is equivalent to the ARIMA(0,1,1) process:

$$(1-B)y_t = (1-\theta B)e_t, \tag{2.4.64}$$

where B is the backward shift operator, θ is a scalar parameter and e_t is the white noise process that is generally assumed to follow a normal distribution. The equivalence condition is $\theta = 1-\alpha_1$. Since the invertibility condition for an ARIMA(0,1,1) process is $0 < \theta < 1$, it follows that $0 < \alpha_1 < 2$ (see Box and Jenkins, 1970, p.107).

By taking first differences of equation (2.4.23), the following expression can be obtained:

$$y_t - y_{t-1} = e_t - (1-\alpha_1)e_{t-1}, \tag{2.4.65}$$

$$\text{or,} \quad (1-B)y_t = (1-(1-\alpha_1)B)e_t, \tag{2.4.66}$$

$$\text{or,} \quad (1-B)y_t = (1-\theta B)e_t, \tag{2.4.67}$$

where $\theta = 1 - \alpha_1$, and hence, the state space form of the local level model is equivalent to an ARIMA(0,1,1) process. Clearly, the SES method and its state space framework have the same ARIMA equivalence.

2.4.3.2 ARIMA Equivalence of Holt's Method

Harrison (1967) showed that Holt's exponential smoothing method is equivalent to the ARIMA(0,2,2) process:

$$(1-B)^2 y_t = (1 - \theta_1 B - \theta_2 B^2) e_t, \tag{2.4.68}$$

where θ_1 and θ_2 are scalar parameters and the equivalence conditions are $\theta_1 = 2 - \alpha_1 - \alpha_2$ and $\theta_2 = \alpha_1 - 1$. McClain and Thomas (1973) showed that the stable conditions of Holt's linear method are $0 < \alpha_1 < 2$, $\alpha_2 > 0$ and $2\alpha_1 + \alpha_2 < 4$. Stability of Holt's method means the equivalent ARIMA process is invertible.

First differences of equation (2.4.31), gives

$$\begin{aligned} (1-B)y_t &= l_{t-1} + b_{t-1} + e_t - l_{t-2} - b_{t-2} - e_{t-1}, \\ &= b_{t-1} + e_t + (\alpha_1 - 1)e_{t-1}. \end{aligned} \tag{2.4.69}$$

Again taking differences of equation (2.4.69), one can obtain

$$\begin{aligned} (1-B)^2 y_t &= b_{t-1} + e_t + (\alpha_1 - 1)e_{t-1} \\ &\quad -b_{t-2} - e_{t-1} - (\alpha_1 - 1)e_{t-2}, \qquad (2.4.70) \\ &= e_t - (2 - \alpha_1 - \alpha_2)e_{t-1} - (\alpha_1 - 1)e_{t-2}, \\ \text{or,} \quad (1-B)^2 y_t &= (1 - \theta_1 B - \theta_2 B^2)e_t, \qquad (2.4.71) \end{aligned}$$

where the moving average coefficients θ_1 and θ_2 are related to the level and trend smoothing parameters by $\theta_1 = 2 - \alpha_1 - \alpha_2$ and $\theta_2 = \alpha_1 - 1$, respectively. Hence the local trend model (Holt's model) and Holt's algorithm have the same ARIMA equivalence, and therefore, have the same invertible conditions.

When $\alpha_1 = 1$, the local trend model is equivalent to an ARIMA(0,2,1) process and the level at time t is conditional on the observation y_t only. Therefore, the historical information is partially disregarded in level, and if in addition to that, we set $\alpha_2 = 1$, the model is equivalent to an ARIMA(0,2,0) process.

2.4.3.3 ARIMA Equivalence of the Additive Holt-Winters' Method

The additive version of the algorithm is equivalent to a very complex ARIMA$(0, 1, s+1)(0, s, 0)_s$ process, where s is the length of seasonality. The process is given by

$$(1 - B)(1 - B^s)y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_s B^s - \theta_{s+1} B^{s+1})e_t, \quad (2.4.72)$$

where $\theta_i, i = 1, 2, \cdots, s + 1$, are scalar parameters (for more details see McKenzie, 1976; Roberts, 1982; Abraham and Ledolter, 1986).

It can be shown that the additive Holt-Winters' model is equivalent to the following ARIMA process:

$$\begin{aligned}(1 - B)(1 - B^s)y_t = e_t - (1 - \alpha_1 - \alpha_2)e_{t-1} - \Sigma_{i=2}^{s-1}(-\alpha_2)e_{t-i} \\ -(1 - \alpha_2 - \alpha_3)e_{t-s} - (\alpha_1 + \alpha_3 - 1)e_{t-s-1}. \quad (2.4.73)\end{aligned}$$

Comparing (2.4.72) and (2.4.73), the equivalence conditions are:

$$\theta_1 = 1 - \alpha_1 - \alpha_2, \quad (2.4.74)$$

$$\theta_s = 1 - \alpha_2 - \alpha_3, \quad (2.4.75)$$

$$\theta_{s+1} = -1 + \alpha_1 + \alpha_3, \quad (2.4.76)$$

$$\theta_i = -\alpha_2, \quad i = 2, \cdots, s - 1. \quad (2.4.77)$$

Thus, the additive Holt-Winters' algorithm and its equivalent state space model have the same ARIMA equivalence.

2.4.3.4 ARIMA Equivalence of the Additive Damped Trend Model

Holt's damped trend method proposed by Gardner and McKenzie (1985) has at least six equivalent ARIMA processes depending on different values of damping as well as exponential smoothing parameters. A brief discussion of these processes is given below.

If $0 < \phi < 1$, the trend is damped and the equivalent process is the ARIMA(1,1,2) process given by

$$(1-B)(1-\phi B)y_t = \left[1-(1+\phi-\alpha_1-\phi\alpha_1\alpha_2)B-(\alpha_1-1)B^2\right]e_t. \quad (2.4.78)$$

Setting $\alpha_1 = 1$ in (2.4.78), we can also obtain an ARIMA(1,1,1) process and for $\alpha_1 = \alpha_2 = 1$, the process is ARIMA(1,1,0).

A linear trend method can be obtained if $\phi = 1$ and the process is the ARIMA(0,2,2) process:

$$(1-B)^2 y_t = \left[1-(2-\alpha_1-\alpha_1\alpha_2)B-(\alpha_1-1)B^2\right]e_t. \quad (2.4.79)$$

If $\phi = 0$, we get the SES method and the equivalent process is the ARIMA(0,1,1) process:

$$(1-B)y_t = \left[1-(1-\alpha_1)B\right]e_t. \quad (2.4.80)$$

The random walk model, or the ARIMA(0,1,0) process can be obtained by setting $\alpha_1 = 1$ in equation (2.4.80).

Due to the parameter restrictions shown in Gardner and McKenzie (1985), the above ARIMA processes for Holt's damped trend model are only a subset of the possible ARIMA processes of the same order. For example, in (2.4.78), ϕ ranges from 0 to 1, but ϕ can range from -1 to 1 in the general ARIMA(1,1,2) process.

Holt-Winters' multiplicative seasonal method has no known ARIMA equivalence, and the same is true for its corresponding damped trend algorithm. Holt-Winters' additive seasonal algorithm is linear and has ARIMA equivalence. This is also true for the damped trend version of this algorithm. However, the equivalent ARIMA processes are so complex as to be of little practical use (see, Gardner and McKenzie, 1989). It can be shown that Holt's damped trend method and its state space form have the same ARIMA equivalence, as discussed above.

One of the advantages of ARIMA equivalence of different exponential smoothing methods (models) is that it helps the practitioner in choosing the method's (model's) parameter space. Lack of reasonable choice of smoothing parameter(s) of a exponential smoothing method (model) deteriorates the method's (model's) forecasting performance (see Archibald, 1990; Bartolomei and Sweet, 1989). Therefore, care should be taken when selecting the value(s) of the parameter(s). The selection of parameter space, particularly for the SES method (local level model), Holt's algorithm (Holt's model) and Holt-Winters' method (Holt-Winters' model) are discussed in the next subsection.

2.4.4 Parameter Selection

The optimum choice of smoothing parameter(s) for various exponential smoothing methods (models) is important to ensure their efficient forecasting performance. Among those who have worked on the parameter space of exponential smoothing methods are Archibald (1990), Roberts (1982), Montgomery and Johnson (1976), McClain (1974), Chatfield (1978) and Gardner (1985). A brief literature survey of this topic is given below.

2.4.4.1 The Local Level Model (SES method)

In most applications of the SES model, the smoothing parameter value α_1 is assumed to lie between 0 and 1. Also, a more restricted range of 0.1 to 0.3 is not uncommon in practice. It is widely understood that a more complicated model should be considered if the best α_1 value is found to be above 0.3 during the model fitting process (Montgomery and Johnson, 1976).

However, there is no evidence to support a restricted range of parameters as above. A wider range of parameter values are suggested both from theoretical and empirical work. Exponential smoothing is equivalent to a difference equation which is stable in the range $0 < \alpha_1 < 2$. Gardner (1985) pointed out that this important property is frequently overlooked in applications. Among others, Muth (1960) was the first who proved that SES is optimal for the ARIMA(0,1,1) process. The ARIMA(0,1,1) model implies through its inevitability condition that $0 < \alpha_1 < 2$ (Box and Jenkins, 1970, p.107).

In spite of the theoretical justification, a value of α_1 more than one is not usually recommended in most expositions of the SES model. However, in a time series, if the irregular component is dominated by the business cycle, parameter values in this range are quite appropriate. For example, in an upward phase of a business cycle, the next period's forecast can be brought to the current level of the series only if the value of the smoothing parameter is unity. This type of forecast will tend to lag and a value of α_1 above unity is required to eliminate the lag.

In the study by Makridakis et al. (1982), the values of α_1 were estimated mostly above 0.3 during the model fitting process. However, rarely was it bigger than unity. Large parameter values were also found in a limited study conducted by Chatfield (1978). From these studies, it is clear that to guess at values of smoothing parameters can be dangerous. According to Harrison (1967), overestimation of the

optimal values of the parameters are less harmful than underestimation. The values should come from the data (Gardner, 1985).

2.4.4.2 Holt's Model (Method)

For Holt's exponential smoothing model, a parameter value less than 0.3 has been recommended by many researchers. Among those who recommended such a parameter value are Brown (1967), Harrison (1967) and Montgomery and Johnson (1976). Chatfield (1978) has criticized these parameter ranges as arbitrary. However, they are suitable in the application of inventory control where forecasts are generated automatically. More support for these parameter values can be found in Gardner (1983, 1984).

A wider range of parameter values is recommended in other applications. From the studies of Makridakis et al. (1982) and Chatfield (1978), the most accurate parameters were frequently found to be in the range of 0 to 1. The choice of this limit for the smoothing parameters confirms that the level and slope are linear functions of past observations with coefficients that decrease as the data get older. Also McClain and Thomas (1973) showed that the trend model is suitable (invertible) over the range $0 < \alpha_1 < 2$ and $0 < \alpha_2 < 4 - 2\alpha_1$.

As was discussed earlier, Holt's exponential smoothing model is equivalent to an ARIMA(0,2,2) model (Harrison, 1967). Hence, the invertibility conditions for both models are identical. Therefore, compared to an ARIMA(0,2,2) model, the invertibility conditions for the local trend model are $\alpha_1 > 0$, $\alpha_2 > 0$ and $2\alpha_1 + \alpha_2 < 4$ (Box, Jenkins and Reinsel, 1994).

In an analysis of serial variation functions, Harrison (1967) showed that underestimation of the optimal parameters in the trend model is always more serious than overestimation. This finding is similar to that for simple exponential smoothing.

He also found that the departure of α_1 from the optimum affects more seriously the variance of the forecast errors than does such a departure for α_2. Again, this is supported by McClain and Thomas (1973).

2.4.4.3 Holt-Winters' Model (Method)

In Holt-Winters' exponential smoothing model, the smoothing parameters are usually restricted to the interval (0,1). However, this type of choice may have theoretical weaknesses. McClain (1974) has shown that for the additive seasonal model, some parameters within this interval cannot produce invertible models. Consequently, when applied to forecasting, the implicit weight given to the older series values is much larger than that given to more recent values (Archibald, 1990). Intuition indicates that such parameter values could produce poor forecasts.

McKenzie (1976) and then Roberts (1982) have shown that for additive seasonality, the Holt-Winters' model is equivalent to an ARIMA$(0,1,s+1)(0,s,0)_s$ model. The invertible region of this ARIMA model or equivalently, the additive seasonal model is developed by Archibald (1990) as:

$$\alpha_3(1-\alpha_1) > 0, \tag{2.4.81}$$

$$s\alpha_1 + \alpha_3 > \alpha_1\alpha_3, \tag{2.4.82}$$

$$\alpha_1 + \alpha_3(1-\alpha_1) < 2, \tag{2.4.83}$$

$$\alpha_1\alpha_2 > 0, \tag{2.4.84}$$

$$\alpha_1\alpha_2 < \alpha_2^*(\alpha_1,\alpha_3), \tag{2.4.85}$$

where

$$\alpha_2^*(\alpha_1,\alpha_3) \equiv [2-\alpha_1-\alpha_3(1-\alpha_1)](1-\cos\phi), \tag{2.4.86}$$

and ϕ is the smallest non-negative solution to:

$$\frac{\alpha_1}{\alpha_1+\alpha_3(1-\alpha_1)} = \frac{0.5\{1-\cos\phi-\cos((s-1)\phi)+\cos(s\phi)\}}{1-\cos(s-1)\phi}. \tag{2.4.87}$$

The regions for the smoothing parameters of the seasonal models suggested so far are complex and not easy to implement. From the empirical studies of Sweet (1983a, 1983b), it is observed that if there are four seasons (periods) in a cycle, the model is invertible for parameters between 0 and 1. However, with the parameters in this range, the model is not necessarily invertible if the cycle is 12 periods. In the latter case, the conditions for invertibility are complex and procedures for checking the invertibility of any set of parameters can be found in Sweet (1983a, 1983b). Note that Ord et al. (1997) have given invertibility conditions for the multiplicative seasonal model. However, as they point out, these conditions are necessary and may not be sufficient. As was mentioned earlier, there is no ARIMA equivalent of this model.

The forecasting performance of each exponential smoothing method (model) mainly depends on two factors, namely, the values of smoothing parameters and initial values of the seed vector. Different methods which have been suggested in the literature for initialization of the seed vector are outlined in the following section.

2.4.5 Initialization Methods

The state space models discussed in Subsection 2.4.2 contain unknown seed values. For example, for the additive Holt-Winter's model (equations (2.4.56) to (2.4.59)), the values of l_{t-1}, b_{t-1} and c_{t-s}, at time $t = 1$, are called seed values and the vector $(l_0, b_0, c_0, c_{-1}, \cdots, c_{-(s-1)})'$ is called the seed state vector. Usually, the seed state vector is unknown and needs to be initialized for estimation of the model. This will be discussed in this section.

The question of how to initialize the seed values has always been a subject of discussion since the introduction of exponential smoothing (Gardner, 1985; Cogger, 1973; McClain, 1981; Taylor, 1981; Wade, 1967; Makridakis and Hibon, 1991 and

Archibald, 1990). Several alternative methods have been suggested in the literature. However, there is little advice as to which to choose (see Chatfield and Yar, 1988). Brown (1963) suggested a method to estimate the initial value(s). This method advocates that the initial value(s) should come through comparison of the time series of interest with other similar series. Also, the value of the smoothing constant needs to be changed to reflect the uncertainty about these estimated initial value(s). The drawback of this method is that it is not easy to compare time series and to know which of them have similar behavior. Some of the methods used for initialization are discussed below.

2.4.5.1 Least Squares Estimates (OLS)

In practice, this is the most widely used approach for estimating initial value(s). Snyder (1985b) has discussed this method very clearly. He showed that the structural model, with appropriate transformations of the data can be converted to a classical regression model and proposed that the conventional least squares procedure be used to estimate the seed regression coefficient.

2.4.5.2 Convenient Initial Values

To initialize the smoothing equations, some convenient values can also be used. For example, the first data value can be used to initialize the level, i.e., $l_0 = y_1$. The difference between the first and second actual values ($b_0 = \frac{1}{2}(y_1 + y_2)$) or the average of the second minus the first and the fourth minus the third ($b_0 = \frac{1}{2}(y_2 - y_1) + (y_4 - y_3)$) (Makridakis and Wheelwright, 1978) can be used to initialize the trend. For the seasonal model, the seasonal components are initialized proportionally to the observations before being normalized.

2.4.5.3 Backcasting

A different type of approach used, for example, in the M competition (Makridakis et al., 1982), is backcasting. In this case, the time-order of the data is inverted and the most recent data value becomes period one, while the least recent becomes period n. Then, the seed vector is found by using the OLS method and the appropriate equations are used to forecast. The forecasted last value of the state vector is used as the initial estimate of the seed vector, except that the sign of the estimated value of trend is reversed.

2.4.5.4 Training Set

In this approach, the data are divided into two parts. The first part (usually the smaller of the two) is used to estimate the initial values for the exponential smoothing equation(s) used with the second part of the data (Makridakis et al., 1983).

2.4.5.5 Winters' Method

In this approach, the data are also divided into two parts and the first part (which is much smaller than the second part) is used to estimate the seed state vector. In the series, the periods are numbered $t = 1, 2, \cdots, n$. Denoting the length of the first part of the series by n^*, the number of years of data associated with this part is $\kappa = n^*/s$. Winters' (1960) suggested calculating the initial trend from the average change per period between the first and last year data of the first part, namely

$$\hat{b}_0 = \frac{1}{(\kappa - 1)s}(\bar{y}_\kappa - \bar{y}_1), \tag{2.4.88}$$

where s is the number of seasons, $\bar{y}_\kappa$ is the mean of observations in year κ and $\bar{y}_1$ is the mean of observations in the first year.

The initial estimate for l_0 is given by

$$\hat{l}_0 = \bar{y}_1 - \frac{1}{2}s\hat{b}_0. \tag{2.4.89}$$

Seasonal factors are computed for each season by

$$c^*_{ij} = \bar{y}_i - (\frac{1}{2}(s+1) - j)b_0, \tag{2.4.90}$$

where $i = 1, 2, \cdots, \kappa$, is the year and $j = 1, 2, \cdots, s$, is the position of the period within the year, e.g., for January, $j = 1$; for February $j = 2$; etc. in the case of monthly data.

Seasonal factors for corresponding periods in each of the initial years are averaged to obtain one seasonal factor for each season in a year. For example, the c^*_{ij} are averaged for all Januarys to get one January seasonal factor.

Assuming $\bar{c}^*_j = \sum_{i=1}^{\kappa} c^*_{ij}$ is the average for season j, the seasonals are normalized so that they sum to m for the multiplicative seasonal model and to 0 for the additive seasonal model. For example, the normalized seasonal values for the multiplicative seasonal model are

$$\hat{c}_{j-s} = \frac{\bar{c}^*_j}{\sum_{j=1}^{s} \bar{c}^*_j} \times s. \tag{2.4.91}$$

This adjustment ensures that over a cycle, the seasonal factors would make only seasonal adjustments and not increase or decrease the average level of the data.

2.4.5.6 Granger and Newbold Method

Granger and Newbold (1986) suggested setting the initial value of the level l_0 to the average observation in the first year, namely

$$l_0 = \bar{y}_1 = \frac{1}{s}\sum_{t=1}^{s} y_i \tag{2.4.92}$$

and b_0 to zero. The seasonal factors $c_0, c_1, \cdots, c_{-s+1}$ are calculated by comparing the appropriate observation in the first year with $\bar{y}_1$. For example, $c_0 = y_s - \bar{y}_1$ in the additive case and $c_0 = y_s/\bar{y}_1$ in the multiplicative case.

Some limited empirical findings on this approach are given by Chatfield (1978, Section 7). Other authors suggest calculating the initial values from the first two or three years.

2.4.5.7 Zero Values

All the initial values can be set to be zero, or one can be chosen as zero and the other(s) can be initialized using one of the alternatives described in Subsubsections 2.4.5.1 to 2.4.5.6 Although this approach seems to be unreasonable compared to other alternatives, it provides an advantage in terms of large initial errors. These large errors force the estimated values to approach the actual values much faster than alternative initialization procedures (see Makridakis and Hibon, 1991).

The study of Makridakis and Hibon (1991), shows that, in general, the post-sample forecasting accuracies do not depend on the initialization method. Shami (1997) compared the performances of different initialization methods (including OLS) in the analysis of M competition data and found OLS performs better than the others. Moreover, OLS is the most widely used initialization method and is easy to understand. Therefore, in this thesis, OLS will be considered as the initialization method for the seed state vector.

2.5 Forecasting Accuracy Performance

Forecasting methods are mainly classified into two broad groups, namely, judgemental methods and quantitative procedures. In forecasting the future, the first choice to be made is whether to rely on judgemental methods or use a quantitative

procedure. From the psychological literature, it is evident that in repetitive situations, quantitative methods outperform clinical judgement. Some early studies on the comparisons of quantitative methods and judgemental procedures can be found in Hogarth (1975), Makridakis et al. (1993), Lawrence et al. (1985), Chatfield (1989, section 5.4.3) and Collopy and Armstrong (1992), among others. The main limitations of judgemental forecasts are described as forecasting based on irrelevant information, lack of reliability, lack of application of valid principles and regression biases.

A number of researchers, outside the psychological literature, have analyzed in some detail the performance of judgemental and quantitative forecasts. Mabert (1975) compared the performance of judgemental and quantitative methods such as exponential smoothing, harmonic smoothing and Box-Jenkins and reported that judgemental forecasts give less accurate results, cost more and take more time. In a more complete study, Adam and Ebert (1976) found that Winters' method produced forecasts that were statistically more accurate than those of subjective forecasters.

There is less agreement as to which quantitative method is best in terms of forecasting accuracy. Since many more methods are available in the literature, in general, a comparison of the relative accuracy of different time series methods is not easy. Further, different sets of methods are used by different researchers when they conducted their comparisons. For example, in an early studiy, Kirby (1966), compared regression (trend fitting), moving average and exponential smoothing for monthly forecasting and reported in favor of exponential smoothing for shorter forecasting horizons. The trend fitting model did better for forecasting horizons of 12 or more. Newbold and Granger (1974, p.143) compared Box-Jenkins and Holt-Winters' forecasting methods and concluded in favor of the Box-Jenkins approach.

From the accuracy studies of McNees (1976) and Makridakis and Wheelwright (1978), it is evident that the most accurate method varies from one time period to the next and from one set of data to another.

The above seemingly contradictory findings about the accuracy performance of forecasting methods continued until the studies of Makridakis and Hibon (1979) and Makridakis et al. (1982). The findings of these studies assisted academics and practitioners greatly in choosing between alternative forecasting methods. Makridakis and Hibon (1979) considered 111 time series for comparing the forecasting performance of 22 different forecasting methods. The major findings of this particular study was that simple methods, such as exponential smoothing, perform better than the statistically sophisticated ones. However, this conclusion was in conflict with the recognized view of the time. More details about this can be found in the discussion following the Makridakis and Hibon (1979) paper. Makridakis et al. (1982) considered a bigger data set (total number of series was 1001) in order to incorporate the suggestions for improvements and to respond to the criticisms. The total number of forecasting methods used in Makridakis et al.'s (1982) study was 24. This extensive numerical study is known as the M competition. The results of the earlier study (Makridakis and Hibon) and those from the M competition were similar. More details about the four major findings from the M competition can be found, particularly in Makridakis et al. (1982) and also in Makridakis and Hibon (2000).

The major criticism of Makridakis et al.'s (1982) study was that in real situations, forecasters can use additional information to improve the forecasting accuracy of quantitative methods. In response to this criticism, Makridakis et al. (1993) performed another study which is known as the M2 competition. Incorporating all available additional judgemental information of forecasters and including

their knowledge about forthcoming economic and industry conditions, Makridakis et al. (1993) determined the post-sample accuracy of various methods/forecasters. Although few differences were found between conclusions of the M2 competition and those of the two previous studies of Makridakis and Hibon (1979) and Makridakis et al. (1982), the M2 competition clearly showed that simple methods can be used to find better predictions for real-life series. Further, in the context of the telecommunications data of Fildes (1992), Fildes et al. (1998) examined the robustness of the conclusion of the M competition data and reported that the findings of the M competition carry through for their telecommunications data. One additional conclusion drawn by Fildes et al. (1998) was that, in determining the relative performance of forecasting methods, the characteristics of the data series are an important factor. Therefore, developing any procedure that will identify and use the most appropriate method from a set of possible choices may be an important contribution to the field of forecasting.

Many researchers (Clemen, 1989; Geurts and Kelly, 1986; Fildes et al., 1998) have introduced new methods for forecasting and found that the results of their studies agree with those of the M competition. Also, the results from some additional studies (Armstrong and Collopy, 1992, 1993; Makridakis et al., 1993; Fildes et al., 1998) using new time series data agreed with the conclusions of the M competition. However, criticisms and emotional objections to empirical accuracy studies have continued in the literature. A detailed discussion of such criticisms/objections can be found in Fildes and Makridakis (1995). Similar results to those of the M and M2 competitions from other studies (Armstrong and Collopy, 1993; Fildes, 1992) increased the confidence in the conclusions made from Makridakis et al. (1982) and Makridakis et al. (1993). Makridakis and Hibon (2000) made an another attempt (M3 competition) to replicate and extend the M and M2 competitions by including

more methods, more researchers and also more data series (total number of series used was 3003). The results of this study confirmed the original conclusions of the M competition and again demonstrated that simple methods (e.g. SES and Holt's damped trend exponential smoothing method), in general, do better than statistically sophisticated methods. Further, some new methods such as the theta method (Assimakopoulos and Nikolopoulos, 2000) and robust trend method (Meade, 2000) also did well.

Bartolomei and Sweet (1989) compared the forecasting performance of Brown's general exponential smoothing (GES) method (Brown, 1963) and the Holt-Winters' method using 47 of 1001 time series which were used in the M competition of Makridakis et al. (1982). The study showed that although the Holt-Winters' method always fits the data better than the GES method, the former gives better forecasts than the GES method only for 55 percent of the series. Bartolomei and Sweet (1989) conjectured that the use of one of the trend methods proposed by Gardner and McKenzie (1985, 1989) may give better forecasts.

Gardner and McKenzie (1985) proposed a damped trend method as an extension of Holt's linear trend method (Holt's method). They compared Holt's method and Holt's damped trend method using 1001 series of Makridakis et al. (1982) and found that in general, the damped trend method performs better than Holt's method. Lewandowski (1979) and Parzen (1979) developed methods for time series forecasting which are known as the Lewandowski method and Parzen method, respectively. Gardner and McKenzie (1985) also analyzed a sample of 111 series from the population of 1001 series and compared the forecasting performance of Lewandowski, Parzen and Holt's damped trend methods. The results of this study show that both the Lewandowski method and Parzen method perform better than the damped trend method at longer forecast horizons. Further, Gardner and

McKenzie (1989) introduced the strategy of trend damping to the popular Holt-Winters exponential smoothing algorithm for seasonal time series. They considered a sample of 60 series from 1001 series of the M competition data and compared the forecasting performance of the following methods: the multiplicative Holt-Winters, the additive Holt-Winters' damped trend, the multiplicative Holt-Winters' damped trend, the Lewandowski method and the Parzen method. The results of the study showed that the Holt-Winters' damped trend method performs better than its competitors.

Chen (1997) investigated the robustness properties (in terms of forecasting performance) of four major forecasting methods for seasonal time series and concluded that the Holt-Winters method and the ARIMA fitting approach based on suitable parsimonious models have satisfactory robustness for a wide class of time series. They suggested that if the probabilistic structure of the data generating process is not clearly known to forecasters, the Holt-Winters method and parsimonious ARIMA methods are worth trying in practical situations. In another study, Chen (1993) investigated the robustness properties of the SES method and various AR processes and found that if the time series is stationary, the SES method is not as good as an AR process, though it is not too poor. However, if the time series deviates from a stationary processes, the SES can provide reasonable forecasts. Since many real life time series in various fields have complicated non-stationarities and structural changes that are not easy to detect, the SES method should be the choice in such situations. Further, from the study of Chen (1996), it is evident that the Holt-Winters method performs robustly for a wide class of time series that have a stochastic/deterministic linear trend and seasonality components. However, this method performs somewhat poorly for longer forecasting horizons (for more details, see Chen, 1996). This behavior of the Holt-Winters' method shows the necessity

of modifying the algorithm to accommodate a damped trend. Discussion on the damped trend method can be found in Gardner (1985) and Gardner and McKenzie (1985, 1989).

In conclusion, the ability of empirical studies to find forecasting methods which more accurately predict real life data should not be ignored. Makridakis and Hibon (2000, p.461) stated:

> We are convinced that those criticizing competitions, and empirical studies in general, should stop doing so and instead concentrate their efforts in explaining the anomalies between theory and practice and in working to improve the accuracy of forecasting methods ..... May be the time has come to follow the example of a recent conference on the Future of Economics (see *The Economist*, March 4th, 2000, p.90) and start debating, in a serious and scientific manner, the future of forecasting.

In this thesis, we aim to develop a numerical approach called individual selection method (ISM) to select models for forecasting real life time series data such as the M3 competition data of Makridakis and Hibon (2000).

2.6 Model Selection for Exponential Smoothing

The literature on IC based model selection for various types of models is vast. However, because of the lack of an appropriate modeling framework and likelihood based estimation methods for exponential smoothing algorithms, automatic model selection procedures such as AIC or BIC in the context of exponential smoothing are difficult to find in the literature.

Snyder (1985a) derived a state space formulation for simple exponential smoothing, Holt's linear trend and additive Holt-Winters' methods. Ord et al. (1997)

proposed a general class of nonlinear state space models with a single source of error. This model also underpins the multiplicative Holt-Winters method. Ord et al. (1997) also derived a conditional likelihood (CL) method for estimating these models. The development of Ord et al. (1997) opened the door for IC based model selection procedures for exponential smoothing methods. Following the framework of Ord et al. (1997), Hyndman et al. (2000) derived the state space formulation for the other exponential smoothing algorithms in Table 2.1. Hyndman et al. (2000) proposed an AIC based automatic model selection procedure using 1001 and 3003 series of the M and M3 competition data, respectively. They also considered a sample of 111 series from the population of the M competition. AIC was used to select an appropriate forecast model for each data series from a group of 24 models (both additive and multiplicative models). The results of this study demonstrated that the AIC based automatic forecasting model selection procedure performs well for short term forecasts.

Koehler et al. (1999) chose three more models from the original multiplicative Holt-Winters' model of Ord et al. (1997) and compared the model selection performance of CL and correlation methods in selecting from the original Holt-Winters' and new models for simulated time series (Koehler et al., 1999, provide more details about the correlation method). The results of this study showed that both methods perform reasonably well in selecting the true models.

An extensive study on IC based model selection procedures in the context of exponential smoothing methods (models) has been due since the development of exponential smoothing methods in the 1950s. In this thesis, we investigate the selection of exponential smoothing methods (models) from a set of simple and widely used methods (models).

2.7 Conclusions

In this chapter, we have reviewed some widely used IC procedures and exponential smoothing methods, along with their corresponding state space models. We have also discussed a global optimization algorithm called SA. There is a vast amount of literature on IC based model selection procedures, and this limits our ability to review all statistical properties associated with each IC procedure. Therefore, we have discussed only the main points for some important criteria namely, AIC, BIC, HQ, MCp, GCV, FPE and RVC. We have also discussed some of their asymptotic as well as finite sample properties. Most of these statistical properties of various IC procedures were investigated for model selection problems in linear regression models and ARMA processes.

Exponential smoothing methods and their corresponding state space models were briefly discussed, along with the ARIMA equivalence and smoothing parameter space of some selected smoothing methods. A number of initialization methods for the seed state vector have been proposed in the literature. We have briefly discussed some of the leading initialization methods. A survey of the forecast accuracy performance of various forecasting procedures was also presented in this chapter. The survey shows that the simple methods such as exponential smoothing methods do better than or as well as complicated and statistically sophisticated methods (see Makridakis et al., 1982; Makridakis et al., 1993; Fildes et al., 1998; Makridakis and Hibon, 2000; Hyndman et al., 2000). Although the idea of exponential smoothing methods has been available since the 1950s, there has not been any extensive IC based model selection procedures attempted in the literature. Therefore, one of the objectives of this thesis is to investigate the use of IC based model selection for exponential smoothing methods.

If the penalty value of an IC procedure is large, all else being equal, then smaller

models are favored. However, for a small penalty value the larger models are favored. Thus, assessing which IC is best for a given problem is very difficult. A well designed Monte Carlo study which covers many different DGPs will never uniformly favor one information criterion over another. In spite of these shortcomings, improvements on the existing IC approach may be possible by estimating penalty functions numerically. The SA algorithm can be used to examine such a possibility. The main aim of this thesis is, therefore, to propose a PEM by exploring the use of SA for maximizing OAPCS as well as minimizing forecast error in the context of time series models. Further, this thesis aims to investigate the use of PEM on the M3 competition data.

Chapter 3

Model Selection for Exponential Smoothing Methods

3.1 Introduction

In this chapter, we apply the existing IC procedures discussed in Chapter 2 for selecting exponential smoothing methods (models). There are many methods that fall in the general category of exponential smoothing. There are as many as 24 different versions for linear and non-linear exponential smoothing methods (see Chapter 2 and Hyndman et al., 2000). One reason for so many variations is that, in the real world, different types of time series behavior do occur and for the best forecasting of different types of time series, different methods have been designed. Some researchers have perceived the existence of particular time series behaviour which can be solved by some of these variations. However, in reality there are only a few time series for which these variations are relevant. Therefore, some of the variations can be ignored. On the other hand, many of the versions have some merit. This is clear from the popularity of exponential smoothing as a forecasting technique. Thus, for the practitioner in need of a forecast method for a particular time series and wishing to select it from the range of exponential smoothing approaches, an important question is: which of the exponential smoothing methods is the best for

this time series.

As discussed by Hillmer (1985), in order to select the best version, the forecaster needs experience, skill and some knowledge of the data series to be forecast. The practitioner needs to decide on whether the data are seasonal. If they are seasonal, then the practitioner needs to find out whether they are additive or multiplicative seasonality. Finally, a decision needs to be made as to whether the model needs a linear trend, a damped trend or is trend free. Some of these decisions are very complex. For example, if the data are seasonal, it is difficult to separate the problems of seasonal type and trend type. Some other technical questions also arise about the values chosen for the smoothing parameters and the starting value of the seed state vector. A variety of opinions about these issues can be found in an excellent review paper by Gardner (1985). However, because of the complexity of these issues, Hillmer (1985) suspected that the simplicity of exponential smoothing may be illusory.

A number of exponential smoothing methods have been mentioned in Chapter 2. An obvious question arises: how can we choose the right model for a particular data series? A good solution to this problem would be an automatic selection procedure for any application. This need is also recognized by Gardner (1985, p.38) who observed that "more research is needed on model identification and validation in exponential smoothing". Similar suggestions have also been made by McKenzie (1985). In other words, all we need in order to create a truly automatic exponential smoothing forecasting system is the development of a robust model selection procedure. For an automatic model selection procedure, a well developed modeling framework is necessary that incorporates stochastic models, likelihood calculation, prediction intervals and procedures for model selection. Although exponential smoothing methods have been around since the 1950s, the above gap

in the literature has limited the development of an automatic model selection approach. Snyder (1985a) and recently, Ord et al. (1997) and Hyndman et al. (2000) derived state space formulation for different exponential smoothing methods. Also, following Ord et al. (1997), exponential smoothing models can be put on a conditional likelihood (CL) footing, and hence, IC procedures can be applied to select exponential smoothing models. Thus, the aim of this chapter is to propose IC based automatic model selection procedures, which to the best of our knowledge have not been considered in the context of exponential smoothing models. In evaluating different IC procedures, the probability of correct selection depends on the model and its parameter values. Therefore, we compare the performance of IC procedures with respect to average probabilities of correct selection (APCS) as well as overall APCS (OAPCS) (see Section 3.5 for more details).

Previous studies (Tunnicliffe Wilson, 1989; Grose and King, 1994) showed that the MGL based IC procedures give improved probabilities of correct selection (PCS). Therefore, another aim of this chapter is to investigate whether we can use MGL methods to improve the quality of correctly model selection property. However, because the maximal invariants of different exponential smoothing models have different distributions, MGL cannot be used for IC based model selection procedures (for more details, see Section 3.4 and King, 1980). We solve this problem by proposing improved conditional likelihood (ICL), based on the MGL methods, and then apply ICL to IC procedures. We also consider CL based IC procedures, and compare the performance of CL and ICL with respect to APCS as well as OAPCS.

The plan of this chapter is as follows. The relationship between structural models and exponential smoothing models is outlined in Section 3.2. Section 3.3 shows how structural models can be transformed to a simple regression model. This

section also presents the OLS estimator of the seed state vector. The estimation methods of the models are discussed in Section 3.4. Theory for calculation of OAPCS using the IC approach is discussed in Section 3.5. A simulation study to calculate the PCS, APCS and OAPCS by using the existing IC procedures is outlined in Section 3.6, and the results of this study are discussed in Section 3.7. The chapter ends with some concluding remarks in Section 3.8.

3.2 Structural Form of Exponential Smoothing Models

In this section, we discuss structural (state space/dynamic) forms of exponential smoothing models. This framework of exponential smoothing models is the key to the development of the likelihood function for the exponential smoothing models (see Ord et al., 1997; Snyder, 1985a). Also, this framework allows the derivation of the general form of forecasting formulae for exponential smoothing models.

According to Harvey and Phillips (1979) and Harrison and Stevens (1976), Kalman (1960) filtering is the central to time series analysis. For example, models that lie within its jurisdiction are the classical and dynamic regression models, exponential smoothing and Box-Jenkins (1970) models. In the literature, Duncan and Horn's (1972) version of the dynamic linear model has been used traditionally to present the univariate cases of the above models. As shown by Snyder (1985a), their model, however, can be replaced by a simpler form of a dynamic linear model which is outlined as follows.

Let us assume that $y_1, y_2, \cdots, y_n$ are observed time series at time $t = 1, 2, \cdots, n$. All the past information contained in this time series can be condensed into the so called state vector β_t with small order k. This state vector, in turn, is used to provide information about future values of the series. The innovations form of the

structural model considered by Snyder (1985a) is given by

$$y_t = x_t'\beta_{t-1} + e_t, \tag{3.2.1}$$

$$\beta_t = T^*\beta_{t-1} + \alpha e_t, \tag{3.2.2}$$

$$\beta_0 = \beta_0^*, \tag{3.2.3}$$

where y_t is the observed value, x_t is a k-vector of independent variables, β_t is a unknown state vector of order k representing regression coefficients that change over time, e_t is an unobservable disturbance term which is independently and identically distributed with mean zero and variance σ^2, T^* is a $k \times k$ transition matrix of known parameters, α is a k-vector of smoothing parameters called the permanent effect vector and β_0^* is a k-vector of unknown seed values (called the seed state vector) for the regression coefficients. Equation (3.2.1) is called the measurement equation and equation (3.2.2) is called the transition equation.

Since the model has independent disturbances e_t, the model is stochastic. The regression coefficients β_t are governed by the transition equation and change over time. Hence, the model is dynamic and structural. The transition matrix, T^*, helps the model to update to systematic change. The model allows the regression coefficients to respond to unanticipated changes and consequently takes into account semi-systematic changes such as business cycles. The extent of the response is determined by the adjustment parameter vector α. Thus, the role of α is similar to the smoothing constants in exponential smoothing (Holt, 1957; Brown, 1959).

The framework (3.2.1) to (3.2.3) suggested by Snyder (1985a) is a mixture of the traditional multiple regression and exponential smoothing models. Box and Jenkins (1976, p.157) suggested transition equations for updating the coefficients of integrated versions of their model. The transition equations suggested by Snyder (1985a) are like those suggested by Box and Jenkins. It is also similar to the

Duncan and Horn (1972) framework for Kalman filtering. However, unlike Kalman filtering, it depends on only one primary source of randomness. It is still not known whether there is any resulting loss of generality in the framework suggested by Snyder (1985a). Even if there is a loss of generality, it should not be serious because the framework suggested by Snyder possesses counterparts for all the 'guidelines' listed in Harrison and Stevens (1976). The beauty of Snyder's framework is that it is very simple and easy to apply. Following Snyder (1985a), some examples of structural models are as follows.

Example 3.2.1 *Local Level Model*

The local level model can be expressed as a state space model with x_t as a unit vector of dimension one which is constant over time and the transition matrix T^* as a unit matrix of order 1×1. The only component of β_t is l_t and that of α is α_1. The state space model corresponding to the local level model is now as follows:

$$y_t = x_t'\beta_{t-1} + e_t, \tag{3.2.4}$$

$$\beta_t = T^*\beta_{t-1} + \alpha e_t, \tag{3.2.5}$$

or equivalently, it can be written as

$$y_t = l_{t-1} + e_t, \tag{3.2.6}$$

$$l_t = l_{t-1} + \alpha_1 e_t. \tag{3.2.7}$$

Example 3.2.2 *Local Trend Model*

The same structure can be applied to the local trend model. For this model, x_t is a 2×1 unit vector and T^* is a 2×2 upper triangular matrix with all elements equal to unity. The equivalent state space model of the local trend model is given by

$$y_t = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} l_{t-1} \\ b_{t-1} \end{bmatrix} + e_t, \tag{3.2.8}$$

$$\begin{bmatrix} l_t \\ b_t \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} l_{t-1} \\ b_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} e_t. \tag{3.2.9}$$

Equations (3.2.8) and (3.2.9) respectively, can be written as

$$y_t = x_t'\beta_{t-1} + e_t, \tag{3.2.10}$$

$$\beta_t = T^*\beta_{t-1} + \alpha e_t, \tag{3.2.11}$$

where $\beta_t' = (l_t,\ b_t)$ and $\alpha = (\alpha_1,\ \alpha_2)$.

Example 3.2.3 *Additive Holt-Winters' Model*

The additive Holt-Winters' model can be expressed as a state space model in the same way with x_t as an $(s+1) \times 1$ vector which is constant over time. The vector x_t is defined as

$$x_t' = (1,\ 1,\ -1, \cdots, -1). \tag{3.2.12}$$

The $(s+1) \times (s+1)$ transition matrix T^* is defined by

$$T^* = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & -1 & \cdots & -1 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}. \tag{3.2.13}$$

The unknown $(s+1) \times 1$ smoothing parameter vector α and the coefficient vector β_t are given by

$$\alpha' = (\alpha_1,\ \alpha_2,\ \alpha_3,\ 0, \cdots, 0), \tag{3.2.14}$$

$$\beta_t' = (l_t,\ b_t,\ c_t,\ c_{t-1}, \cdots, c_{t-s+2}), \tag{3.2.15}$$

where l_t, b_t and c_{t-i}, $i = 0, 1, 2, \cdots, s-2$, are the smoothing level, trend and seasonality respectively, and α_i, $i = 1, 2, 3$, are the smoothing parameters for the level, trend and seasonality, respectively.

3.3 OLS Estimate for the Seed Vector

The structural model can be converted to a classical linear regression model by an appropriate transformation of data (Snyder, 1985b). Therefore, instead of the Kalman filter, the conventional least squares procedures, particularly Gauss' (1821) recursive version of it, can be applied to estimate the seed state vector.

In this approach, all the observed values are written in terms of the seed values and the errors. Then, an estimate of the seed values can be obtained by applying OLS to these resulting transformations. Snyder (1985b) has derived a procedure for the above purpose. We consider the same procedure in our setting. The structural model of our concern is:

$$y_t = x_t'\beta_{t-1} + e_t, \tag{3.3.1}$$

$$\beta_t = T^*\beta_{t-1} + \alpha e_t, \tag{3.3.2}$$

$$\beta_0 = \beta_0^*. \tag{3.3.3}$$

Solving the measurement equation (3.3.1) for e_t, we get

$$e_t = y_t - x_t'\beta_{t-1}, \tag{3.3.4}$$

and substituting e_t from (3.3.4) into the transition equation (3.3.2) gives

$$\begin{aligned} \beta_t &= \alpha y_t + (T^* - \alpha x_t')\beta_{t-1} \\ &= \alpha y_t + D\beta_{t-1}, \end{aligned} \tag{3.3.5}$$

where $D = T^* - \alpha x_t'$, and is called the discount matrix. This equation shows a recursive relationship between the β_t. Hence, after back solving, all β_t depend linearly on the seed vector β_0^*. The solution has the general form

$$\beta_t = \Upsilon_t\beta_0^* + \omega_t, \tag{3.3.6}$$

where Υ_t is a $k \times k$ matrix and ω_t is a k-vector. Substituting the value of β_t in (3.3.5) gives:

$$\Upsilon_t \beta_0^* + \omega_t = \alpha y_t + D\beta_{t-1}. \tag{3.3.7}$$

Again substituting this value of β_{t-1} in (3.3.7), we get

$$\Upsilon_t \beta_0^* + \omega_t = \alpha y_t + D(\Upsilon_{t-1}\beta_0^* + \omega_{t-1}). \tag{3.3.8}$$

By equating the like terms of equation (3.3.8), the matrices Υ_t and vectors ω_t can be expressed by the following recurrence relationships:

$$\Upsilon_t = D\Upsilon_{t-1}; \qquad \Upsilon_0 = I, \tag{3.3.9}$$

$$\omega_t = D\omega_{t-1} + \alpha y_t; \qquad \omega_0 = 0. \tag{3.3.10}$$

Substituting the value of β_{t-1} from (3.3.6) into (3.3.1) gives

$$\begin{aligned} y_t &= x_t'(\Upsilon_{t-1}\beta_0^* + \omega_{t-1}) + e_t \\ &= x_t'\Upsilon_{t-1}\beta_0^* + x_t'\omega_{t-1} + e_t \\ &= x_t^{*'}\beta_0^* + x_t'\omega_{t-1} + e_t. \end{aligned} \tag{3.3.11}$$

The above equation can be written as

$$\tilde{y}_t = x_t^{*'}\beta_0^* + e_t, \tag{3.3.12}$$

where $\tilde{y}_t$ is a scalar and x_t^* is a k-vector determined by the respective transformations

$$\tilde{y}_t = y_t - x_t' w_{t-1}, \tag{3.3.13}$$

$$x_t^{*'} = x_t'\Upsilon_{t-1}. \tag{3.3.14}$$

Clearly, (3.3.12) is a classical linear regression model and an OLS procedure can be used for estimating the seed vector β_0^*. More specifically, the OLS estimate of β_0^* is given by

$$\hat{\beta}_0^* = (X'X)^{-1}X'\tilde{y}, \tag{3.3.15}$$

where $X = (x_1^{*'}, x_2^{*'}, \cdots, x_n^{*'})'$ is the matrix formed from the vector x_t^* and $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_n)'$. More details about the equivalent regression model of the structural model can be found in Snyder (1985b, 1985c).

In the above estimation procedure, it has been assumed that the smoothing parameter vector α is known. Usually this assumption is not realistic in applications. Therefore, a potential method for estimating α is needed, and one possibility involves maximum likelihood methodology. On the basis of normality assumptions for the errors, it is possible to find an expression for the likelihood function in terms of the one-step ahead prediction errors. This is outlined in the next section.

3.4 Maximum Likelihood Estimation

A recursive procedure similar to the Kalman filter has been proposed by Synder (1985a) to estimate the dynamic linear model which involves the measurement equation (3.3.1). In this paper, he assumed that the transition matrix T^* and the exponential smoothing parameter vector α are known. When either or both of them are unknown, maximum likelihood methods can be used to estimate the unknown parameters. By using the theory of conditional probability, the likelihood function (LF) can be written in a very simple form in terms of the one-step ahead prediction errors. Also, it can be shown that maximizing the likelihood is equivalent to minimizing the sum of squared one-step ahead prediction errors. The likelihood estimate must, therefore, be identical to the results from the Kalman filter described by Snyder (1985a).

Let us assume that $y_1, y_2, \cdots, y_n$ are random variables that are independent and identically distributed with probability density function $pr(y|\alpha)$, where α is a vector

of unknown exponential smoothing parameters. Then the LF of α is given by

$$L(\alpha|y) = \prod_{t=1}^{n} pr(y_t|\alpha). \tag{3.4.1}$$

The maximum likelihood estimator (MLE) $\widehat{\alpha}$ of α is defined by

$$L(\widehat{\alpha}|y) = \sup_{\alpha} L(\alpha|y). \tag{3.4.2}$$

3.4.1 MLE for Exponential Smoothing Models

At time t, the past time series values $y_1, y_2, \cdots, y_{t-1}$ are known, while the value y_t at time t is unknown. Consider the inductive hypothesis that the state vector β_{t-1} is fixed and known conditional on the seed state vector β_0^* and the smoothing parameter vector α is given.

From equation (3.3.1), the probability density function of $y_t|y_1, \cdots, y_{t-1}, \alpha, \beta_0^*, \sigma^2$ is identical to the probability density function of the error term e_t. Assuming the error terms, e_t, are independent and identically normally distributed, their probability density function is given by

$$pr(e_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2} e_t^2). \tag{3.4.3}$$

The Jacobian of the transformation between y_t given $y_1, \cdots, y_{t-1}, \alpha, \beta_0^*, \sigma^2$ and e_t is unity, and hence,

$$pr(y_t|y_1, \cdots, y_{t-1}, \alpha, \beta_0^*, \sigma^2) = pr(e_t). \tag{3.4.4}$$

At the end of period t, the y_t value is observed and a fixed value of the error term, the one-step ahead prediction error, can be calculated from (3.3.1) as follows:

$$e_t = y_t - x_t' \beta_{t-1}, \tag{3.4.5}$$

where $e_1 = y_1 - x_1' \beta_0^*$. Then, by using equation (3.3.2), a fixed value of β_t can be calculated.

The observations $y_1, y_2, \cdots, y_n$ are not independent for structural models. However, knowledge of the past values of the series can help to calculate the probability density function of y_t. The repeated application of the probability law $pr(A^* \bigcap B^*) = pr(A^*|B^*)pr(B^*)$ gives the following relationship:

$$pr(y_1, y_2, \cdots, y_n|\alpha, \beta_0^*, \sigma^2) = \prod_{t=2}^{n} pr(y_t|y_1, \cdots, y_{t-1}, \alpha, \beta_0^*, \sigma^2) \times pr(y_1|\alpha, \beta_0^*, \sigma^2). \quad (3.4.6)$$

From equations (3.4.4) and (3.4.6)

$$pr(y_1, \cdots, y_n|\alpha, \beta_0^*, \sigma^2) = \prod_{t=1}^{n} pr(e_t) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \prod_{t=1}^{n} \exp(-\frac{1}{2\sigma^2} e_t^2). \quad (3.4.7)$$

The seed state vector β_0^* has a well defined density [1] for a stationary time series. So for a stationary case, an unconditional density $\psi(y_1, \cdots, y_n|\alpha, \sigma^2)$ can be obtained. However, in most real life applications of exponential smoothing, the time series are non-stationary, and hence, the distribution of β_0^* does not exist. Therefore, the unconditional density of the sample cannot be obtained. The definition of the likelihood must be based upon the density (3.4.7), because it summaries all the information that can be known about the sample generated by a non-stationary stochastic process. Thus, the CL which is a function of β_0^* together with the parameters α and σ is as follows:

$$L(\alpha, \beta_0^*, \sigma^2|y_1, y_2, \cdots, y_n) = pr(y_1, \cdots, y_n|\alpha, \beta_0^*, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp(-\frac{1}{2\sigma^2} \Sigma_{t=1}^{n} e_t^2). \quad (3.4.8)$$

The prediction error decomposition of the likelihood function associated with the Kalman filter (Schwepe, 1965; Harvey, 1991) and the likelihood function (3.4.8)

[1] For example, it can be shown that the seed value $\beta_0^* = l_0$ of the local level model with damped parameter has a finite variance.

looks the same. However, the difference between these two is that the one-step ahead prediction errors in (3.4.8) from the Kalman filter are heteroscedastic and those from exponential smoothing are homoscedastic.

Taking logs on both sides of (3.4.8):

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n} e_t^2. \tag{3.4.9}$$

The estimate of the variance of σ^2 can be obtained by maximizing $\log L$ with respect to σ^2 and the estimate is given by

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{n}\sum_{t=1}^{n} e_t^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}(y_t - x_t'\beta_{t-1})^2. \end{aligned} \tag{3.4.10}$$

The estimates of α and β_0^* are very complex. However, after fixing α, the structural model can be converted to a classical regression model by an appropriate transformation of the data. Then, an OLS method can be used to estimate the seed vector β_0^* (see Section 3.3).

Substitution of (3.4.10) into (3.4.9) yields the maximum likelihood value

$$\begin{aligned} \log L &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\tilde{\sigma}^2 - \frac{n}{2} \\ &= \text{Const} - \frac{n}{2}\log\tilde{\sigma}^2, \end{aligned} \tag{3.4.11}$$

where Const $= -\frac{n}{2}(\log 2\pi + 1)$. Ignoring the constant term, the likelihood depends only upon $\tilde{\sigma}^2$ which is a function of the smoothing parameter α. Hence, calculating maximum likelihood estimate with respect to α is the same as obtaining α so as to minimize the estimate of the error variance σ^2.

$$\text{Max}(\log L) \Longleftrightarrow \text{Min}\left(\frac{n}{2}\log\tilde{\sigma}^2\right). \tag{3.4.12}$$

This is equivalent to minimizing sum of the squared residuals or estimated forecast errors, i.e.,

$$\text{Max}(\log L) \Longleftrightarrow \text{Min}\left(\sum_{t=1}^{n} \tilde{e}_t^2\right). \tag{3.4.13}$$

The likelihood eventually depends only on exponential smoothing parameter vector α, and is highly non-linear. As a result, an analytical solution for the closed form estimate of α is complex and we believe impossible. Therefore, numerical search methods such as the GAUSS constrained optimization technique or SA as outlined in Chapter 2 need to be employed to solve this estimation problem.

3.4.2 ICL Via MGL Estimates

From the study of Grose and King (1994), it is evident that MGL (for details on MGL see Ara, 1995; Tunnicliffe Wilson, 1989; King, 1980) based model selection is better than that based on the profile likelihood (PL). Similar results were also given by Tunnicliffe Wilson (1989) who considered the problem of choosing between alternative time series models. This may be because the MGL provides better estimates than those from the PL (Laskar and King, 1998). Therefore, it is of interest to see how the MGL performs in the exponential smoothing settings.

In matrix and vector notation, regression equation (3.3.12) can be written as:

$$\tilde{y} = X\beta_0^* + e, \tag{3.4.14}$$

where $\tilde{y}$ and X are formed from $\tilde{y}_t$ and $\tilde{x}_t^*$, respectively. It should be noted that $\tilde{y}_t$ and $\tilde{x}_t^*$ were defined in equations (3.3.13) and (3.3.14), respectively. It is noteworthy to observe that the design matrices of different exponential smoothing models such as the local level, local trend and Holt-Winters' are different. For these models, the design matrices are as follows.

From equations (3.3.13) and (3.3.14), the vector x_t^* can be expressed as $x_t^{*'} = x_t' D^{t-1}$, $t = 1, 2, \cdots, n$, where D is the discount matrix defined in Section 3.3. D is different for different exponential smoothing models. Let us denote the vector $x_t^{*'}$ and the discount matrix D for the local level, local trend and additive Holt-Winters' models by $x_{t,lm}^{*'}$, $x_{t,tm}^{*'}$ and $x_{t,sm}^{*'}$, respectively and D_{lm}, D_{tm} and D_{sm}, respectively. Then, the design matrices for the local level, local trend and additive Holt-Winters' models are given by

$$X_1 = \begin{bmatrix} x_{1,lm}^{*'} \\ x_{2,lm}^{*'} D_{lm} \\ \vdots \\ x_{n,lm}^{*'} D_{lm}^{n-1} \end{bmatrix}_{n\times 1}, X_2 = \begin{bmatrix} x_{1,tm}^{*'} \\ x_{2,tm}^{*'} D_{tm} \\ \vdots \\ x_{n,tm}^{*'} D_{tm}^{n-1} \end{bmatrix}_{n\times 2} \text{ and}$$

$$X_3 = \begin{bmatrix} x_{1,sm}^{*'} \\ x_{2,sm}^{*'} D_{sm} \\ \vdots \\ x_{n,sm}^{*'} D_{sm}^{n-1} \end{bmatrix}_{n\times(s+1)}, \tag{3.4.15}$$

respectively.

MGL (Ara, 1995) can be thought of as the density function of the maximal invariant vector ν (see King, 1980), which itself is a function of X. The maximal invariants of different exponential smoothing models are different, because the design matrices of the models are not the same. But, because X is different for different models, the maximal invariant statistics are different. Let us assume that ν_1, ν_2 and ν_3 are maximal invariants for the local level, local trend and Holt-Winters models, respectively, where

$$\nu_i = \frac{M_i^* y}{\|M_i^* y\|^{1/2}} = g_i(y), i = 1, 2, 3, \tag{3.4.16}$$

in which $M_i^* = I_n - X_i(X_i'X_i)^{-1}X_i'$ and I_n is the $n \times n$ identity matrix. We assume that $f_1(\nu_1)$, $f_2(\nu_2)$ and $f_3(\nu_3)$ are the distributions of ν_1, ν_2 and ν_3, respectively. Clearly, ν_1, ν_2 and ν_3 have different dimensions and g_i^{-1} does not exist. This means

that the maximal invariants ν_1, ν_2 and ν_3 have different distributions. Therefore, MGL cannot be used for IC based model selection procedures in the usual way for exponential smoothing models.

It is well known that the MGL based estimates are better (see Tunnicliffe Wilson, 1989; Ara, 1995) than those based on any other competitive methods. Therefore, although in the exponential smoothing framework the application of the MGL to IC is not appropriate, the estimate of α obtained from it can be useful to improve model selection criteria as discussed below. Assuming that the MGL based estimate of α is better than the one estimated from the CL, using the MGL estimate of α in the CL function is likely to give a better likelihood value. We hope that the application of the MGL estimate based CL to IC will give better probabilities of correct model selection. This is investigated in Section 3.6.

3.5 Theory of OAPCS for IC Procedures

Among the existing IC procedures, that which selects the correct model or an appropriate model most often is considered as the best performing IC. However, the choice of the best IC procedure may depend on various factors such as the models to be selected from and values considered for the exponential smoothing parameters of the model. Particularly, probabilities of correct selection are affected largely by the values of exponential smoothing parameters (see Subsection 3.7.1). For some parameter values, the probabilities of correct selection is lower and for other values it is higher. Further, in real life applications, the values of exponential smoothing parameters are neither known nor necessarily fixed. Therefore, it would be better if we can study the APCS. In this section, we outline a method of calculating APCS, and hence, OAPCS by using a particular IC procedure.

Let us assume that $M_1, M_2, \cdots, M_N$ are N competing models. When data are

Table 3.1: Maximized log-likelihood values.

DGP	Estimated Models			
	M_1	M_2	$\cdots$	M_N
M_1	$\log L_1(\hat{\alpha}^1)$	$\log L_2(\hat{\alpha}^2)$	$\cdots$	$\log L_N(\hat{\alpha}^N)$
M_2	$\log L_1(\hat{\alpha}^1)$	$\log L_2(\hat{\alpha}^2)$	$\cdots$	$\log L_N(\hat{\alpha}^N)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
M_N	$\log L_1(\hat{\alpha}^1)$	$\log L_2(\hat{\alpha}^2)$	$\cdots$	$\log L_N(\hat{\alpha}^N)$

simulated from the ith model, $i = 1, \cdots, N$, all the models in the competing set are fitted by maximizing their log-likelihoods. Let $\log L_j(\hat{\alpha}^j)$ be the maximized log-likelihood value for the jth model when data are generated from the ith model, and $\hat{\alpha}^i$ is the corresponding estimated exponential smoothing parameter vector, $i, j = 1, \cdots, N$. Given the true models, all of the estimated log-likelihoods are shown in Table 3.1. The diagonal elements of this table represent the estimated log-likelihoods for the true models, while the off-diagonal elements show the log-likelihoods for the other models.

Let us assume that we have generated R time series for each of the N models in the competing group. At each replication we estimate maximized log-likelihoods as shown in Table 3.1. We then penalize each of these estimated maximized log-likelihoods with a penalty term from a particular IC procedure and compute the number of times the true model is selected.

We choose the model with the largest value of the IC over all models under consideration, i.e., model M_i will be selected if

$$\log L_i(\hat{\alpha}^i) - p_i > \log L_j(\hat{\alpha}^j) - p_j, \quad \text{for all} \quad j \neq i, \quad j = 1, \cdots, N,$$

where $\log L_i(\hat{\alpha}^i)$ is the maximized log-likelihood function for model M_i with $\hat{\alpha}^i$

being the maximum likelihood estimate of α^i and p_i is the corresponding penalty.

Let us assume that $Pr(CSM_i|M_i, \alpha^i, p_1, \cdots, p_N)$ denotes the probability of correct selection that model M_i is true with parameter vector α^i when penalties $p_1, \cdots, p_N$ are used. This probability can be given by

$$\begin{aligned} &Pr(CSM_i|M_i, \alpha^i, p_1, \cdots, p_N) \\ &= Pr[\log L_i(\widehat{\alpha}^i) - p_i > \log L_j(\widehat{\alpha}^j) - p_j, j \neq i, j = 1, \cdots, N|M_i, \alpha^i]. \end{aligned} \tag{3.5.1}$$

The value of (3.5.1) changes with α^i changes. An obvious way to overcome this problem is to work with the APCS. This requires a weighting density function for different values of α^i, similar to a prior density function used in Bayesian statistics. Assuming $\zeta(\alpha^i)$ is the weighting density function for α^i, the APCS for the ith true model M_i is given by

$$\text{APCS}_i = \int Pr[CSM_i|M_i, \alpha^i, p_1, \cdots, p_N]\zeta(\alpha^i)d\alpha^i. \tag{3.5.2}$$

Given (3.5.1), we can estimate (3.5.2) through Monte Carlo integration by drawing α^i randomly from the distribution given by $\zeta(\alpha^i)$ and using these parameter values to generate R simulated data sets $y_{t\ell}$, $t = 1, \cdots, n$; $\ell = 1, \cdots, R$, of sample size n. The only remaining problem is that (3.5.1) is unknown. It can easily be estimated by Monte Carlo methods. In a different setting, King and Bose (2000) have investigated how best to estimate expressions such as (3.5.2) via Monte Carlo simulation. The question they addressed was the best split of a total number of replications between those needed to estimate (3.5.1) and those needed to estimate (3.5.2). They found the best results are obtained by maximizing the number of drawings of α^i and using only one iteration for each estimate of (3.5.1) in (3.5.2). If

$$\begin{aligned} &I_\ell(M_i, \alpha^{i\ell}, p_1, \cdots, p_N) \\ &\quad = I\left(\log L_i^\ell(\widehat{\alpha}^i) - p_i > \log L_j^\ell(\widehat{\alpha}^j) - p_j, j \neq i, j = 1, \cdots, N|M_i, \alpha^{i\ell}\right) \end{aligned}$$

denote the indicator function for the event that M_i is correctly chosen at the ℓth iteration, where $L_j^{\ell}(\widehat{\alpha}^j)$ is the maximized log-likelihood for M_j, based on the data set of the ℓth iteration, and $\alpha^{i\ell}$ is the ℓth drawing from $\zeta(\alpha^i)$, then the estimated value of $\mathrm{APCS}_i(p_1,\cdots,p_N)$ is:

$$\widehat{\mathrm{APCS}}_i(p_1,\cdots,p_N)=\frac{1}{R}\sum_{\ell=1}^{R} I_{\ell}(M_i,\alpha^{i\ell},p_1,\cdots,p_N). \tag{3.5.3}$$

The OAPCS for the set of penalties $p_1,\cdots,p_N$, is obtained by

$$\widehat{\mathrm{OAPCS}}(p_1,\cdots,p_N)=\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathrm{APCS}}_i(p_1,\cdots,p_N). \tag{3.5.4}$$

The determination of the best IC can be made by evaluating the OAPCS for each IC procedure.

3.6 Design of the Monte Carlo Study

A Monte Carlo experiment was conducted to investigate how the existing IC procedures perform for selecting the true model from a group of exponential smoothing models. Another aim was to compare the CL and ICL based selection probabilities. Further, the effects of changing factors such as sample size n, standard deviation σ and exponential smoothing parameter vector α were also observed. Because the idea of applying the IC approach to select a model from a group of exponential smoothing models is fairly new, only two models, namely, the local level and local trend were considered for the first part of the experiment, and data were generated from the underlying models using selected values for exponential smoothing parameters. Then, for the second part of the experiment an additive seasonal model, which has much use in applications, was included into the above group of competing models. The local level, local trend and seasonal models are denoted by SM1, SM2 and SM3, respectively. In this part, at each replication, exponential

smoothing parameters were generated randomly from a weighting distribution. For comparison, we considered the following six IC procedures: AIC, BIC, HQ, MCp, GCV and FPE.

3.6.1 First Part of the Experiment

In this part, the existing IC procedures were used to select between the SM1 and SM2 models to see how they perform in choosing the true models. The data were generated by the underlying models for some specified values of the exponential smoothing parameter vector. The effect of changing the exponential smoothing parameters was observed in this part of the experiment. The sample size and parameter choices in the data generating processes were as follows: $n = 36$, $\sigma = 1$ and 5. For the SM1 model the smoothing parameter values were $\alpha_1 = 0.0$, 0.1, 0.2, 0.5, 0.7 and 0.9 and those for the SM2 model were $\alpha_1 = 0.0$, 0.1, 0.2, 0.5, 0.7 and 0.9 and $\alpha_2 = 0.01$ and 0.1. The initial seed level l_0 and seed growth rate (trend) b_0 were set to 100 and 5.0, respectively. The results are discussed in Subsection 3.7.1.

3.6.2 Second Part of the Experiment

The second part of the simulation study was conducted in the context of selection between SM1 and SM2 models, SM2 and SM3 models, and SM1, SM2 and SM3 models. For all of these models, the parameter of interest is the vector of exponential smoothing parameters. Although, theoretically the value of smoothing parameters can be greater than unity, in most applications they are considered to be within 0 and 1. However, more restricted smoothing parameter values were considered for trend and seasonalitity (see for example, Hyndman et al., 2000). Let us assume that α^i, $i = 1, 2, 3$, are the exponential smoothing parameter vector for the SM1, SM2 and SM3 models respectively, where $\alpha^1 = \alpha_1$, $\alpha^2 = (\alpha_1,\ \alpha_2)'$ and $\alpha^3 = (\alpha_1,\ \alpha_2\ \alpha_3)'$. We also assume that $\zeta(\alpha^i)$ is the weighting distribution (uniform) for the ith model.

At each replication, the values of exponential smoothing parameters were generated randomly from independent uniform distributions with ranges as follows:

Weighting Distribution	Smoothing Parameter Value		
	α_1	α_2	α_3
SM1 : $\zeta(\alpha^1)$:	[0, 0.999]		
SM2 : $\zeta(\alpha^2)$:	[0, 0.999]	[0, 0.2]	
SM3 : $\zeta(\alpha^3)$:	[0, 0.999]	[0, 0.2]	[0, 0.1]

This helps to estimate APCS, and hence, OAPCS. The performance of the IC procedures was compared in terms of APCS as well as OAPCS.

The necessary steps of the Monte Carlo study are as follows:

Step 1: Draw $\alpha^1 = \alpha_1$, $\alpha^2 = (\alpha_1,\ \alpha_2)'$ and $\alpha^3 = (\alpha_1,\ \alpha_2,\ \alpha_3)'$ values from their respective $\zeta(\alpha)$ distributions at each replication.

Step 2: Generate three samples of size n from each of the SM1, SM2 and SM3 models for these parameter values.

Step 3: For each data set, compute the maximized log-likelihood for all of the competing models as shown in Table 3.1.

Step 4: Repeat steps 1 to 3 R times, where R is the number of replication.

Step 5: Use maximized log-likelihoods obtained in step 4 for calculating APCS (equation (3.5.3)) for each true model by using various IC procedures. Also, using (3.5.4) calculate OAPCS by averaging APCS over the true models.

For comparison, we calculated OAPCS for each IC procedure considered in the experiment, based on both CL and ICL. The sample size and σ value were considered: $n = 24$, 48, 72, 96, 120 and 200 and $\sigma = 1$, 5 and 20, respectively. We calculated the initial seasonal values by using the equation $c_{j-s} = 1 + A\sin(2j\pi/s)$, $j = 1, 2, \cdots, s$, where A is the seasonal amplitude and s is the number of seasons in a year. The values of A and s were set to 0.3 and 4, respectively (assuming quarterly time series). For this choice of A, see Koehler et al. (1999). The initial

seed state vector was estimated by using the OLS method on the transformed model as discussed in Section 3.3.

Traditionally, the values of smoothing parameters are obtained through making a good guess in most applications. The studies of Chatfield (1978) and Bartolomei and Sweet (1989) show that it is dangerous to guess the values of the smoothing parameters. In fact, the parameters should be estimated from data. Therefore, both CL and MGL estimation methods were used to estimate the exponential smoothing parameters by using the data series under consideration. Since none of the likelihoods (CL and MGL) have a closed form estimator for α, the GAUSS (see Aptech, 1996) constrained optimization algorithm was used for maximizing both the CL and MGL.

3.7 Results and Discussion

The results of the simulation study are organized as follows. The probabilities of correct selection (PCS) for the SM1 and SM2 models from the first part of the experiment are presented in Tables 3.2 to 3.4. From the second part of the study, the APCS and OAPCS are presented in Tables 3.5 to 3.7, 3.8 to 3.10 and 3.11 to 3.16 when correct models are selected from the competing groups SM1 and SM2, SM2 and SM3 and SM1, SM2 and SM3, respectively. Tables 3.17 to 3.20 present the effect of sample size on APCS and OAPCS. The effect of sample size is obtained by taking the average of APCS and OAPCS across all σ values for a particular n. Similarly, Tables 3.21 to 3.23 show the effect of σ value on APCS and OAPCS. This means that for a particular value of σ, the APCS and OAPCS were averaged across all sample sizes. For a clearer view, selected results from Tables 3.5, 3.8 and 3.11 are presented in Figures 3.1 to 3.3, respectively. It should be noted that the results are discussed below in terms of correctly selected models only.

3.7.1 First Part of the Experiment

Table 3.2 presents the PCS for the SM1 model. This table shows that irrespective of the value of α_1, BIC correctly selects the SM1 model most often, followed by HQ, and the third position is occupied by GCV. The fourth and fifth positions are occupied by either AIC or FPE, while MCp correctly selects the SM1 model least often. The PCSs for all existing IC procedures used in this experiment are affected by the choice of smoothing parameter values. For example, Table 3.2 shows that for $n = 36$ and $\sigma = 1$, the PCS for ICL based BIC are 97.9%, 90.4%, 79.9%, 84.4%, 90.8% and 93.6% for $\alpha_1 = 0.0$, 0.1, 0.2, 0.5, 0.7 and 0.9, respectively. Clearly, for smoothing parameter values close to zero and unity, all the IC procedures perform very well at correctly selecting the SM1 model. In general, the PCS for the SM1 model is not affected by changing the σ value.

Tables 3.3 and 3.4 show the PCS for the SM2 model. According to these tables, the existing IC procedures perform in the opposite direction to their performance for correctly selecting the SM1 model. More clearly, MCp selects the SM2 model most often, followed by either AIC or FPE. The fourth and fifth positions are occupied by GCV and HQ respectively, and BIC selects the SM2 model least often. The PCS for the SM2 model decreases with increase of values of α_1 and α_2. Also, the PCS decreases with σ increases.

When correctly selecting the SM1 model, irrespective of the values of α_1 and σ, the ICL based existing IC procedures perform better than those based on CL. Table 4.2 shows that for $\sigma = 5$, ICL based AIC dominates CL based AIC by 9.86% to 33.84% for various values of α_1 in the interval 0.0 to 0.9. However, the CL based IC procedures perform better than those based on ICL when correctly selecting the SM2 model. For example, CL dominates ICL by 0.04% to 9.52% for $\sigma = 5$, $\alpha_2 = 0.1$ and different values of α_1 ranging from 0.0 to 0.9.

Clearly, the results of the first part of the experiment show that the PCSs of the SM1 and SM2 models depend on the choice of the values of exponential smoothing parameters. Therefore, we calculated APCS as well as OAPCS in the second part of the experiment. The results are discussed as follows.

3.7.2 Second Part of the Experiment

As shown in Tables 3.5 to 3.16, in the context of APCS, it is clear that MCp most often selects the largest (in terms of the number of smoothing parameters involved) model in the competing group than any other existing IC procedures considered in this chapter. MCp is followed by either AIC or FPE. The fourth and fifth positions are occupied by GCV and HQ respectively, and BIC correctly selects the largest model least often. On the other hand, when selecting other models in the group, the existing IC procedures are found to perform in the reverse order to their performance of selecting the largest model.

In terms of OAPCS, in general, the performance of the existing IC procedures follow the same order as they did in terms of APCS when correctly selecting the models other than the largest model. When correctly selecting from the competing models SM1 and SM2, AIC, FPE and MCp perform equally, particularly for large sample sizes (see for example, Table 3.7).

It is interesting to observe that the above ranking between the existing IC procedures with respect to APCS and OAPCS carries through for both CL and ICL based model selection criteria. In terms of APCS, ICL performs better than CL if correctly selecting models other than the largest model. According to Table 3.11 (competing models are SM1, SM2 and SM3), for $n = 24$ and $\sigma = 5$, with respect to APCS, the ICL based IC procedures perform better than those based on CL by 17.1% to 31.8% and 3.1% to 11.0% when correctly selecting the SM1

and SM2 models, respectively. However, when correctly selecting the SM3 model, the APCSs for CL based IC procedures are 0.4% to 1.9% larger than that of those based on ICL. In terms of OAPCS, ICL uniformly dominates CL. For example, the ICL based BIC (best among the existing IC procedures) performs better than that based on CL by 0.12% to 6.08% (see Tables 3.5 to 3.16). The OAPCS for $n = 24$ and σ =1, 5 and 20 from Tables 3.5, 3.8 and 3.11 are presented in Figures 3.1 to 3.3, respectively. These figures clearly show the difference between the performance of CL and ICL based selection criteria in terms of OAPCS.

3.7.3 Effects of n and σ on APCS and OAPCS

Tables 3.17 to 3.20 show that the APCS for each model increases as the sample size increases. However, the improvement in APCS for the SM2 model is not remarkable when it is selected from the competing groups SM2 and SM3, and SM1, SM2 and SM3. When selecting from SM1, SM2 and SM3, the APCS for the SM1 model is low for small sample sizes, and it improves quickly as the sample size increases. For all IC procedures, the OAPCS increases with increases in sample size.

The effects of σ on APCS as well as on OAPCS can be observed from Tables 3.21 to 3.23. In terms of APCS, both CL and ICL based correct selection probabilities of the smallest model (in terms of the number of parameters involved) of each plausible group remains almost unchanged as σ increases. However, APCS corresponding to other models decreases with σ increases. Irrespective of σ value, ICL uniformly performs better than CL with respect to OAPCS, and its superiority to CL decreases with σ increases.

3.8 Conclusions

The likelihood approach to exponential smoothing models allows the application of IC procedures for selecting the best model from a group of competing models. The results of this chapter show that the selection performance of the various IC procedures depends on the value of the exponential smoothing parameter, sample size n, σ and the models considered in the competing group. The OAPCS for IC procedures decreases as σ value increases. However, as was expected, OAPCS increases as the sample size increases. In terms of OAPCS, BIC performs best, followed by HQ, and MCp is the worst. All the existing IC procedures perform well in selecting exponential smoothing models, but the selection criterion with the largest penalty gives better selection probabilities. Among the existing IC procedures, BIC has the largest penalty value.

We have introduced ICL based IC procedures, where ICL is based on the MGL methods. We compared the performance of the CL and ICL based IC procedures with respect to APCS and OAPCS. The results of our study show that in terms of APCS, when correctly selecting the smaller models, ICL based IC procedures perform better than those based on CL. However, when selecting the largest model in the group, the CL based IC procedures usually perform better than those based on ICL. Irrespective of sample size, σ value and competing group, in terms of OAPCS, the ICL based IC procedures perform better than those based on CL. Therefore, we suggest the use of ICL based BIC criterion when faced with selection problem between different exponential smoothing models.

Table 3.2: The PCS for the SM1 model (when SM1 and SM2 are competing models) for $\sigma = 1$ and 5, $n = 36$, $l_0 = 100$ and different values of exponential smoothing parameter α.

n	σ	α_1	LF	AIC	BIC	HQ	MCp	GCV	FPE
36	1	0.0	CL	0.832	0.933	0.885	0.827	0.846	0.833
			ICL	0.914	0.979	0.938	0.912	0.922	0.915
		0.1	CL	0.629	0.764	0.682	0.625	0.639	0.631
			ICL	0.779	0.904	0.820	0.777	0.784	0.779
		0.2	CL	0.527	0.658	0.579	0.520	0.536	0.528
			ICL	0.702	0.799	0.728	0.700	0.705	0.703
		0.5	CL	0.674	0.792	0.726	0.669	0.687	0.674
			ICL	0.781	0.844	0.811	0.780	0.784	0.781
		0.7	CL	0.758	0.875	0.811	0.752	0.767	0.759
			ICL	0.857	0.908	0.883	0.853	0.861	0.858
		0.9	CL	0.799	0.912	0.851	0.793	0.813	0.799
			ICL	0.878	0.936	0.908	0.876	0.881	0.878
36	5	0.0	CL	0.832	0.933	0.885	0.827	0.847	0.833
			ICL	0.914	0.979	0.938	0.912	0.922	0.915
		0.1	CL	0.629	0.764	0.682	0.625	0.639	0.631
			ICL	0.777	0.904	0.819	0.776	0.783	0.777
		0.2	CL	0.526	0.658	0.579	0.520	0.536	0.528
			ICL	0.704	0.800	0.729	0.701	0.706	0.704
		0.5	CL	0.674	0.793	0.726	0.669	0.687	0.674
			ICL	0.784	0.846	0.813	0.783	0.787	0.784
		0.7	CL	0.758	0.875	0.811	0.753	0.767	0.759
			ICL	0.857	0.909	0.883	0.853	0.861	0.858
		0.9	CL	0.800	0.912	0.852	0.794	0.813	0.800
			ICL	0.880	0.938	0.910	0.878	0.883	0.880

Table 3.3: The PCS for the SM2 model (when SM1 and SM2 are competing models) for $\sigma = 1$, $n = 36$, $l_0 = 100$, $b_0 = 5$ and different values of exponential smoothing parameters α_1 and α_2.

n	σ	α_1	α_2	LF	AIC	BIC	HQ	MCp	GCV	FPE
36	1.0	0.0	0.01	CL	1.000	1.000	1.000	1.000	1.000	1.000
				ICL	0.999	0.999	0.999	0.999	0.999	0.999
		0.1		CL	1.000	1.000	1.000	1.000	1.000	1.000
				ICL	0.999	0.999	0.999	0.999	0.999	0.999
		0.2		CL	1.000	1.000	1.000	1.000	1.000	1.000
				ICL	0.999	0.999	0.999	0.999	0.999	0.999
		0.5		CL	0.999	0.998	0.999	0.999	0.999	0.999
				ICL	0.999	0.991	0.997	0.999	0.999	0.999
		0.7		CL	0.988	0.974	0.985	0.989	0.987	0.988
				ICL	0.975	0.991	0.966	0.976	0.975	0.975
		0.9		CL	0.966	0.907	0.948	0.967	0.962	0.966
				ICL	0.912	0.832	0.890	0.913	0.909	0.911
36	1.0	0.0	0.1	CL	0,998	0.997	0.998	0.998	0.998	0.998
				ICL	0.994	0.986	0.992	0.994	0.993	0.994
		0.1		CL	0.998	0.996	0.998	0.998	0.998	0.998
				ICL	0.992	0.983	0.990	0.992	0.991	0.992
		0.2		CL	0.992	0.986	0.989	0.992	0.991	0.992
				ICL	0.978	0.964	0.972	0.978	0.978	0.978
		0.5		CL	0.931	0.895	0.917	0.932	0.928	0.931
				ICL	0.886	0.853	0.876	0.888	0.885	0.886
		0.7		CL	0.873	0.829	0.858	0.875	0.870	0.873
				ICL	0.832	0.773	0.813	0.834	0.831	0.832
		0.9		CL	0.840	0.760	0.811	0.844	0.836	0.839
				ICL	0.767	0.705	0.745	0.768	0.766	0.767

Table 3.4: The PCS for the SM2 model (when SM1 and SM2 are competing models) for $\sigma = 5$, $n = 36$, $l_0 = 100$, $b_0 = 5$ and different values of exponential smoothing parameters α_1 and α_2.

n	σ	α_1	α_2	LF	AIC	BIC	HQ	MCp	GCV	FPE
36	5.0	0.0	0.01	CL	0.994	0.998	0.988	0.992	0.993	0.994
				ICL	0.979	0.972	0.977	0.979	0.979	0.979
		0.1		CL	0.991	0.989	0.990	0.991	0.990	0.991
				ICL	0.981	0.970	0.981	0.981	0.981	0.981
		0.2		CL	0.982	0.967	0.977	0.982	0.982	0.982
				ICL	0.968	0.949	0.961	0.970	0.967	0.968
		0.5		CL	0.928	0.895	0.914	0.928	0.926	0.928
				ICL	0.899	0.873	0.887	0.901	0.897	0.899
		0.7		CL	0.897	0.846	0.881	0.900	0.895	0.897
				ICL	0.851	0.785	0.833	0.853	0.848	0.851
		0.9		CL	0.864	0.816	0.847	0.868	0.860	0.864
				ICL	0.820	0.753	0.801	0.823	0.817	0.820
36	5.0	0.0	0.1	CL	0.978	0.963	0.974	0.979	0.977	0.978
				ICL	0.958	0.929	0.951	0.958	0.957	0.958
		0.1		CL	0.976	0.962	0.971	0.976	0.976	0.976
				ICL	0.949	0.916	0.943	0.951	0.947	0.949
		0.2		CL	0.966	0.949	0.963	0.968	0.966	0.966
				ICL	0.920	0.872	0.903	0.920	0.918	0.920
		0.5		CL	0.849	0.770	0.824	0.852	0.846	0.849
				ICL	0.753	0.679	0.732	0.753	0.749	0.753
		0.7		CL	0.760	0.653	0.722	0.762	0.753	0.760
				ICL	0.651	0.591	0.626	0.655	0.645	0.650
		0.9		CL	0.703	0.601	0.657	0.704	0.701	0.703
				ICL	0.604	0.514	0.573	0.606	0.599	0.604

Table 3.5: The estimated APCS and OAPCS (where indicated) when competing models are SM1 and SM2 with $\sigma = 1$, 5 and 20, $n = 24$ and 48, $l_0 = 100$ and $b_0 = 5$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
24	1	CL	SM1	0.615	0.695	0.641	0.610	0.625	0.615
			SM2	0.973	0.964	0.972	0.976	0.973	0.973
		ICL	SM1	0.663	0.748	0.686	0.652	0.670	0.664
			SM2	0.966	0.956	0.964	0.968	0.964	0.966
	OAPCS	CL		0.798	0.830	0.807	0.793	0.799	0.794
		ICL		0.815	0.852	0.825	0.810	0.817	0.815
	5	CL	SM1	0.611	0.697	0.638	0.606	0.623	0.611
			SM2	0.849	0.804	0.831	0.854	0.839	0.849
		ICL	SM1	0.690	0.766	0.712	0.679	0.698	0.691
			SM2	0.814	0.782	0.801	0.815	0.810	0.813
	OAPCS	CL		0.730	0.751	0.735	0.730	0.731	0.730
		ICL		0.752	0.774	0.757	0.747	0.754	0.751
	20	CL	SM1	0.612	0.697	0.639	0.607	0.624	0.612
			SM2	0.822	0.767	0.809	0.825	0.817	0.822
		ICL	SM1	0.690	0.768	0.711	0.680	0.698	0.691
			SM2	0.793	0.740	0.781	0.798	0.789	0.793
	OAPCS	CL		0.717	0.732	0.724	0.716	0.721	0.717
		ICL		0.742	0.754	0.746	0.739	0.744	0.742
48	1	CL	SM1	0.734	0.841	0.788	0.731	0.751	0.734
			SM2	0.992	0.989	0.992	0.992	0.992	0.992
		ICL	SM1	0.769	0.884	0.821	0.766	0.775	0.769
			SM2	0.991	0.970	0.991	0.991	0.991	0.991
	OAPCS	CL		0.863	0.905	0.890	0.862	0.872	0.863
		ICL		0.880	0.917	0.906	0.879	0.883	0.880
	5	CL	SM1	0.734	0.862	0.789	0.730	0.751	0.734
			SM2	0.903	0.874	0.893	0.903	0.902	0.900
		ICL	SM1	0.770	0.885	0.822	0.766	0.777	0.770
			SM2	0.890	0.863	0.879	0.893	0.892	0.893
	OAPCS	CL		0.812	0.868	0.841	0.817	0.827	0.819
		ICL		0.830	0.874	0.851	0.830	0.835	0.832
	20	CL	SM1	0.737	0.861	0.789	0.733	0.753	0.737
			SM2	0.876	0.824	0.853	0.878	0.874	0.876
		ICL	SM1	0.771	0.885	0.822	0.768	0.777	0.771
			SM2	0.859	0.804	0.836	0.863	0.857	0.859
	OAPCS	CL		0.807	0.843	0.821	0.806	0.814	0.807
		ICL		0.815	0.845	0.829	0.816	0.817	0.815

Table 3.6: The estimated APCS and OAPCS (where indicated) when competing models are SM1 and SM2 with $\sigma = 1$, 5 and 20, $n = 72$ and 96, $l_0 = 100$ and $b_0 = 5$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
72	1	CL	SM1	0.750	0.882	0.825	0.744	0.756	0.750
			SM2	0.986	0.982	0.985	0.986	0.986	0.986
		ICL	SM1	0.806	0.910	0.862	0.804	0.811	0.806
			SM2	0.984	0.980	0.983	0.984	0.984	0.984
	OAPCS	CL		0.868	0.901	0.905	0.865	0.871	0.868
		ICL		0.895	0.945	0.923	0.894	0.898	0.895
	5	CL	SM1	0.751	0.882	0.826	0.745	0.758	0.751
			SM2	0.927	0.893	0.913	0.928	0.927	0.927
		ICL	SM1	0.808	0.910	0.862	0.805	0.811	0.808
			SM2	0.915	0.884	0.900	0.916	0.912	0.915
	OAPCS	CL		0.839	0.888	0.870	0.837	0.843	0.839
		ICL		0.862	0.897	0.881	0.861	0.862	0.862
	20	CL	SM1	0.752	0.882	0.826	0.746	0.758	0.752
			SM2	0.894	0.846	0.878	0.894	0.894	0.894
		ICL	SM1	0.808	0.910	0.862	0.805	0.812	0.808
			SM2	0.877	0.831	0.852	0.877	0.875	0.876
	OAPCS	CL		0.823	0.864	0.852	0.820	0.826	0.823
		ICL		0.843	0.871	0.857	0.841	0.844	0.842
96	1	CL	SM1	0.773	0.906	0.846	0.771	0.777	0.773
			SM2	0.994	0.993	0.994	0.994	0.994	0.994
		ICL	SM1	0.851	0.946	0.905	0.850	0.853	0.856
			SM2	0.994	0.989	0.992	0.994	0.994	0.994
	OAPCS	CL		0.884	0.950	0.920	0.883	0.886	0.884
		ICL		0.923	0.968	0.949	0.922	0.924	0.923
	5	CL	SM1	0.772	0.906	0.845	0.771	0.776	0.772
			SM2	0.946	0.917	0.934	0.946	0.946	0.946
		ICL	SM1	0.851	0.946	0.905	0.850	0.854	0.851
			SM2	0.932	0.901	0.918	0.933	0.932	0.932
	OAPCS	CL		0.859	0.912	0.890	0.859	0.861	0.859
		ICL		0.892	0.924	0.912	0.892	0.893	0.892
	20	CL	SM1	0.773	0.906	0.845	0.772	0.777	0.773
			SM2	0.921	0.884	0.906	0.921	0.920	0.921
		ICL	SM1	0.852	0.946	0.906	0.851	0.854	0.852
			SM2	0.909	0.871	0.894	0.909	0.909	0.909
	OAPCS	CL		0.847	0.895	0.876	0.847	0.849	0.847
		ICL		0.881	0.909	0.900	0.880	0.882	0.881

Table 3.7: The estimated APCS and OAPCS (where indicated) when competing models are SM1 and SM2 with $\sigma = 1$, 5 and 20, $n = 120$ and 200, $l_0 = 100$ and $b_0 = 5$.

n	σ	LF	DGP	AIC	BIC	HQ	MC_P	GCV	FPE
120	1	CL	SM1	0.774	0.909	0.847	0.774	0.775	0.774
			SM2	0.995	0.992	0.995	0.995	0.995	0.995
		ICL	SM1	0.844	0.943	0.903	0.843	0.844	0.844
			SM2	0.993	0.991	0.992	0.993	0.993	0.993
	OAPCS	CL		0.885	0.951	0.921	0.885	0.885	0.885
		ICL		0.919	0.967	0.948	0.918	0.919	0.919
	5	CL	SM1	0.774	0.911	0.847	0.774	0.775	0.774
			SM2	0.953	0.935	0.947	0.953	0.953	0.953
		ICL	SM1	0.844	0.943	0.902	0.844	0.844	0.844
			SM2	0.946	0.928	0.937	0.946	0.946	0.946
	OAPCS	CL		0.864	0.923	0.897	0.864	0.864	0.864
		ICL		0.895	0.936	0.920	0.895	0.895	0.895
	20	CL	SM1	0.774	0.911	0.847	0.773	0.775	0.774
			SM2	0.938	0.920	0.927	0.938	0.936	0.938
		ICL	SM1	0.845	0.943	0.905	0.845	0.845	0.845
			SM2	0.922	0.898	0.910	0.922	0.922	0.922
	OAPCS	CL		0.856	0.916	0.887	0.856	0.856	0.856
		ICL		0.884	0.921	0.908	0.884	0.884	0.884
200	1	CL	SM1	0.807	0.923	0.870	0.806	0.808	0.807
			SM2	1.000	0.999	1.000	1.000	1.000	1.000
		ICL	SM1	0.899	0.956	0.925	0.899	0.901	0.899
			SM2	1.000	0.999	1.000	1.000	1.000	1.000
	OAPCS	CL		0.904	0.961	0.935	0.903	0.904	0.904
		ICL		0.950	0.978	0.963	0.950	0.951	0.950
	5	CL	SM1	0.808	0.922	0.868	0.807	0.810	0.808
			SM2	0.961	0.951	0.958	0.961	0.961	0.961
		ICL	SM1	0.900	0.956	0.925	0.900	0.902	0.990
			SM2	0.955	0.942	0.952	0.955	0.955	0.955
	OAPCS	CL		0.884	0.937	0.913	0.884	0.886	0.885
		ICL		0.928	0.949	0.939	0.928	0.929	0.928
	20	CL	SM1	0.808	0.920	0.868	0.807	0.810	0.808
			SM2	0.940	0.921	0.933	0.940	0.940	0.940
		ICL	SM1	0.898	0.956	0.925	0.898	0.901	0.898
			SM2	0.925	0.907	0.914	0.926	0.925	0.925
	OAPCS	CL		0.874	0.921	0.901	0.874	0.875	0.874
		ICL		0.912	0.932	0.920	0.912	0.913	0.912

Table 3.8: The estimated APCS and OAPCS (where indicated) when competing models are SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 24$ and 48, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
24	1	CL	SM2	0.411	0.595	0.457	0.384	0.440	0.414
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM2	0.468	0.610	0.502	0.449	0.489	0.468
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.706	0.798	0.729	0.692	0.720	0.707
		ICL		0.734	0.805	0.751	0.725	0.745	0.734
	5	CL	SM2	0.420	0.598	0.463	0.392	0.448	0.423
			SM3	0.905	0.841	0.888	0.915	0.894	0.905
		ICL	SM2	0.467	0.617	0.498	0.448	0.487	0.467
			SM3	0.906	0.838	0.884	0.908	0.892	0.904
	OAPCS	CL		0.663	0.720	0.676	0.654	0.671	0.664
		ICL		0.687	0.728	0.691	0.678	0.690	0.686
	20	CL	SM2	0.423	0.600	0.462	0.394	0.449	0.426
			SM3	0.734	0.592	0.695	0.757	0.703	0.731
		ICL	SM2	0.467	0.605	0.498	0.450	0.487	0.467
			SM3	0.706	0.585	0.673	0.727	0.680	0.702
	OAPCS	CL		0.579	0.596	0.579	0.576	0.576	0.579
		ICL		0.587	0.595	0.586	0.589	0.584	0.585
48	1	CL	SM2	0.441	0.692	0.576	0.427	0.467	0.441
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM2	0.484	0.718	0.601	0.467	0.500	0.485
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.721	0.846	0.788	0.714	0.734	0.721
		ICL		0.742	0.859	0.801	0.734	0.750	0.743
	5	CL	SM2	0.443	0.695	0.580	0.433	0.467	0.443
			SM3	0.956	0.925	0.947	0.956	0.955	0.956
		ICL	SM2	0.482	0.722	0.600	0.466	0.499	0.483
			SM3	0.956	0.906	0.943	0.956	0.952	0.956
	OAPCS	CL		0.699	0.810	0.764	0.695	0.711	0.699
		ICL		0.719	0.814	0.772	0.711	0.726	0.720
	20	CL	SM2	0.443	0.695	0.580	0.433	0.467	0.443
			SM3	0.806	0.634	0.735	0.812	0.792	0.806
		ICL	SM2	0.476	0.722	0.596	0.462	0.494	0.478
			SM3	0.795	0.613	0.719	0.800	0.785	0.795
	OAPCS	CL		0.625	0.665	0.658	0.623	0.630	0.625
		ICL		0.636	0.668	0.658	0.631	0.640	0.637

Table 3.9: The estimated APCS and OAPCS (where indicated) when competing models are SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 72$ and 96, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
72	1	CL	SM2	0.443	0.750	0.597	0.434	0.460	0.443
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM2	0.495	0.773	0.632	0.485	0.509	0.495
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.722	0.875	0.799	0.717	0.730	0.722
		ICL		0.748	0.887	0.816	0.743	0.755	0.748
	5	CL	SM2	0.445	0.750	0.595	0.435	0.461	0.445
			SM3	0.978	0.954	0.973	0.978	0.978	0.978
		ICL	SM2	0.503	0.776	0.639	0.493	0.517	0.503
			SM3	0.976	0.950	0.965	0.976	0.973	0.976
	OAPCS	CL		0.712	0.852	0.784	0.706	0.720	0.712
		ICL		0.740	0.863	0.802	0.735	0.745	0.740
	20	CL	SM2	0.446	0.750	0.597	0.437	0.461	0.446
			SM3	0.863	0.734	0.811	0.868	0.859	0.863
		ICL	SM2	0.502	0.778	0.639	0.492	0.517	0.502
			SM3	0.836	0.687	0.786	0.838	0.834	0.836
	OAPCS	CL		0.655	0.744	0.691	0.653	0.660	0.655
		ICL		0.669	0.733	0.713	0.665	0.676	0.669
96	1	CL	SM2	0.439	0.765	0.616	0.429	0.450	0.439
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM2	0.511	0.799	0.658	0.505	0.524	0.511
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.720	0.883	0.808	0.715	0.725	0.720
		ICL		0.756	0.900	0.829	0.753	0.762	0.756
	5	CL	SM2	0.441	0.768	0.616	0.432	0.451	0.441
			SM3	0.981	0.969	0.977	0.981	0.981	0.981
		ICL	SM2	0.511	0.799	0.657	0.504	0.525	0.512
			SM3	0.986	0.992	0.968	0.980	0.986	0.986
	OAPCS	CL		0.711	0.869	0.797	0.707	0.716	0.771
		ICL		0.749	0.900	0.813	0.742	0.756	0.749
	20	CL	SM2	0.440	0.767	0.616	0.432	0.452	0.440
			SM3	0.898	0.772	0.842	0.899	0.898	0.898
		ICL	SM2	0.510	0.796	0.656	0.504	0.523	0.510
			SM3	0.868	0.744	0.813	0.872	0.867	0.868
	OAPCS	CL		0.669	0.771	0.729	0.666	0.675	0.669
		ICL		0.689	0.770	0.735	0.688	0.695	0.689

Table 3.10: The estimated APCS and OAPCS (where indicated) when competing models are SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 120$ and 200, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
120	1	CL	SM2	0.472	0.796	0.656	0.463	0.480	0.472
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM2	0.553	0.824	0.693	0.548	0.555	0.553
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.735	0.898	0.828	0.732	0.740	0.736
		ICL		0.777	0.912	0.847	0.774	0.778	0.777
	5	CL	SM2	0.469	0.794	0.655	0.462	0.477	0.469
			SM3	0.985	0.973	0.982	0.985	0.985	0.985
		ICL	SM2	0.548	0.825	0.690	0.543	0.550	0.548
			SM3	0.987	0.977	0.981	0.987	0.987	0.987
	OAPCS	CL		0.727	0.884	0.819	0.724	0.731	0.727
		ICL		0.768	0.901	0.836	0.765	0.769	0.768
	20	CL	SM2	0.469	0.794	0.654	0.462	0.477	0.469
			SM3	0.926	0.829	0.883	0.928	0.925	0.926
		ICL	SM2	0.550	0.825	0.692	0.545	0.552	0.550
			SM3	0.903	0.801	0.866	0.903	0.902	0.903
	OAPCS	CL		0.698	0.812	0.769	0.695	0.701	0.698
		ICL		0.727	0.813	0.779	0.724	0.727	0.727
200	1	CL	SM2	0.489	0.829	0.680	0.486	0.497	0.489
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM2	0.537	0.861	0.707	0.536	0.544	0.537
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.745	0.915	0.840	0.743	0.749	0.745
		ICL		0.769	0.931	0.854	0.768	0.772	0.769
	5	CL	SM2	0.493	0.829	0.680	0.490	0.499	0.493
			SM3	0.986	0.983	0.985	0.986	0.986	0.986
		ICL	SM2	0.537	0.862	0.709	0.536	0.543	0.537
			SM3	0.992	0.989	0.990	0.992	0.992	0.992
	OAPCS	CL		0.740	0.906	0.833	0.738	0.743	0.740
		ICL		0.765	0.926	0.850	0.764	0.768	0.765
	20	CL	SM2	0.492	0.831	0.680	0.489	0.499	0.492
			SM3	0.931	0.863	0.898	0.931	0.931	0.931
		ICL	SM2	0.538	0.861	0.710	0.536	0.544	0.538
			SM3	0.933	0.855	0.906	0.933	0.932	0.933
	OAPCS	CL		0.712	0.847	0.789	0.710	0.715	0.712
		ICL		0.736	0.858	0.808	0.735	0.738	0.736

Table 3.11: The estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $\sigma = 1, 5$ and 20, $n = 24$, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
24	1	CL	SM1	0.390	0.571	0.452	0.379	0.425	0.392
			SM2	0.399	0.570	0.443	0.373	0.427	0.402
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM1	0.489	0.669	0.539	0.484	0.531	0.490
			SM3	0.451	0.593	0.484	0.430	0.471	0.450
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.596	0.714	0.632	0.584	0.617	0.598
		ICL		0.647	0.754	0.674	0.638	0.667	0.647
	5	CL	SM1	0.391	0.574	0.452	0.381	0.426	0.393
			SM2	0.363	0.480	0.387	0.338	0.381	0.365
			SM3	0.900	0.810	0.878	0.910	0.885	0.900
		ICL	SM1	0.461	0.714	0.547	0.502	0.533	0.514
			SM2	0.403	0.495	0.411	0.378	0.407	0.391
			SM3	0.896	0.795	0.868	0.898	0.878	0.894
	OAPCS	CL		0.551	0.621	0.572	0.543	0.564	0.553
		ICL		0.577	0.654	0.609	0.593	0.606	0.600
	20	CL	SM1	0.392	0.575	0.453	0.381	0.428	0.394
			SM2	0.355	0.466	0.384	0.332	0.377	0.358
			SM3	0.702	0.536	0.659	0.722	0.672	0.699
		ICL	SM1	0.507	0.679	0.553	0.497	0.530	0.510
			SM2	0.379	0.466	0.399	0.368	0.395	0.379
			SM3	0.659	0.528	0.626	0.679	0.639	0.656
	OAPCS	CL		0.483	0.526	0.499	0.478	0.492	0.482
		ICL		0.515	0.558	0.526	0.515	0.521	0.515

Table 3.12: The estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 48$, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
48	1	CL	SM1	0.555	0.816	0.678	0.546	0.572	0.555
			SM2	0.437	0.684	0.571	0.423	0.463	0.437
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM1	0.580	0.848	0.695	0.574	0.605	0.580
			SM2	0.479	0.710	0.594	0.462	0.495	0.480
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.664	0.833	0.750	0.656	0.678	0.664
		ICL		0.686	0.852	0.763	0.679	0.700	0.687
	5	CL	SM1	0.557	0.819	0.680	0.548	0.574	0.557
			SM2	0.398	0.586	0.502	0.390	0.420	0.398
			SM3	0.954	0.918	0.945	0.954	0.953	0.954
		ICL	SM1	0.578	0.847	0.696	0.573	0.605	0.578
			SM2	0.461	0.651	0.557	0.457	0.475	0.462
			SM3	0.955	0.900	0.941	0.955	0.951	0.955
	OAPCS	CL		0.636	0.774	0.709	0.631	0.649	0.636
		ICL		0.665	0.800	0.731	0.662	0.677	0.665
	20	CL	SM1	0.557	0.819	0.680	0.548	0.574	0.557
			SM2	0.382	0.552	0.483	0.374	0.399	0.382
			SM3	0.784	0.600	0.707	0.789	0.766	0.783
		ICL	SM1	0.578	0.843	0.693	0.571	0.605	0.578
			SM2	0.440	0.621	0.531	0.427	0.453	0.441
			SM3	0.770	0.586	0.696	0.777	0.761	0.770
	OAPCS	CL		0.574	0.657	0.623	0.570	0.580	0.574
		ICL		0.596	0.683	0.640	0.592	0.606	0.596

Table 3.13: The estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 72$, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
72	1	CL	SM1	0.624	0.876	0.762	0.621	0.636	0.624
			SM2	0.434	0.737	0.588	0.425	0.451	0.434
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM1	0.608	0.862	0.759	0.604	0.614	0.600
			SM2	0.487	0.762	0.623	0.477	0.501	0.487
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.686	0.871	0.783	0.682	0.700	0.686
		ICL		0.698	0.875	0.794	0.694	0.705	0.696
	5	CL	SM1	0.626	0.877	0.764	0.623	0.635	0.626
			SM2	0.415	0.670	0.544	0.405	0.428	0.415
			SM3	0.976	0.946	0.968	0.976	0.976	0.976
		ICL	SM1	0.607	0.892	0.758	0.603	0.614	0.607
			SM2	0.472	0.692	0.593	0.463	0.485	0.472
			AM3	0.975	0.945	0.961	0.975	0.972	0.975
	OAPCS	CL		0.672	0.831	0.759	0.668	0.680	0.672
		ICL		0.685	0.843	0.771	0.680	0.690	0.685
	20	CL	SM1	0.626	0.876	0.762	0.623	0.636	0.626
			SM2	0.406	0.627	0.515	0.398	0.418	0.405
			SM3	0.853	0.699	0.793	0.857	0.849	0.853
		ICL	SM1	0.613	0.894	0.762	0.607	0.620	0.613
			SM2	0.472	0.670	0.582	0.463	0.487	0.472
			SM3	0.827	0.659	0.771	0.827	0.825	0.827
	OAPCS	CL		0.628	0.734	0.690	0.626	0.634	0.628
		ICL		0.644	0.748	0.712	0.639	0.651	0.644

Table 3.14: The estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 96$, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
96	1	CL	SM1	0.695	0.919	0.822	0.692	0.698	0.695
			SM2	0.436	0.757	0.611	0.426	0.447	0.436
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM1	0.637	0.916	0.784	0.634	0.642	0.637
			SM2	0.506	0.793	0.653	0.500	0.519	0.506
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.710	0.892	0.811	0.706	0.715	0.710
		ICL		0.714	0.903	0.812	0.711	0.720	0.714
	5	CL	SM1	0.695	0.920	0.820	0.692	0.698	0.695
			SM2	0.413	0.694	0.564	0.404	0.423	0.413
			SM3	0.981	0.964	0.970	0.981	0.981	0.981
		ICL	SM1	0.637	0.915	0.804	0.654	0.662	0.657
			SM2	0.498	0.751	0.633	0.491	0.512	0.499
			SM3	0.986	0.965	0.980	0.986	0.986	0.989
	OAPCS	CL		0.696	0.860	0.785	0.692	0.701	0.696
		ICL		0.707	0.877	0.806	0.710	0.720	0.714
	20	CL	SM1	0.696	0.919	0.821	0.693	0.699	0.696
			SM2	0.392	0.672	0.549	0.384	0.404	0.392
			SM3	0.879	0.749	0.821	0.880	0.879	0.879
		ICL	SM1	0.638	0.915	0.781	0.634	0.644	0.638
			SM2	0.496	0.732	0.626	0.490	0.508	0.496
			SM3	0.862	0.724	0.804	0.865	0.861	0.862
	OAPCS	CL		0.656	0.780	0.730	0.652	0.661	0.656
		ICL		0.669	0.794	0.741	0.667	0.675	0.669

Table 3.15: The estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 120$, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
120	1	CL	SM1	0.684	0.921	0.827	0.682	0.688	0.684
			SM2	0.469	0.789	0.651	0.460	0.477	0.469
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM1	0.625	0.907	0.787	0.623	0.627	0.625
			SM2	0.552	0.818	0.690	0.547	0.554	0.552
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.718	0.903	0.826	0.714	0.722	0.718
		ICL		0.726	0.908	0.826	0.723	0.727	0.726
	5	CL	SM1	0.685	0.922	0.827	0.683	0.690	0.685
			SM2	0.450	0.741	0.615	0.444	0.458	0.450
			SM3	0.984	0.970	0.981	0.984	0.984	0.984
		ICL	SM1	0.623	0.907	0.787	0.621	0.626	0.628
			SM2	0.545	0.792	0.671	0.540	0.547	0.545
			SM3	0.986	0.977	0.980	0.986	0.986	0.986
	OAPCS	CL		0.706	0.878	0.808	0.704	0.711	0.706
		ICL		0.718	0.892	0.813	0.716	0.720	0.718
	20	CL	SM1	0.685	0.922	0.829	0.683	0.689	0.685
			SM2	0.437	0.723	0.600	0.432	0.445	0.437
			SM3	0.916	0.814	0.871	0.918	0.914	0.916
		ICL	SM1	0.640	0.904	0.781	0.619	0.626	0.622
			SM2	0.537	0.783	0.662	0.532	0.538	0.537
			SM3	0.895	0.787	0.851	0.895	0.894	0.895
	OAPCS	CL		0.679	0.820	0.767	0.678	0.683	0.680
		ICL		0.677	0.825	0.765	0.682	0.686	0.685

Table 3.16: The estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $\sigma = 1$, 5 and 20, $n = 200$, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
200	1	CL	SM1	0.773	0.944	0.877	0.770	0.775	0.774
			SM2	0.489	0.828	0.680	0.486	0.497	0.489
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	SM1	0.659	0.941	0.837	0.659	0.665	0.659
			SM2	0.537	0.860	0.707	0.536	0.544	0.537
			SM3	1.000	1.000	1.000	1.000	1.000	1.000
	OAPCS	CL		0.754	0.924	0.852	0.752	0.757	0.754
		ICL		0.732	0.934	0.848	0.732	0.736	0.732
	5	CL	SM1	0.773	0.944	0.877	0.771	0.775	0.773
			SM2	0.483	0.796	0.663	0.480	0.489	0.483
			SM3	0.986	0.983	0.985	0.986	0.986	0.986
		ICL	SM1	0.657	0.929	0.817	0.656	0.662	0.657
			SM2	0.528	0.833	0.694	0.527	0.534	0.528
			SM3	0.992	0.989	0.990	0.992	0.992	0.992
	OAPCS	CL		0.747	0.908	0.842	0.746	0.750	0.747
		ICL		0.736	0.927	0.844	0.735	0.740	0.736
	20	CL	SM1	0.773	0.944	0.877	0.769	0.775	0.773
			SM2	0.464	0.767	0.634	0.461	0.471	0.464
			SM3	0.925	0.853	0.893	0.925	0.925	0.925
		ICL	SM1	0.662	0.940	0.833	0.660	0.665	0.662
			SM2	0.514	0.800	0.669	0.512	0.520	0.514
			SM3	0.929	0.849	0.900	0.929	0.928	0.929
	OAPCS	CL		0.721	0.853	0.801	0.718	0.724	0.721
		ICL		0.712	0.873	0.811	0.711	0.714	0.712

Table 3.17: The average (taken on $\sigma = 1$, 5 and 20) of estimated APCS and OAPCS (where indicated) when competing models are SM1 and SM2 with $n = 24$, 48, 72, 96, 120 and 200, $l_0 = 100$ and $b_0 = 5$.

n	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
24	SM1	CL	0.613	0.696	0.639	0.608	0.624	0.613
		ICL	0.681	0.761	0.703	0.670	0.689	0.682
	SM2	CL	0.881	0.845	0.871	0.885	0.876	0.881
		ICL	0.858	0.826	0.849	0.860	0.854	0.857
Average		CL	0.747	0.771	0.755	0.746	0.750	0.747
OAPCS		ICL	0.769	0.793	0.776	0.765	0.772	0.770
48	SM1	CL	0.735	0.855	0.789	0.731	0.752	0.735
		ICL	0.770	0.885	0.822	0.767	0.776	0.770
	SM2	CL	0.924	0.896	0.913	0.924	0.923	0.924
		ICL	0.913	0.879	0.902	0.916	0.913	0.914
Average		CL	0.829	0.875	0.851	0.828	0.837	0.829
OAPCS		ICL	0.842	0.882	0.862	0.841	0.845	0.842
72	SM1	CL	0.751	0.882	0.826	0.745	0.757	0.751
		ICL	0.807	0.910	0.862	0.805	0.811	0.807
	SM2	CL	0.936	0.907	0.925	0.936	0.936	0.936
		ICL	0.925	0.898	0.912	0.926	0.924	0.925
Average		CL	0.843	0.895	0.876	0.841	0.847	0.843
OAPCS		ICL	0.866	0.904	0.887	0.865	0.868	0.866
96	SM1	CL	0.773	0.906	0.845	0.771	0.777	0.773
		ICL	0.851	0.946	0.905	0.850	0.854	0.851
	SM2	CL	0.954	0.931	0.945	0.954	0.953	0.954
		ICL	0.945	0.920	0.935	0.945	0.945	0.945
Average		CL	0.863	0.919	0.895	0.863	0.865	0.863
OAPCS		ICL	0.898	0.933	0.920	0.898	0.899	0.898
120	SM1	CL	0.774	0.910	0.847	0.774	0.775	0.774
		ICL	0.844	0.943	0.903	0.844	0.844	0.844
	SM2	CL	0.962	0.949	0.956	0.962	0.961	0.962
		ICL	0.954	0.939	0.946	0.954	0.954	0.954
Average		CL	0.868	0.930	0.902	0.868	0.868	0.868
OAPCS		ICL	0.899	0.941	0.925	0.899	0.899	0.899
200	SM1	CL	0.808	0.922	0.869	0.807	0.809	0.808
		ICL	0.899	0.956	0.925	0.899	0.901	0.929
	SM2	CL	0.967	0.957	0.964	0.967	0.967	0.967
		ICL	0.960	0.949	0.955	0.960	0.960	0.960
Average		CL	0.887	0.939	0.916	0.887	0.888	0.887
OAPCS		ICL	0.930	0.953	0.940	0.930	0.931	0.945

Table 3.18: The average (taken on $\sigma = 1$, 5 and 20) of estimated APCS and OAPCS (where indicated) when competing models are SM2 and SM3 with $n = 24$, 48, 72 96, 120 and 200, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
24	SM2	CL	0.418	0.598	0.461	0.387	0.446	0.421
		ICL	0.473	0.611	0.499	0.449	0.488	0.467
	SM3	CL	0.880	0.811	0.861	0.891	0.866	0.879
		ICL	0.871	0.808	0.852	0.878	0.857	0.869
Average		CL	0.649	0.704	0.661	0.639	0.656	0.650
OAPCS		ICL	0.672	0.709	0.676	0.664	0.673	0.668
48	SM2	CL	0.442	0.598	0.579	0.431	0.467	0.442
		ICL	0.481	0.611	0.599	0.465	0.498	0.482
	SM3	CL	0.921	0.811	0.894	0.923	0.916	0.921
		ICL	0.917	0.808	0.887	0.919	0.912	0.917
Average		CL	0.682	0.774	0.736	0.677	0.691	0.682
OAPCS		ICL	0.699	0.780	0.743	0.692	0.705	0.700
72	SM2	CL	0.445	0.750	0.596	0.435	0.461	0.445
		ICL	0.500	0.776	0.637	0.490	0.514	0.500
	SM3	CL	0.947	0.896	0.928	0.949	0.946	0.947
		ICL	0.937	0.879	0.917	0.938	0.936	0.937
Average		CL	0.696	0.823	0.762	0.692	0.703	0.696
OAPCS		ICL	0.719	0.827	0.777	0.714	0.725	0.719
96	SM2	CL	0.440	0.767	0.616	0.431	0.451	0.440
		ICL	0.511	0.798	0.657	0.504	0.524	0.511
	SM3	CL	0.960	0.914	0.940	0.960	0.960	0.960
		ICL	0.951	0.912	0.927	0.951	0.951	0.951
Average		CL	0.700	0.840	0.778	0.696	0.705	0.700
OAPCS		ICL	0.731	0.855	0.792	0.728	0.738	0.73
120	SM2	CL	0.470	0.795	0.655	0.462	0.478	0.470
		ICL	0.550	0.825	0.692	0.545	0.552	0.550
	SM3	CL	0.970	0.934	0.955	0.971	0.970	0.970
		ICL	0.963	0.926	0.949	0.963	0.963	0.963
Average		CL	0.720	0.864	0.805	0.717	0.724	0.720
OAPCS		ICL	0.757	0.875	0.820	0.754	0.758	0.757
200	SM2	CL	0.491	0.830	0.680	0.488	0.498	0.491
		ICL	0.537	0.861	0.709	0.536	0.544	0.537
	SM3	CL	0.972	0.949	0.961	0.972	0.972	0.972
		ICL	0.975	0.948	0.965	0.975	0.975	0.975
Average		CL	0.732	0.889	0.821	0.730	0.735	0.732
OAPCS		ICL	0.756	0.905	0.837	0.756	0.759	0.756

Table 3.19: The average (taken on $\sigma = 1$, 5 and 20) of estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $n = 24$, 48 and 72, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
24	SM1	CL	0.391	0.573	0.452	0.380	0.426	0.393
		ICL	0.486	0.687	0.546	0.494	0.531	0.505
	SM2	CL	0.372	0.505	0.405	0.348	0.395	0.375
		ICL	0.411	0.518	0.431	0.392	0.424	0.407
	SM3	CL	0.862	0.782	0.846	0.877	0.852	0.866
		ICL	0.852	0.774	0.831	0.859	0.839	0.850
Average		CL	0.544	0.620	0.568	0.535	0.558	0.545
OAPCS		ICL	0.583	0.660	0.603	0.582	0.598	0.587
48	SM1	CL	0.556	0.818	0.679	0.547	0.573	0.556
		ICL	0.579	0.846	0.695	0.573	0.605	0.579
	SM2	CL	0.406	0.607	0.519	0.396	0.427	0.406
		ICL	0.460	0.661	0.561	0.449	0.474	0.461
	SM3	CL	0.913	0.839	0.884	0.914	0.906	0.912
		ICL	0.908	0.829	0.879	0.911	0.904	0.908
Average		CL	0.625	0.755	0.694	0.619	0.636	0.625
OAPCS		ICL	0.649	0.778	0.711	0.644	0.661	0.649
72	SM1	CL	0.625	0.876	0.763	0.622	0.636	0.625
		ICL	0.609	0.883	0.760	0.605	0.616	0.607
	SM2	CL	0.418	0.678	0.549	0.409	0.432	0.418
		ICL	0.477	0.708	0.599	0.468	0.491	0.477
	SM3	CL	0.943	0.882	0.920	0.944	0.942	0.943
		ICL	0.934	0.868	0.911	0.934	0.932	0.934
Average		CL	0.662	0.812	0.744	0.659	0.670	0.662
OAPCS		ICL	0.673	0.820	0.757	0.669	0.680	0.673

Table 3.20: The average (taken on $\sigma = 1, 5$ and 20) of estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $n = 96$, 120 and 200, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

n	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
96	SM1	CL	0.695	0.919	0.821	0.692	0.698	0.695
		ICL	0.637	0.915	0.790	0.641	0.649	0.644
	SM2	CL	0.414	0.708	0.575	0.405	0.425	0.414
		ICL	0.500	0.759	0.637	0.494	0.513	0.500
	SM3	CL	0.953	0.904	0.930	0.954	0.953	0.953
			0.949	0.896	0.928	0.950	0.949	0.949
Average		CL	0.687	0.844	0.775	0.684	0.692	0.687
OAPCS		ICL	0.696	0.857	0.785	0.695	0.704	0.698
120	SM1	CL	0.685	0.922	0.828	0.683	0.689	0.685
		ICL	0.629	0.906	0.785	0.621	0.626	0.623
	SM2	CL	0.452	0.751	0.622	0.445	0.460	0.452
		ICL	0.545	0.798	0.674	0.540	0.546	0.545
	SM3	CL	0.967	0.928	0.951	0.967	0.966	0.967
		ICL	0.960	0.921	0.944	0.960	0.960	0.960
Average		CL	0.701	0.867	0.800	0.698	0.705	0.701
OAPCS		ICL	0.711	0.875	0.801	0.707	0.711	0.709
200	SM1	CL	0.773	0.944	0.877	0.770	0.775	0.773
		ICL	0.659	0.937	0.829	0.658	0.664	0.659
	SM2	CL	0.479	0.797	0.659	0.476	0.486	0.479
		ICL	0.526	0.831	0.690	0.525	0.533	0.526
	SM3	CL	0.970	0.945	0.959	0.970	0.970	0.970
		ICL	0.974	0.946	0.963	0.974	0.973	0.974
Average		CL	0.741	0.895	0.832	0.739	0.744	0.741
OAPCS		ICL	0.720	0.905	0.827	0.719	0.723	0.720

Table 3.21: The average (taken on sample sizes) of estimated APCS and OAPCS (where indicated) when competing models are SM1 and SM2 with $\sigma = 1$, 5 and 20, $l_0 = 100$ and $b_0 = 5$.

σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
1	SM1	CL	0.742	0.859	0.803	0.739	0.749	0.742
		ICL	0.805	0.898	0.850	0.802	0.809	0.806
	SM2	CL	0.990	0.987	0.990	0.991	0.990	0.990
		ICL	0.988	0.981	0.987	0.988	0.988	0.988
Average		CL	0.866	0.923	0.896	0.865	0.869	0.866
OAPCS		ICL	0.897	0.939	0.919	0.895	0.898	0.897
5	SM1	CL	0.742	0.863	0.802	0.739	0.749	0.74
		ICL	0.811	0.901	0.855	0.807	0.814	0.826
	SM2	CL	0.923	0.896	0.913	0.924	0.921	0.923
		ICL	0.909	0.883	0.898	0.910	0.908	0.909
Average		CL	0.832	0.880	0.857	0.832	0.835	0.832
OAPCS		ICL	0.860	0.892	0.876	0.859	0.861	0.867
20	SM1	CL	0.743	0.863	0.802	0.740	0.750	0.743
		ICL	0.811	0.901	0.855	0.808	0.815	0.811
	SM2	CL	0.899	0.860	0.884	0.899	0.897	0.899
		ICL	0.881	0.842	0.865	0.883	0.880	0.881
Average		CL	0.821	0.862	0.843	0.820	0.823	0.821
OAPCS		ICL	0.846	0.872	0.860	0.845	0.847	0.846

Table 3.22: The average (taken on sample sizes) of estimated APCS and OAPCS (where indicated) when competing models are SM2 and SM3 with $\sigma = 1$, 5 and 20, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
1	SM2	CL	0.449	0.738	0.597	0.436	0.466	0.450
		ICL	0.511	0.764	0.632	0.498	0.520	0.508
	SM3	CL	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	1.000	1.000	1.000	1.000	1.000	1.000
Average		CL	0.725	0.869	0.799	0.718	0.733	0.725
OAPCS		ICL	0.756	0.882	0.816	0.749	0.760	0.754
5	SM2	CL	0.452	0.739	0.598	0.441	0.467	0.452
		ICL	0.508	0.767	0.632	0.498	0.520	0.508
	SM3	CL	0.965	0.941	0.959	0.967	0.963	0.965
		ICL	0.967	0.942	0.955	0.967	0.964	0.967
Average		CL	0.709	0.840	0.778	0.704	0.715	0.709
OAPCS		ICL	0.738	0.854	0.794	0.732	0.742	0.738
20	SM2	CL	0.452	0.740	0.598	0.441	0.468	0.453
		ICL	0.507	0.765	0.632	0.498	0.520	0.508
	SM3	CL	0.860	0.738	0.811	0.866	0.851	0.859
		ICL	0.840	0.714	0.794	0.846	0.833	0.840
Average		CL	0.656	0.739	0.704	0.654	0.659	0.656
OAPCS		ICL	0.674	0.739	0.713	0.672	0.676	0.674

Table 3.23: The average (taken on sample sizes) of estimated APCS and OAPCS (where indicated) when competing models are SM1, SM2 and SM3 with $\sigma = 1$, 5 and 20, $l_0 = 100$, $b_0 = 5$ and $A = 0.3$.

σ	LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE
1	SM1	CL	0.620	0.841	0.736	0.615	0.632	0.621
		ICL	0.600	0.857	0.734	0.596	0.614	0.599
	SM2	CL	0.444	0.728	0.591	0.432	0.460	0.445
		ICL	0.502	0.756	0.625	0.492	0.514	0.502
	SM3	CL	1.000	1.000	1.000	1.000	1.000	1.000
		ICL	1.000	1.000	1.000	1.000	1.000	1.000
Average		CL	0.688	0.856	0.776	0.682	0.698	0.688
OAPCS		ICL	0.701	0.871	0.786	0.696	0.709	0.700
5	SM1	CL	0.621	0.843	0.737	0.616	0.633	0.629
		ICL	0.594	0.867	0.735	0.602	0.617	0.606
	SM2	CL	0.420	0.661	0.546	0.410	0.433	0.421
		ICL	0.485	0.702	0.593	0.476	0.493	0.483
	SM3	CL	0.964	0.932	0.955	0.965	0.961	0.964
		ICL	0.965	0.929	0.953	0.965	0.961	0.965
Average		CL	0.668	0.812	0.745	0.664	0.676	0.669
OAPCS		ICL	0.681	0.833	0.760	0.681	0.690	0.685
20	SM1	CL	0.622	0.843	0.737	0.616	0.634	0.622
		ICL	0.606	0.863	0.734	0.598	0.615	0.604
	SM2	CL	0.406	0.635	0.528	0.397	0.419	0.406
		ICL	0.473	0.679	0.578	0.465	0.484	0.473
	SM3	CL	0.843	0.709	0.791	0.849	0.834	0.843
		ICL	0.824	0.689	0.775	0.829	0.818	0.823
Average		CL	0.624	0.729	0.685	0.621	0.629	0.624
OAPCS		ICL	0.634	0.743	0.696	0.631	0.639	0.633

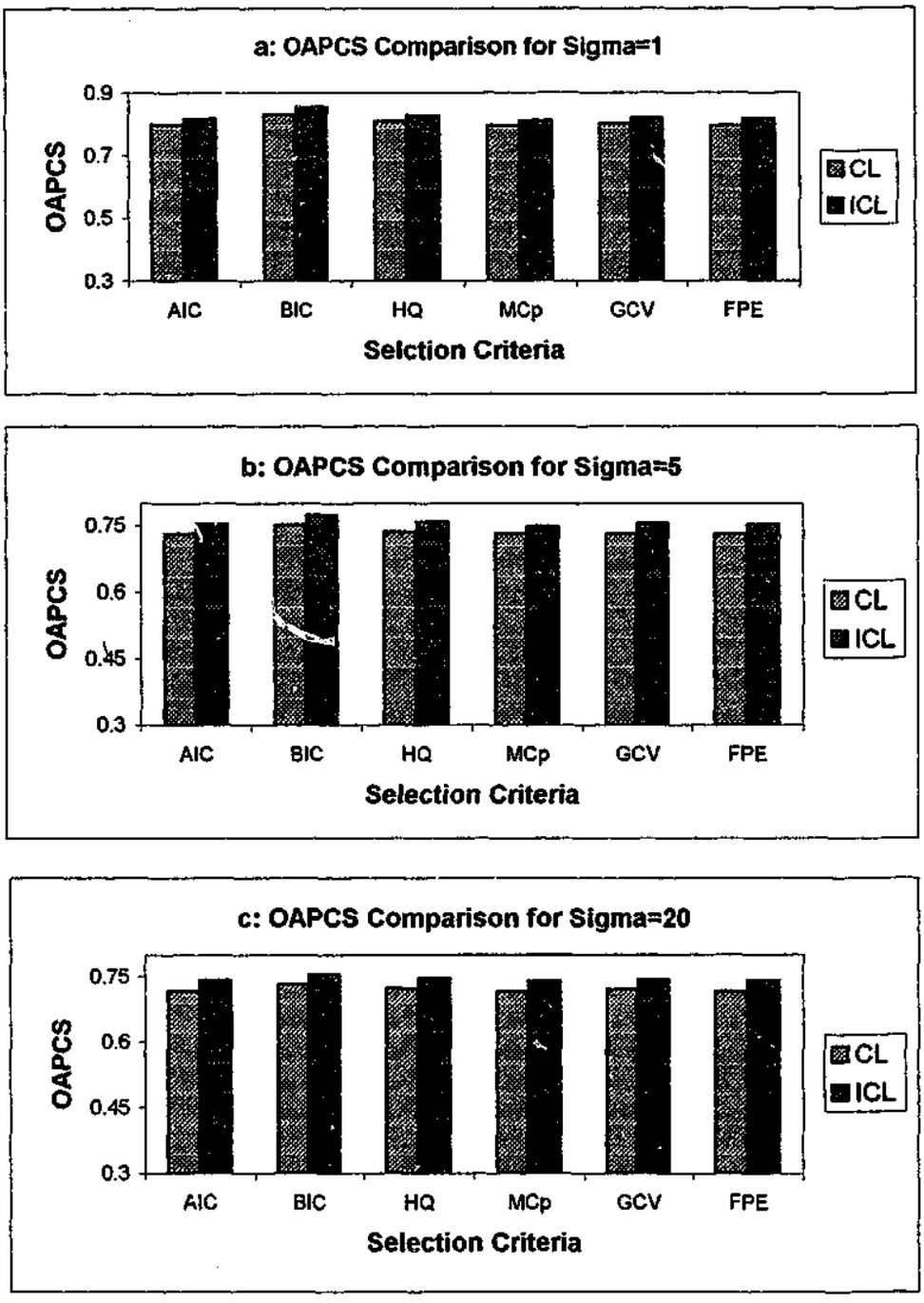

Figure 3.1: Comparison of the CL and ICL based IC procedures with respect to OAPCS when the competing models are SM1 and SM2, and $n = 24$.

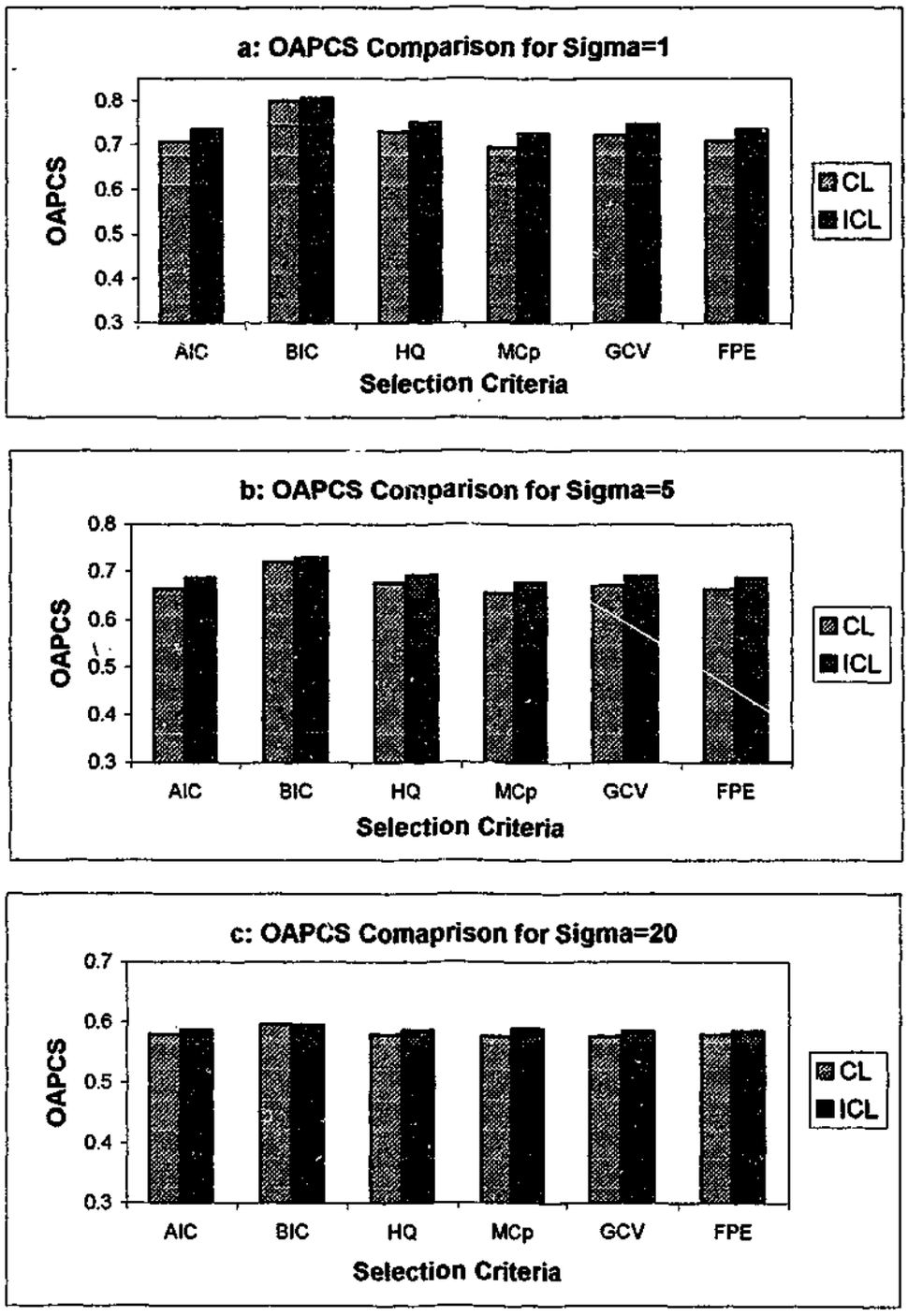

Figure 3.2: Comparison of the CL and ICL based IC procedures with respect to OAPCS when the competing models are SM2 and SM3, and $n = 24$.

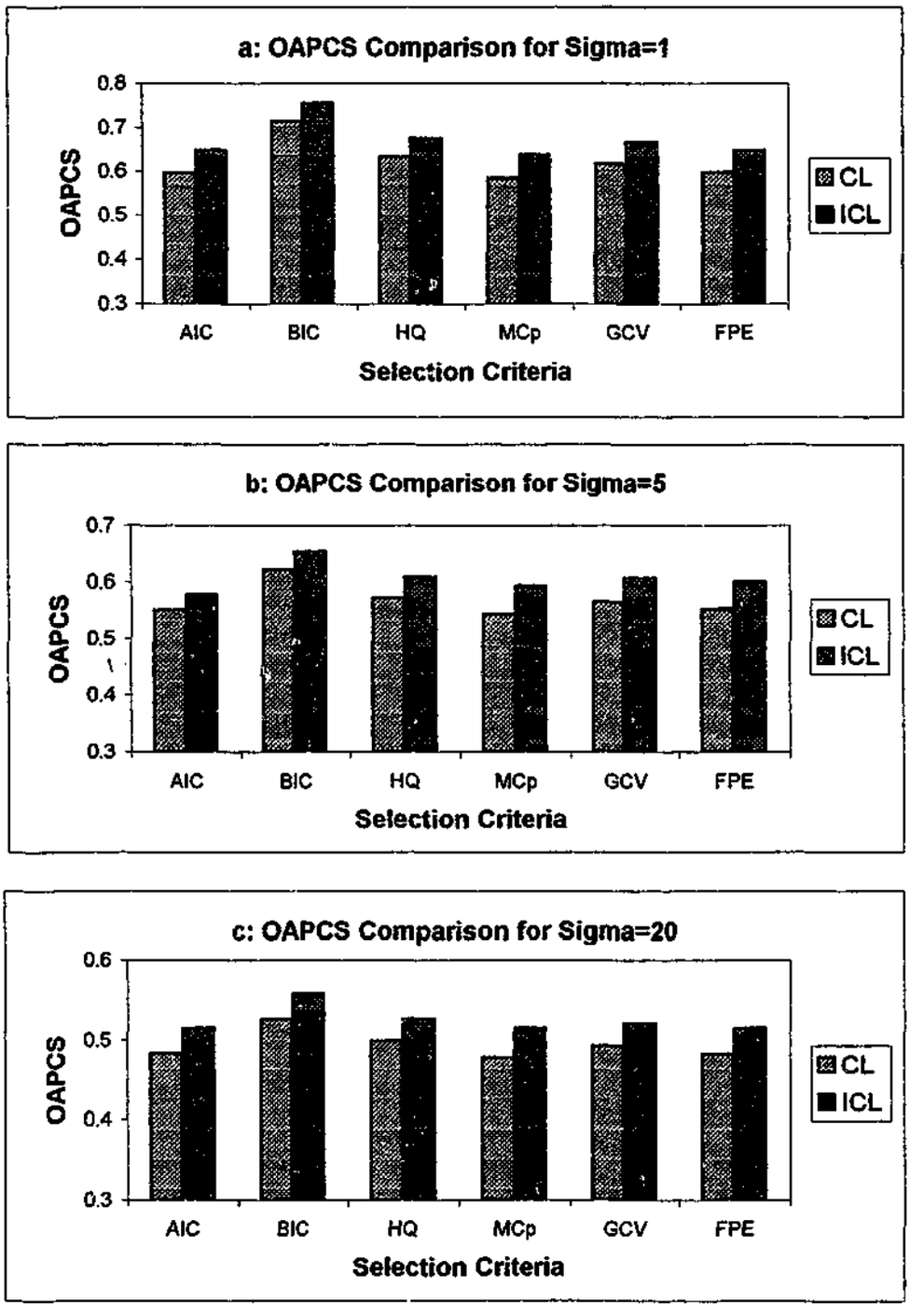

Figure 3.3: Comparison of the CL and ICL based IC procedures with respect to OAPCS when the competing models are SM1, SM2 and SM3, and $n = 24$.

Chapter 4

Using PEM for Regression Error Model Selection[1]

4.1 Introduction

As mentioned in Chapter 2, a number of Monte Carlo studies have been conducted to compare the small sample performance of various IC procedures, in order to select the correct or appropriate model from a set of models. These previous studies on small sample model selection show that there is no existing IC procedure which is uniformly better than the other criteria. With an emphasis on small sample performance, Hurvich and Tsai (1989) proposed bias corrected AIC (AICc) for regression and AR time series models and pointed out that this criterion will perform better when the sample size is small. However, the small sample performance of AICc is not consistent for all model selection problems.

In the context of choosing between first-order autoregressive (AR(1)) and first-order moving average (MA(1)) disturbances in the general linear regression model, Grose and King (1994) have shown that a particular model can be unfairly favored because of the shape or functional form of its log-likelihood. They also found that the presence of nuisance parameters can adversely affect the probabilities of correct

[1] A paper based on the material in this chapter has been published in the *Journal of Statistical Research*. See Billah and King (2000b).

selection and recommended the application of an IC procedure to the MGL rather than the classical likelihood. Their model selection problem is one in which both models have the same number of parameters. In this chapter, we address a related model selection problem of selecting from a range of different disturbance processes with differing numbers of parameters. In particular, we consider the problem of selecting between white noise (WN), AR(1), second-order AR (AR(2)) and MA(1) disturbances in the general linear regression model using profile likelihood (PL) as well as MGL. Note that a complete study of ARMA disturbance models under all possible IC procedures is not feasible because of the size of the task. Therefore, we have had to narrow the number of criterion and the number of models.

The aim of this chapter is to look at the application of OAPCS to time series model selection problems. We investigate a new approach called PEM, which involves the use of those penalty function values that maximizes OAPCS for the particular sample size and set of models under consideration. Simulation methods have been used to estimate the required APCS when different models are the true model. This means that the OAPCS which we wish to maximize, is a step function that might not always be particularly well behaved. This makes it very difficult to use standard methods to maximize the OAPCS. The PEM based on GS (PEM-GS) is one possible method, but sometimes the computation involved becomes unrealistic as the number of competing models increases. We, therefore, have turned to a relatively new global optimization algorithm called SA. As mentioned in Chapter 2, this algorithm performs well, even in the presence of a large number of local maxima. Because the algorithm accepts both up hill and down hill moves, transitions out of a local maximum are possible. Compared to classical methods, it requires less rigorous assumptions regarding the objective function and consequently functions like ours with ridges and plateaux can be dealt with more

easily. In this chapter, we modify the SA algorithm in the context of model selection problems so that the modified algorithm (PEM-SA) will estimate the penalty values which maximize OAPCS.

The plan of this chapter is as follows. The models and the two estimation methods we use are discussed in Section 4.2. The theory of the new model selection approach (PEM) is outlined in Section 4.3, while the penalty estimation algorithm is presented in Section 4.4. The design of the Monte Carlo study to evaluate PEM is given in Section 4.5, with the results of the study discussed in Section 4.6. The final section contains some concluding remarks.

4.2 The Model and the Methods of Estimation

Consider the linear regression model,

$$y = X\beta + u, \tag{4.2.1}$$

where y is an $n \times 1$ vector of observations, X is an $n \times k$ non-stochastic matrix of rank $k < n$, β is a $k \times 1$ unknown parameter vector and u is an $n \times 1$ disturbance vector such that $u \sim N(0, \sigma^2\Omega(\theta))$ in which $\Omega(\theta)$ is an $n \times n$ positive definite matrix function and θ is a $q \times 1$ vector of unknown parameters.

Several methods have been suggested in the vast array of literature for estimation of the model (4.2.1). Examples include the use of MGL (Kalbfleish and Sprott, 1970; Tunnicliffe Wilson, 1989; Ara, 1995 and Ara and King, 1993), the conditional profile likelihood (CPL; Cox and Reid, 1987 and Laskar and King, 1998) and the approximate conditional profile likelihood (ACPL; Cox and Reid, 1993). From this literature, it is evident that the MGL based estimates, tests and model selection procedures perform better than those based on any other likelihoods. However, for comparison, we also consider the PL.

We are interested in choosing between four different covariance matrices $\Omega(\theta)$ which result from WN, AR(1), MA(1) and AR(2) disturbance processes in the context of (4.2.1). The form of $\Omega(\theta)$ in these four cases is well known, see for example, Box and Jenkins (1976). The log-likelihood function of (4.2.1) is

$$\begin{aligned} & f(\beta, \sigma^2, \theta|y) \\ & = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2\sigma^2}(y - X\beta)'\Omega(\theta)^{-1}(y - X\beta). \end{aligned} \quad (4.2.2)$$

For any given value of θ, the values of β and σ^2 that maximize (4.2.2) are

$$\widehat{\beta}_\theta = \left(X'\Omega(\theta)^{-1}X\right)^{-1} X'\Omega(\theta)^{-1}y \quad (4.2.3)$$

and

$$\widehat{\sigma}^2_\theta = \frac{1}{n}(y - X\widehat{\beta}_\theta)'\Omega(\theta)^{-1}(y - X\widehat{\beta}_\theta). \quad (4.2.4)$$

If these estimates of β and σ^2 are substituted back into (4.2.2), we get the profile or concentrated log-likelihood:

$$f_p(\theta|y) = -\frac{n}{2}\log(2\pi\widehat{\sigma}^2_\theta) - \frac{1}{2}\log|\Omega(\theta)| - \frac{n}{2}. \quad (4.2.5)$$

The problem of maximizing (4.2.2) with respect to the unknown parameters β, σ^2 and θ therefore becomes one of maximizing the profile log-likelihood (4.2.5) with respect to θ. In the case of WN disturbances, $\Omega(\theta) = I_n$, the $n \times n$ identity matrix, and the maximized value of (4.2.2) is

$$-\frac{n}{2}\log(2\pi\widehat{\sigma}^2) - \frac{n}{2}$$

in which $\widehat{\sigma}^2$ is (4.2.4) with $\Omega(\theta)^{-1} = I_n$ and $\widehat{\beta}_\theta$ is replaced by the OLS estimator of β from (4.2.1).

The problem of choosing between the four different disturbance processes is invariant to transformations of the form

$$y \rightarrow \eta_0 y + X\eta, \quad (4.2.6)$$

where η_0 is a positive scalar and η is a $k \times 1$ vector. If $m = n - k$, $z = My$ is the OLS residual vector from (4.2.1), $M = I_n - X(X'X)^{-1}X'$ and P is an $m \times n$ matrix such that $P'P = M$ and $PP' = I_m$, then the $m \times 1$ vector

$$\nu = \frac{Pz}{(z'P'Pz)^{1/2}}$$

is a maximal invariant under the group of transformations of the form of (4.2.6). The principle of invariance allows one to work with the maximal invariant as though it is the observed data in constructing invariant tests. This is because all invariant test statistics can be written as functions of the maximal invariant (see Lehmann, 1991). Similarly, if we restrict our attention only to model selection procedures that are invariant to transformations of the form of (4.2.6), then, because the maximal invariant takes the same value for all sample points connected by a transformation (and which, hence, have the same outcome) and different values for points not connected by transformations, we can treat the maximal invariant as our observed data and build a model selection procedure on this basis.

The joint density function of ν can be shown (see King, 1980) to be,

$$h(\nu)d\nu = \frac{1}{2}\Gamma(m/2)\pi^{-m/2}\left|P\Omega(\theta)P'\right|^{-1/2}\left(\nu'(P\Omega(\theta)P')^{-1}\nu\right)^{-m/2}d\nu, \qquad (4.2.7)$$

where $d\nu$ denotes the uniform measure on the surface of the unit m-sphere. Note that when $\Omega(\theta) = I_n$, (4.2.7) reduces to

$$h(\nu)d\nu = \frac{1}{2}\Gamma(m/2)\pi^{-m/2}d\nu$$

which is the density of the uniform distribution on the surface of the unit m-sphere.

Two useful results in evaluating (4.2.7) are from King (1980, lemma 2)

$$\nu'(P\Omega(\theta)P')^{-1}\nu = \frac{\widehat{u}'_\theta\Omega(\theta)^{-1}\widehat{u}_\theta}{z'z},$$

in which $\widehat{u}_\theta$ is the generalized least squares (GLS) residual vector of (4.2.1) assuming covariance matrix $\sigma^2\Omega(\theta)$ and from Verbyla (1990)

$$|P\Omega(\theta)P'| = |X'X|^{-1}\,|\Omega(\theta)|\left|X'\Omega(\theta)^{-1}X\right|.$$

Therefore, an IC procedure for model selection of different disturbance processes in (4.2.1) can be based on treating (4.2.7) as the likelihood to be maximized and penalized. Ara and King (1993) have shown that the likelihood of θ based on (4.2.7) is equivalent to the MGL for θ from Tunnicliffe Wilson (1989) which is given by

$$h_m(\theta|y) = |\Omega(\theta)|^{-1/2}\left|X'\Omega(\theta)^{-1}X\right|^{-1/2}\left(\widehat{u}_\theta'\Omega(\theta)^{-1}\widehat{u}_\theta\right)^{-m/2}. \qquad (4.2.8)$$

Thus, applying IC procedures to (4.2.7) is equivalent to applying them to (4.2.8) when the model selection problem involves choosing different $\Omega(\theta)$ specifications for the distribution of u in (4.2.1). Either way, the resultant procedure is invariant to transformations of the form of (4.2.6).

4.3 The Theory of Optimal Penalties in Small Samples

From the literature, it is evident that the probability of correct selection of an IC procedure depends on factors other than sample size and the number of parameters, q, in the model. In particular, it depends on the characteristics of the models in the set of models from which the selection is made. For choosing between AR(1) and MA(1) disturbance models which both have the same number of parameters, and hence, penalties, Grose and King (1994) showed that the probability of correct selection is influenced by the shape of the LF. The latter depends on the error process and the design matrix. A good penalty function should, in our view, be able to adjust to different design matrices and error structures in addition to n and

q. In this section, we outline the theory towards obtaining optimal penalties in small samples. The process for estimating APCS and OAPCS for a given optimal penalty set is the same as discussed in Chapter 3.

Let us assume that $M_1, M_2, \cdots, M_N$ denote N models each with different disturbance processes of (4.2.1), and $\log L_i(\widehat{\theta}_i)$ is the maximized log-likelihood function for the ith model M_i with $\widehat{\theta}_i$ being the maximum likelihood estimate of θ_i. The probability of correct selection for the ith true model can be given by equation (3.5.1) of Section 3.5, which changes as θ_i changes. As was done in Section 3.5, this problem can be solved through calculating the APCS for the ith true model by using $\zeta(\theta_i)$ as the weighting density function for θ_i. Thus, APCS_i is given by

$$\text{APCS}_i(p_1, \cdots, p_N) = \int Pr[CSM_i | M_i, \theta_i, p_1, \cdots, p_N] \zeta(\theta_i) d\theta_i. \tag{4.3.1}$$

In general, calculation of (4.3.1) is not easy, because $Pr[CSM_i | M_i, \theta_i, p_1, \cdots, p_N]$ is unknown. However, following the procedures in Section 3.5, $\text{APCS}_i(p_1, \cdots, p_N)$ can be estimated as follows:

$$\widehat{\text{APCS}}_i(p_1, \cdots, p_N) = \frac{1}{R} \sum_{\ell=1}^{R} I_\ell(M_i, \theta_{i\ell}, p_1, \cdots, p_N), \tag{4.3.2}$$

where $\theta_{i\ell}$ is the ℓth drawing from $\zeta(\theta_i)$ and $I_\ell(M_i, \theta_{i\ell}, p_1, \cdots, p_N)$, $\ell = 1, \cdots, R$, is defined in Section 3.5.

Then, the OAPCS for the set of penalties $p_1, \cdots, p_N$, is estimated by

$$\widehat{\text{OAPCS}}(p_1, \cdots, p_N) = \frac{1}{N} \sum_{i=1}^{N} \widehat{\text{APCS}}_i(p_1, \cdots, p_N). \tag{4.3.3}$$

Our proposed IC procedure involves finding the penalties $p_1, \cdots, p_N$ which maximize (4.3.3). Without loss of generality, we can set $p_1 = 0$. Then, (4.3.3) will be maximized with respect to the penalties $p_2, \cdots, p_N$. The APCS and hence, OAPCS changes with changes in $p_2, \cdots, p_N$. Thus, the optimal penalties are found when OAPCS is at its global maximum.

4.4 Suitable Methods of Optimization

In this section, we describe two possible methods of optimizing (4.3.3), namely PEM-GS and PEM-SA.

4.4.1 PEM-GS Algorithm for Model Selection

Let b_{il} and b_{iu} be preselected lower and upper limits for p_i, $i = 2, \cdots, N$, chosen by the researcher. For each value of i, a grid of S_i values of p_i is generated as $b_{il}, b_{il} + \delta, b_{il} + 2\delta, \cdots, b_{il} + (S_i - 1)\delta = b_{iu}$, where δ is decided by the researcher. The smaller δ is, the finer will be the GS, with a cost of higher computational time. From these grids of p_i values, $K = S_2 \times S_3 \times \cdots \times S_N$ possible penalty sets are constructed and for each set, (4.3.3) is calculated. The penalty values, denoted $p_2, \cdots, p_N$, corresponding to the highest calculated value of (4.3.3), are recorded. Then, a new but much finer grid of penalty values centered at $p_2, \cdots, p_N$ is calculated and the process is repeated. The whole process is repeated several times. The GS ends by comparing the last maximum OAPCS with the most recent maximum OAPCS. If the difference is negligible, the algorithm can be terminated. The GS should converge to the global maximum of (4.3.3).

GS is one of the successful ways to optimize a very complicated and relatively ill-behaved function. However, the drawback of GS is the required computational time which increases sharply with the number of parameters and so can be extremely high for a refined search.

4.4.2 PEM-SA Algorithm for Model Selection

A discussion of the SA algorithm has been presented in Chapter 2, where it was noted that this algorithm is very robust even with respect to estimating the parameters of difficult functions. In this subsection, we discuss how PEM-SA can be

developed from the SA algorithm for model selection problems.

Let us assume that p be the vector of penalties, $p = (p_2, \cdots, p_N)'$ and let $f(p)$ denote (4.3.3), the function to be optimized. Also, we assume that the optimal value of p is such that $b_{il} < p_i < b_{iu}$, where the values of b_{il} and b_{iu} are nominated by the user. Note that in our case, $f(p)$ is a discrete valued but bounded function.

Starting from an initial point p_0, the algorithm randomly chooses a new point p' in the neighborhood (within step length V, an $(N-1) \times 1$ vector of maximum step lengths selected by the user) of p_0. The value of the objective function is evaluated at this trial point and is compared to its value at the initial point. In a maximization problem such as ours, all up hill moves are accepted, i.e., if the change $\Delta f = f(p') - f(p_0)$ represents an increase in the value of the objective function, then the new point is accepted and the algorithm continues from this trial point. Note that the step length is always centered at the trial point. If the change represents a reduction in the objective function value, then the down hill moves may be accepted with probability $\exp(-\Delta f/T)$, where T is a parameter called temperature. If the trial point is rejected, another point is chosen for a trial evaluation.

Each element of the step length vector V is adjusted periodically so that about half of all points are accepted. The algorithm also requires the specification of a cooling schedule. An initial value T_0 is set for the temperature parameter T. It should be relatively high, in general, so that most trials are accepted and there is little chance of the algorithm zooming in on a local maximum in the early stages. A fall in temperature is imposed upon the system with a temperature reduction factor r_T ranging from 0 to 1. Finally, a stopping criterion is imposed to terminate the algorithm.

The following is a detailed description of how the algorithm can be implemented

for our model selection problem.

Step 1 (Initialization of parameters)

Decide on initial values for:

p_0, the initial penalty vector of order $N-1$,

V_0, the starting step length vector of order $N-1$,

T_0, the initial temperature,

r_T, the temperature reduction factor; the suggested value by Corana et al. (1987) is 0.85,

N_s, the number of cycles; after $N_s(N-1)$ function evaluations, each element of the step length vector is adjusted,

N_T, the number of iterations to each temperature reduction,

ϵ, the error tolerance for termination,

N_ϵ, number of final function values used to decide upon termination; suggested value is 4,

b_{il}, the lower bound for the penalty p_i, which is set to zero for our purposes,

b_{iu}, the upper bound for the penalty p_i,

$c = (c_2, \cdots, c_N)'$, the vector that controls the step length adjustment.

In the steps which follow, whenever the objective function $f(p)$ needs to be evaluated at $p = (p_2, \cdots, p_N)'$, this involves a series of $N-1$ simulations of length R as outlined in Section 4.3 to obtain $\widehat{\text{OAPCS}}(p_2, \cdots, p_N)$ as given by (4.3.3) which is the required value of $f(p)$.

Step 2 (Calculate the objective function)

Calculate $f(p_0)$ and store p_0 as p and $f(p_0)$ as f.

Step 3 (Searching for a new penalty)

Generate a new penalty p' by varying element i of p as

$$p_i' = p_i + u^* v_i,$$

where u^* is a uniformly distributed number from [-1,1], generated by a random number generator and v_i is the $(i-1)$th element of the step vector V, $i = 2, \cdots, N$. If p_i' is outside $b_{il} \leq p_i' \leq b_{iu}$, then repeat step 3 until a p_i' is found that is within these bounds.

Step 4 (Metropolis criterion)

Compute $f' = f(p')$. If $f' > f$, then accept the new penalty, i.e., store p' as p and f' as f, the optimum value of the function. If $f' \leq f$, the Metropolis criterion decides on acceptance or rejection of the penalty with acceptance probability

$$p_r = \exp\left(\frac{f' - f}{T}\right).$$

This is done by generating p_u, a uniformly distributed random number from [0,1]. If $p_r > p_u$, the new penalty is accepted, otherwise it is rejected (and there is no change in p and f).

Step 5 (Adjustment of V)

After N_s steps through all elements of p, i.e., after $N_s(N-1)$ function evaluations, the step length vector V is adjusted so that approximately half of all function evaluations are accepted. The $(i-1)$th element of the new step vector V' is

$$\begin{aligned} v_i' &= v_i\left(1 + c_i\frac{m_i/N_s - 0.6}{0.4}\right) \qquad \text{if} \quad m_i > 0.6N_s, \\ v_i' &= \frac{v_i}{1 + c_i\frac{0.4 - m_i/N_s}{0.4}} \qquad \text{if} \quad m_i < 0.4N_s, \\ v_i' &= v_i \qquad \text{otherwise,} \end{aligned}$$

where m_i is the number of moves accepted in step 4 after N_s steps through the $(i-1)$th element of p and c_i is the $(i-1)$th element of c.

Step 6 (Temperature reduction)

After repeating steps 3 to 5 N_T times, i.e., after $N_T N_s(N-1)$ function evaluations, the temperature is reduced by the temperature reduction factor r_T:

$$T' = r_T T.$$

Set T equal to T' and return to step 3.

Step 7 (Termination criterion)

Let us assume that f_k be the most recent function value from the kth temperature reduction, $f_{k-k'}$, $k' = 1, 2, \cdots, N_\epsilon$, be the last N_ϵ values of the largest function values at the temperature reduction step. Then, if

$$\begin{aligned} |f_k - f_{k-k'}| &\leq \epsilon, \qquad k' = 1, 2, \cdots, N_\epsilon, \\ |f_k - f_{opt}| &\leq \epsilon, \end{aligned}$$

stop the search.

4.5 Design of the Monte Carlo Study

We conducted a Monte Carlo experiment in order to evaluate the feasibility of the above two approaches to penalty value computation and to compare the performance of the resultant IC procedures (denoted PEM-GS and PEM-SA) with a range of existing IC procedures. The latter procedures are AIC, BIC, HQ, MCp, GCV and FPE. These criteria and their penalty functions can be summarized as

follows:

Criterion	Penalty Function
AIC	q
BIC	$q\log(n)/2$
HQ	$q\log(\log(n))$
MCp	$n\log(1+2q/m^*)/2$
GCV	$-n\log(1-q/n)$
FPE	$(n\log(n+q)-n\log(n-q))/2$

where n is the sample size, q is the total number of free parameters in the model and $m^* = n - q^*$, where q^* is the number of free parameters in the smallest model which nests all models under consideration.

The experiment was conducted in the context of selecting between WN, AR(1), MA(1) and AR(2) disturbance processes in the general linear regression model (4.2.1). In this situation, β and σ^2 are nuisance parameters and θ is the vector of parameters of interest. The probabilities of correct selection are influenced by a number of factors, but particularly by the value of θ. We need $\zeta(\theta)$, the weighting function (similar to a prior density function) discussed in Section 4.3. Its purpose is to weight different parameter vector values when calculating the APCS. This weighting function is not necessary for WN disturbances.

In the case of AR(1) disturbances given by

$$u_t = \rho u_{t-1} + e_t, \qquad t = 1, \cdots, n, \tag{4.5.1}$$

where $e = (e_1, \cdots, e_n)' \sim N(0, \sigma^2 I_n)$, $\theta = \rho$ and $\zeta(\rho)$ was taken as the uniform distribution on [-1,1]. For MA(1) disturbances given by

$$u_t = \gamma e_{t-1} + e_t, \qquad t = 1, \cdots, n, \tag{4.5.2}$$

in which $e^* = (e_0, e_1, \cdots, e_n)' \sim N(0, \sigma^2 I_{n+1})$, $\theta = \gamma$ and $\zeta(\gamma)$ was also taken as the

uniform distribution on [-1,1]. For AR(2) disturbances given by

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + e_t, \qquad t = 1, \cdots, n, \tag{4.5.3}$$

in which $e \sim N(0, \sigma^2 I_n)$, then $\theta = (\phi_1, \phi_2)'$. Because $\phi_1 = \rho_1 + \rho_2$ and $\phi_2 = -\rho_1\rho_2$ where ρ_1 and ρ_2 are the roots of $z^2 - \phi_1 z - \phi_2 = 0$ which lie inside the unit circle, generating ρ_1 and ρ_2 satisfying the condition $|\rho_1|, |\rho_2| < 1$ identifies ϕ_1 and ϕ_2. Therefore, ρ_1 and ρ_2 were generated randomly from two independent uniform distributions with range [-1,1].

The relevent steps of the experiment are as follows:

Step 1: At each replication, draw ρ, γ and $(\phi_1, \phi_2)'$ values from their respective $\zeta(\theta)$ distributions as outlined above.

Step 2: Setting $\beta = 0$ and $\sigma^2 = 1$ (which can be done without loss of generality because our model selection problem is invariant to transformations of the form (4.2.6), and therefore, to the values taken by β and σ^2), generate four samples of size n from each of the WN, AR(1), MA(1) and AR(2) disturbance models.

Step 3: For each data set, compute the maximized log-likelihood for each of the four models.

Step 4: Repeat steps 1-3 $R = 2000$ times. The total number of maximized log-likelihoods thus calculated is $4 \times 4 \times 2000 = 32000$.

Step 5: Using these maximized log-likelihoods, estimate the APCS for each IC procedure and for each of the four models. Also, calculate the OAPCS in each case.

Step 6: Finally, estimate p_2, p_3 and p_4 (setting $p_1 = 0$ without loss of generality) by using PEM-GS as well as PEM-SA, as outlined in Section 4.4. Then, use these penalties on the maximized log-likelihoods to estimate the APCS for the PEM-GS and PEM-SA procedures and also their OAPCS.

For comparison purposes, we used both the PL (4.2.5) and the MGL (4.2.8). Except for the WN model, a closed form maximum likelihood estimator for θ is

not available for either likelihood. The GAUSS (see Aptech, 1996) constrained optimization algorithm was used to maximize log-likelihoods where required.

The following four design matrices with $n = 20$ and 30 were used in the experiment.

$X1 : n \times 3$. The eigenvectors corresponding to the three smallest eigenvalues of the Durbin-Watson (DW) $n \times n$ A_1 matrix, and hence, regressors corresponding to the upper bound of the DW statistic. A_1 is a tridiagonal matrix with 2's down the main diagonal, -1's on the off-diagonals and 1 as the top left and bottom right elements.

$X2 : n \times 3$. The first n observations of Durbin and Watson's (1951, p.159) consumption of spirits example.

$X3 : n \times 3$. A constant dummy, the quarterly Australian Consumer Price Index commencing 1959(1) and the same index lagged one quarter.

$X4 : n \times 3$. A constant dummy, quarterly Australian private capital movements and quarterly Australian Government capital movements commencing 1968(1).

These design matrices cover a range of behavior. A plausible estimate of the autocorrelation of the error term is given by $(2 - d)/2$, where d is the familiar DW test statistic. Therefore, the results for $X1$ may show some extreme behavior. $X2$ is comprised of annual data, while $X3$ and $X4$ are constructed from quarterly data. $X4$ contains strong seasonal regressors with two seasonal peaks per annum in addition to some large fluctuations.

4.6 Results and Discussion

The results from the simulation study are presented in Tables 4.5 to 4.8. These tables give estimated APCSs for each of the 4 models in turn for each of the selection procedures, based on both the PL and MGL. The estimated OAPCS is also provided. The OAPCS for $n = 20$ and 30 are also presented in Figures 4.1 and 4.2, respectively for design matrices $X1$ to $X3$. The following is a detailed discussion of the results.

From these results, it is clear that in terms of individual APCS, no one procedure performs better than the others and relative performances change with changes in sample size and design matrix. The existing IC procedures and PEM-SA are ranked for each true model (DGP) with respect to APCS for all design matrices and sample sizes considered in this chapter. Then, the overall rank for each criterion is calculated by taking the average of all corresponding ranks. These overall ranks are presented in Table 4.1. This table shows that among the PL based existing IC procedures, BIC performs better when correctly selecting WN and AR(1) models. For correctly selecting MA(1) and AR(2) error models respectively, FPE and MCp perform better. Among the MGL based IC procedures, BIC, GCV, AIC/FPE and MCp perform better for selecting WN, AR(1), MA(1) and AR(2) error models, respectively.

Compared to the existing IC procedures, in terms of APCS, the PL based PEM-SA performs better when correctly selecting WN, AR(1) and AR(2) error models and it correctly selects the MA(1) error model least often. MGL based PEM-SA performs very well at correctly selecting AR(2) error models. Compared to PEM-SA, the MGL based existing IC procedures perform very poorly for correctly selecting AR(2) error models. At correctly selecting WN, AR(1) and MA(1) error models, PEM-SA occupies second, sixth and third positions, respectively.

Table 4.1: Overall ranks for the existing IC procedures and PEM-SA, based on estimated APCS.

LF	DGP	AIC	BIC	HQ	MCp	GCV	FPE	PEM-SA
PL	WN	6	2	3	4	5	7	**1**
	AR(1)	6	2	3	7	4	5	**1**
	MA(1)	2	6	5	4	3	**1**	7
	AR(2)	3	7	6	2	3	5	**1**
MGL	WN	7	**1**	3	4	5	6	2
	AR(1)	3	5	4	7	**1**	2	6
	MA(1)	**1**	7	6	5	4	**1**	3
	AR(2)	3	7	6	2	3	5	**1**

Table 4.2: Overall ranks for the existing IC procedures and PEM-SA, based on estimated OAPCS.

LF	AIC	BIC	HQ	MCp	GCV	FPE	PEM-SA
PL	5	7	6	2	3	3	1
MGL	3	7	6	2	5	3	1

Table 4.3: Comparison of computation times (elapsed time for Pentium 3, running at 333 MHz) required by the PL, based PEM-GS and PEM-SA for estimating optimal penalty values.

n	Optimization Method	$X1$	$X2$	$X3$	$X4$
20	PEM-GS	105.00	103.50	104.12	95.65
	PEM-SA	5.53	5.25	5.41	5.85
30	PEM-GS	101.30	98.66	93.01	96.12
	PEM-SA	5.15	5.37	5.44	5.55

The IC procedures and PEM-SA were also ranked with respect to OAPCS. Table 4.2, which reports these ranks, shows that among the PL based existing IC procedures, MCp performs best, followed by either GCV or FPE. The fifth and sixth positions are occupied by AIC and HQ respectively, and BIC is the worst. All the MGL based existing IC procedures carry through this ranking order, except for AIC and GCV. In this case, AIC and GCV occupied the third and fifth positions, respectively.

Of the existing IC procedures, it is interesting to note that the one with the highest OAPCS depends very much on the design matrix and sample size, while typically BIC has the lowest OAPCS, although the differences are not great in some cases (see Figures 4.1 and 4.2). The reason for BIC being ranked so low seems to be due to its poor performance at selecting the AR(2) error model when it is the true model.

The percentage improvement (with respect to OAPCS) of PEM-GS and PEM-SA over the existing IC procedures are presented in Table 4.4. This table shows that on average, the PL and MGL based PEM-SA perform better than those based

Table 4.4: Percentage improvement of PEM-GS and PEM-SA over the best and lowest performing existing IC procedures with respect to OAPCS.

Design Matrix	n	PEM-GS		PEM-SA	
		PL	MGL	PL	MGL
$X1$	20	11.28–**15.33**	6.26–9.60	12.34–**16.62**	6.05–9.56
	30	10.34–12.86	4.76–6.92	10.14–12.66	4.58–6.73
$X2$	20	8.63–13.53	6.93–**11.38**	8.39–13.28	6.93–**11.38**
	30	8.37–11.24	5.37–8.59	8.37–11.24	5.37–8.59
$X3$	20	5.23–6.34	7.87–10.15	6.90–8.03	7.87–10.15
	30	**4.45**–5.77	**4.05**–7.07	**4.44**–6.33	**4.05**–7.07
$X4$	20	6.82–7.93	6.82–11.42	6.82–7.93	6.63–10.99
	30	5.40–6.16	5.39–8.50	5.22–5.98	5.39–8.50

existing IC procedures by $4.44 - 16.62\%$ and $4.05 - 11.38\%$, respectively. Also, the OAPCS of the PL and MGL based PEM-GS method are respectively, on average $4.45 - 15.33\%$ and $4.05 - 11.38\%$ larger than those based existing IC procedures. Therefore, the main feature of the results is that both the PEM-GS and PEM-SA procedures perform uniformly better than the existing IC procedures with respect to estimated OAPCS (see Figures 4.1 and 4.2). However, in some situations, PEM-SA is marginally dominated by PEM-GS, but with a cost of high computational time. The required computation times for PL based PEM-GS and PEM-SA are presented in Table 4.3. It should be noted that the time required by the PL and MGL based PEM-GS and PEM-SA are approximately the same. Tables 4.6 and 4.3 show that for $n = 20$ and design matrix $X2$, the OAPCS of PL based PEM-SA and PEM-GS are 45.2% and 45.3%, respectively, and the corresponding computational

times are 5.53 and 105 minutes, respectively.

Comparing the performance of PL and MGL based selection procedures in terms of APCS, MGL typically out-performs PL when selecting WN and AR(1) error models as well as AR(2) error models for $X1$ and $X2$. PL generally performs better than MGL when the MA(1) error model is the true model and for AR(2) error models for $X3$ and $X4$. In terms of OAPCS, the MGL based IC procedures pick the true model more frequently for design matrices of non-lagged regressors ($X1$ and $X2$). For example, Table 4.5 shows that for $X1$ with $n = 20$, the OAPCS corresponding to the MGL based existing IC procedures are on average $21 - 22\%$ larger than those based on PL. As the sample size increases, any advantages from using MGL over PL diminishes gradually. In the case of $X3$ and $X4$ (lagged regressor), there is little difference between the OAPCSs of the two different base likelihoods. Overall, one would have to recommend the use of MGL because its use can result in a significant improvement in the OAPCS and never in a substantial drop.

4.7 Conclusions

The results of our experiment support the use of the MGL rather than the PL in IC model selection procedures applied to choosing the disturbance process in the general linear regression model. Of the existing IC procedures we considered, no one procedure stands out as the best, although it does seem that BIC has the worst overall performance. In contrast, PEM-GS consistently dominates all existing IC procedures used in this chapter in terms of OAPCS, although at a high computational cost. To avoid the computational limits imposed by PEM-GS, we considered the PEM-SA, whose performance is similar to that of PEM-GS, while its computational cost is much lower. We suggest its use in conjunction with maximized MGLs

when faced with choosing between different regression disturbance models.

Table 4.5: Estimated APCS and OAPCS (where indicated) for design matrix $X1$ and $n = 20$ and 30.

n	DGP	LF	AIC	BIC	HQ	MCp	GCV	FPE	PEM-GS	PEM-SA
20	WN	PL	0.440	0.487	0.458	0.449	0.442	0.440	0.561	0.556
		MGL	0.756	0.863	0.780	0.771	0.764	0.756	0.782	0.784
	AR(1)	PL	0.166	0.170	0.164	0.167	0.164	0.169	0.370	0.378
		MGL	0.383	0.368	0.383	0.373	0.388	0.383	0.461	0.463
	MA(1)	PL	0.729	0.677	0.722	0.726	0.727	0.729	0.485	0.495
		MGL	0.483	0.402	0.463	0.467	0.474	0.483	0.421	0.409
	AR(2)	PL	0.180	0.138	0.176	0.184	0.177	0.180	0.281	0.283
		MGL	0.228	0.159	0.217	0.235	0.217	0.228	0.304	0.296
	OAPCS	PL	0.379	0.367	0.381	0.381	0.379	0.379	0.424	0.428
		MGL	0.463	0.448	0.461	0.462	0.461	0.463	0.492	0.491
30	WN	PL	0.651	0.740	0.688	0.657	0.653	0.651	0.756	0.758
		MGL	0.780	0.904	0.836	0.791	0.787	0.780	0.856	0.850
	AR(1)	PL	0.293	0.307	0.300	0.302	0.292	0.296	0.410	0.409
		MGL	0.501	0.494	0.499	0.489	0.502	0.502	0.417	0.408
	MA(1)	PL	0.727	0.673	0.712	0.724	0.727	0.728	0.573	0.569
		MGL	0.568	0.497	0.545	0.562	0.566	0.568	0.550	0.551
	AR(2)	PL	0.289	0.215	0.254	0.292	0.279	0.289	0.438	0.434
		MGL	0.333	0.246	0.295	0.334	0.322	0.332	0.465	0.463
	OAPCS	PL	0.493	0.482	0.489	0.491	0.489	0.490	0.544	0.543
		MGL	0.546	0.535	0.544	0.544	0.544	0.546	0.572	0.571

Table 4.6: Estimated APCS and OAPCS (where indicated) for design matrix $X2$ and $n = 20$ and 30.

n	DGP	LF	AIC	BIC	HQ	MCp	GCV	FPE	PEM-GS	PEM-SA
20	WN	PL	0.521	0.601	0.539	0.532	0.562	0.521	0.628	0.636
		MGL	0.756	0.858	0.782	0.769	0.763	0.756	0.752	0.800
	AR(1)	PL	0.230	0.227	0.230	0.223	0.231	0.230	0.368	0.356
		MGL	0.390	0.381	0.387	0.380	0.391	0.390	0.377	0.347
	MA(1)	PL	0.678	0.615	0.666	0.672	0.674	0.678	0.456	0.471
		MGL	0.494	0.420	0.480	0.492	0.492	0.494	0.467	0.458
	AR(2)	PL	0.210	0.156	0.200	0.219	0.202	0.210	0.346	0.343
		MGL	0.255	0.170	0.231	0.263	0.235	0.254	0.439	0.431
	OAPCS	PL	0.409	0.399	0.408	0.412	0.417	0.409	0.453	0.452
		MGL	0.474	0.457	0.470	0.476	0.470	0.474	0.509	0.509
30	WN	PL	0.616	0.753	0.692	0.668	0.666	0.662	0.809	0.802
		MGL	0.756	0.858	0.782	0.769	0.763	0.756	0.801	0.804
	AR(1)	PL	0.315	0.323	0.325	0.311	0.317	0.315	0.439	0.440
		MGL	0.488	0.483	0.490	0.479	0.491	0.488	0.457	0.452
	MA(1)	PL	0.712	0.662	0.691	0.710	0.710	0.712	0.555	0.553
		MGL	0.570	0.493	0.541	0.563	0.567	0.570	0.535	0.537
	AR(2)	PL	0.315	0.233	0.287	0.320	0.308	0.315	0.371	0.372
		MGL	0.343	0.260	0.319	0.350	0.338	0.343	0.471	0.473
	OAPCS	PL	0.489	0.493	0.499	0.502	0.500	0.501	0.544	0.544
		MGL	0.539	0.524	0.533	0.540	0.539	0.539	0.569	0.569

Table 4.7: Estimated APCS and OAPCS (where indicated) for design matrix $X3$ and $n = 20$ and 30.

n	DGP	LF	AIC	BIC	HQ	MCp	GCV	FPE	PEM-GS	PEM-SA
20	WN	PL	0.685	0.777	0.702	0.679	0.691	0.685	0.778	0.772
		MGL	0.769	0.866	0.792	0.782	0.780	0.769	0.839	0.837
	AR(1)	PL	0.339	0.337	0.345	0.329	0.349	0.339	0.251	0.250
		MGL	0.397	0.390	0.398	0.389	0.402	0.398	0.356	0.358
	MA(1)	PL	0.579	0.536	0.570	0.571	0.577	0.579	0.498	0.500
		MGL	0.488	0.433	0.475	0.480	0.486	0.488	0.463	0.461
	AR(2)	PL	0.303	0.242	0.288	0.311	0.293	0.302	0.523	0.522
		MGL	0.272	0.201	0.254	0.279	0.257	0.272	0.424	0.420
	OAPCS	PL	0.477	0.473	0.477	0.477	0.478	0.476	0.503	0.511
		MGL	0.482	0.473	0.479	0.483	0.481	0.482	0.521	0.521
30	WN	PL	0.756	0.873	0.809	0.764	0.758	0.757	0.879	0.884
		MGL	0.797	0.904	0.846	0.804	0.801	0.797	0.869	0.872
	AR(1)	PL	0.458	0.472	0.469	0.449	0.458	0.458	0.492	0.488
		MGL	0.507	0.514	0.516	0.502	0.514	0.507	0.554	0.558
	MA(1)	PL	0.630	0.585	0.619	0.629	0.630	0.630	0.609	0.618
		MGL	0.586	0.523	0.573	0.584	0.585	0.586	0.574	0.572
	AR(2)	PL	0.374	0.292	0.350	0.380	0.366	0.373	0.358	0.359
		MGL	0.363	0.267	0.336	0.364	0.358	0.361	0.365	0.361
	OAPCS	PL	0.555	0.556	0.562	0.556	0.553	0.555	0.587	0.588
		MGL	0.563	0.552	0.568	0.564	0.565	0.563	0.591	0.591

Table 4.8: Estimated APCS and OAPCS (where indicated) for design matrix $X4$ and $n = 20$ and 30.

n	DGP	LF	AIC	BIC	HQ	MCp	GCV	FPE	PEM-GS	PEM-SA
20	WN	PL	0.602	0.700	0.628	0.620	0.613	0.603	0.694	0.691
		MGL	0.744	0.837	0.767	0.756	0.754	0.744	0.689	0.687
	AR(1)	PL	0.362	0.362	0.369	0.356	0.372	0.365	0.453	0.450
		MGL	0.430	0.421	0.429	0.426	0.434	0.430	0.365	0.360
	MA(1)	PL	0.647	0.607	0.638	0.642	0.644	0.647	0.589	0.591
		MGL	0.488	0.408	0.467	0.478	0.482	0.488	0.512	0.515
	AR(2)	PL	0.312	0.248	0.300	0.316	0.301	0.312	0.331	0.333
		MGL	0.270	0.188	0.252	0.276	0.257	0.270	0.501	0.499
	OAPCS	PL	0.481	0.479	0.479	0.484	0.483	0.482	0.517	0.517
		MGL	0.483	0.464	0.478	0.484	0.482	0.483	0.517	0.515
30	WN	PL	0.730	0.853	0.774	0.738	0.734	0.730	0.876	0.875
		MGL	0.784	0.899	0.826	0.793	0.790	0.784	0.898	0.903
	AR(1)	PL	0.461	0.469	0.463	0.454	0.466	0.461	0.450	0.444
		MGL	0.503	0.509	0.510	0.498	0.508	0.503	0.443	0.440
	MA(1)	PL	0.653	0.603	0.633	0.649	0.652	0.653	0.615	0.617
		MGL	0.579	0.519	0.563	0.578	0.580	0.579	0.575	0.574
	AR(2)	PL	0.379	0.284	0.347	0.384	0.371	0.379	0.404	0.403
		MGL	0.355	0.262	0.318	0.359	0.349	0.354	0.431	0.429
	OAPCS	PL	0.556	0.552	0.554	0.556	0.556	0.556	0.586	0.585
		MGL	0.555	0.547	0.554	0.541	0.557	0.555	0.587	0.587

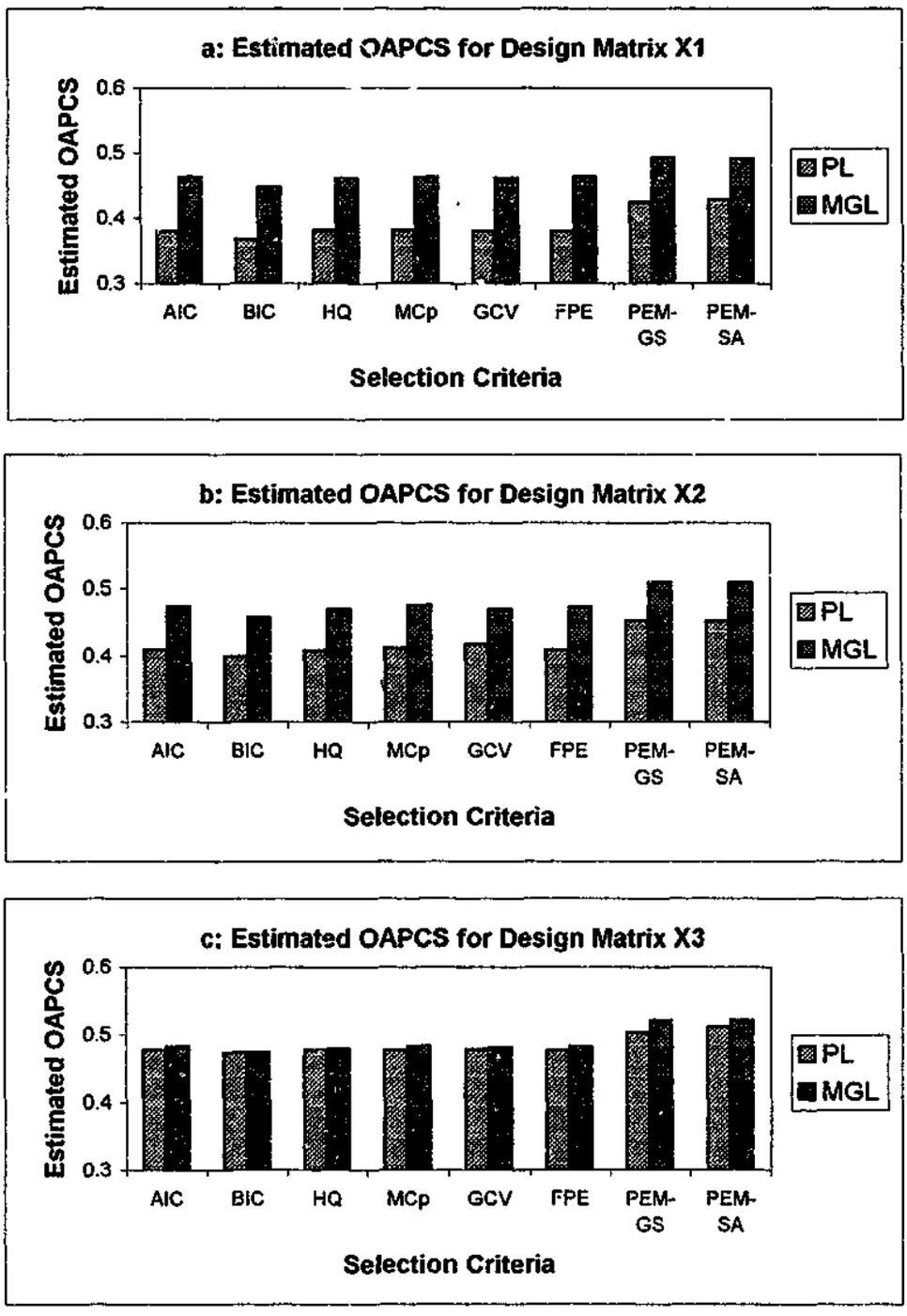

Figure 4.1: Estimated OAPCS for $n = 20$ and design matrics $X1$ to $X3$.

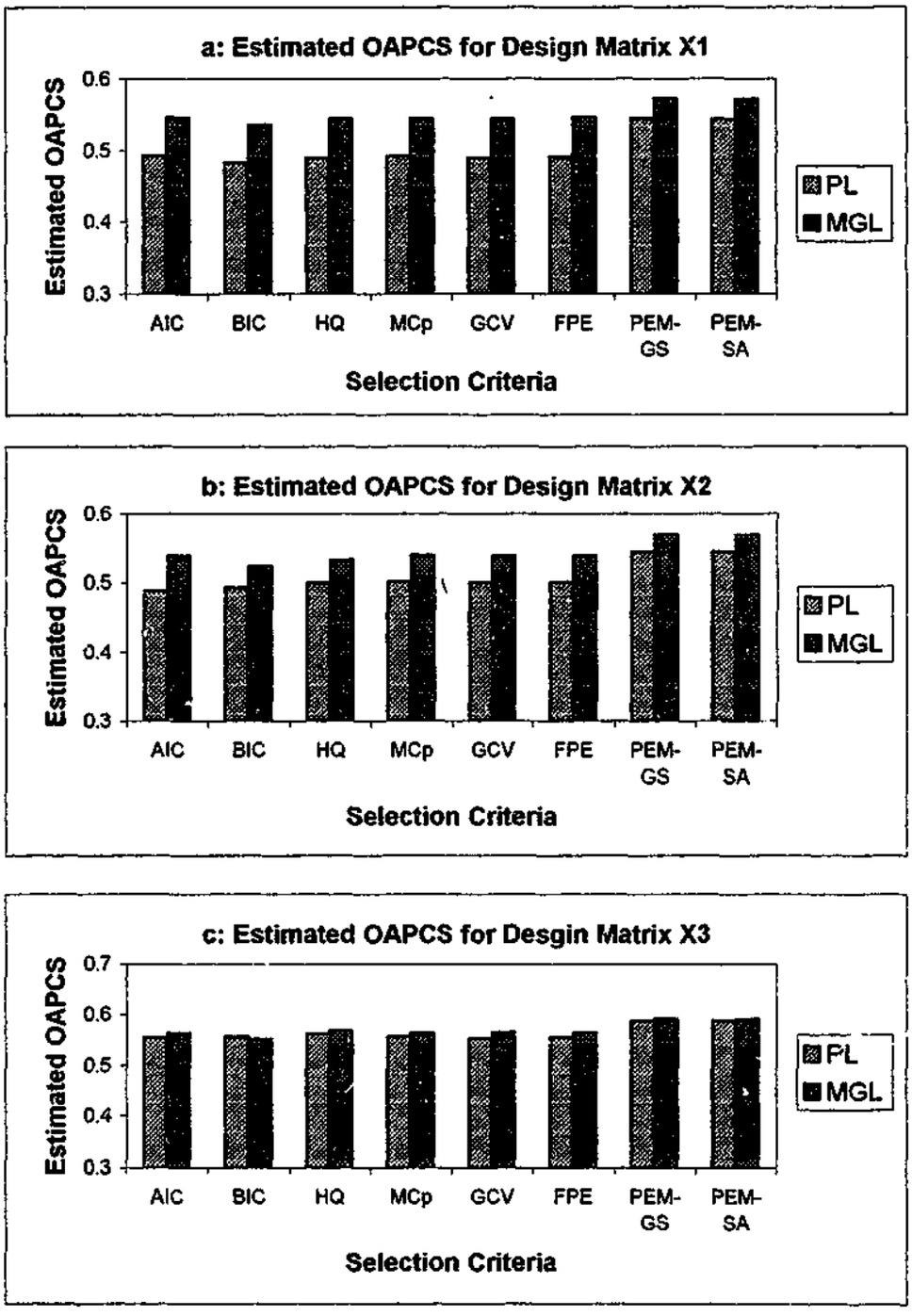

Figure 4.2: Estimated OAPCS for $n = 30$ and design matrices $X1$ to $X3$.

Chapter 5

Regression Error Model Selection for Forecasting Via PEM[1]

5.1 Introduction

In applied work, it is often assumed that a given set of time series data are generated by an AR process. On the basis of this assumption, the data generation process can be modeled and used, for example, for forecasting future values of the time series. The order determination of the AR process is usually the most delicate part of an analysis of this kind. Recently, there has been a substantial amount of literature on this problem and various different criteria have been proposed to help in choosing the order of an AR process. A number of IC procedures have been presented in Chapter 2. The most prominent and widely used are AIC and BIC. Among the remaining alternatives, GCV, HQ and MCp are also familiar to practitioners. Asymptotic theory leads to some precise results regarding the performance properties of these model selection criteria in large samples. However, the practitioner (e.g., see Geweke and Meese, 1981; Koehler and Murphree, 1988; Lutkepohl, 1985; Engle and Brown, 1986; Koreisha and Pukkila, 1995) usually faces the problem of making a choice on the basis of a limited data set. In this study,

[1] A paper based on the material in this chapter has been published in a special issue of the *Pakistan Journal of Statistics*. See Billah and King (2000a)

we are concerned with the performance of the various criteria in the context of the linear regression model with AR disturbances in small samples.

To select a model which best represents a time series, it is necessary to be clear about the purpose of the model. Is its main objective to explain the nature of the system generating the series? Or is the model to be judged on its ability to forecast future values of the time series? This chapter is concerned with the forecasting abilities of a model. Usually, the proportion of times that each criterion selects the true data generating model has been used as the main performance characteristic for evaluating a criterion. However, there are other considerations. For example, Makridakis (1986), and Mills and Prasad (1992) observe that a model having the best within-sample fit for a given series does not necessarily mean it is the best forecasting model. In general, the best forecasting model for a series is the one which has the most accurate post-sample forecasting performance. This makes it useful to compare model selection criteria in terms of out-of-sample forecast errors of the chosen model. Previous studies concerning AR errors in the linear regression model have not addressed the model selection issue in detail from the stand-point of forecasting. Therefore, we evaluate the relative performance of the model selection criteria with the application of forecasting clearly in mind.

From previous studies (e.g., Crato and Ray, 1996; Ray, 1993; Engle and Brown, 1986), it is not clear which criteria, if any, should be used for the selection of a forecasting model and the conditions under which different criteria can be expected to perform well. As in model selection, the forecasting performance of a criterion is sensitive to the variation of factors such as sample size, number of parameters in the model, forecasting horizon, design matrix and data series being forecast. However, the existing IC based selection procedures take into account only sample size and number of parameters. Therefore, another objective of this chapter is to extend

the PEM (PEM-GS and PEM-SA) based model selection procedures proposed in Chapter 4 so that the selected model's average forecast error as measured by the overall average forecast mean square error (OAMSE) is minimized (see Section 5.4 for the definition of OAMSE). We expect the extended procedures will result in uniformly better forecasting performance irrespective of the variation of any (or all) of the above factors.

The plan of this chapter is as follows. The models, their estimation methods and forecasting equations are discussed in Section 5.2. Section 5.3 contains a description of the theory for model selection and forecasting using existing IC procedures. The theory of small sample penalty estimation for forecasting is discussed in Section 5.4. Section 5.5 presents the PEM-GS and PEM-SA methods for forecasting. The design of the Monte Carlo study is given in Section 5.6, while the results of the study are presented in Section 5.7. Finally, in Section 5.8, we draw some conclusions.

5.2 The Model and Estimation Method

Independent error processes are commonly assumed for applications of the linear regression model. However, for economic time series applications, the assumption of independent errors may be unrealistic and the possible presence of autocorrelation in the disturbance term is well recognized. In this chapter, our interest is in forecasting for models with AR error terms. Since their introduction, these models have proved to be useful tools in time series analysis and have been extensively used in many applications.

We consider the linear regression model,

$$y = X\beta + u, \tag{5.2.1}$$

where y is an $n \times 1$ vector of observations, X is an $n \times k$ non-stochastic matrix of

rank $k < n$, β is a $k \times 1$ parameter vector and u is an $n \times 1$ disturbance vector such that $u \sim N(0, \sigma^2\Omega(\theta))$ in which $\Omega(\theta)$ is an $n \times n$ positive definite matrix function and θ is a vector of unknown parameters.

In the context of (5.2.1), a p^*th-order autoregressive error process, denoted by AR(p^*), is given by

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_{p^*} u_{t-p^*} + e_t, \qquad t = 1, 2, \cdots, n, \tag{5.2.2}$$

where $e \sim N(0, \sigma^2 I)$. A stationary AR(p^*) process is assumed because it guarantees that the u_t's have finite variance and the covariance matrix is positive definite. Stationarity requires ϕ_i, $i = 1, \cdots, p^*$, to be such that the characteristic roots of the equation

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_{p^*} z^{p^*} = 0 \tag{5.2.3}$$

lie inside the unit circle.

The matrix $\Omega(\theta)$ can be given as follows:

$$\Omega(\theta) = [\mathcal{L}'\mathcal{L} - \mathcal{N}\mathcal{N}']^{-1}, \tag{5.2.4}$$

where $\mathcal{L}$ is the $n \times n$ matrix of the form

$$\mathcal{L} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ -\phi_1 & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ -\phi_2 & -\phi_1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & & \vdots & & \vdots \\ -\phi_{p^*} & -\phi_{p^*-1} & \cdots & -\phi_1 & 1 & 0 & \cdots & 0 \\ 0 & -\phi_{p^*} & -\phi_{p^*-1} & \cdots & -\phi_1 & 1 & & 0 \\ \vdots & \vdots & \ddots & & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & & -\phi_{p^*} & \cdots & -\phi_1 & 1 \end{bmatrix}$$

and $\mathcal{N}$ is the $n \times p^*$ matrix of zeros whose top $p^* \times p^*$ block is

$$\begin{bmatrix} -\phi_{p^*} & -\phi_{p^*-1} & -\phi_{p^*-2} & \cdots & -\phi_1 \\ 0 & -\phi_{p^*} & -\phi_{p^*-1} & \cdots & -\phi_2 \\ 0 & 0 & -\phi_{p^*} & \cdots & -\phi_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\phi_{p^*} \end{bmatrix}.$$

For more detail, see Van der Leeuw (1994) or Ljung and Box (1979).

The Cholesky decomposition matrix $\mathcal{D}(\theta) = \otimes(\theta)^{-\frac{\infty}{\epsilon}}$ is useful in this context because of the relationship $\Omega(\theta)^{-\frac{1}{2}}u = e$, where $e \sim N(0, \sigma^2 I)$. In fact, the matrix $\mathcal{D}(\theta)$ is equal to $\mathcal{L}$, but with the top left $p^* \times p^*$ block replaced by the lower triangle matrix,

$$\begin{bmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ d_{21} & d_{22} & 0 & \cdots & 0 \\ d_{31} & d_{32} & d_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ d_{p^*1} & d_{p^*2} & \cdots & \cdots & d_{p^*p^*} \end{bmatrix}.$$

Following Ara (1995), the d_{ij}, $i, j = 1, \cdots, p^*$, values of the above matrix can be calculated recursively in the following order:

$$\begin{aligned}
d_{p^*p^*} &= \left(1 - \phi_{p^*}^2\right)^{\frac{1}{2}}, \\
d_{p^*p^*-1} &= -\frac{1}{d_{p^*p^*}}\left(\phi_i + \phi_{p^*-i}\phi_{p^*}\right), \\
d_{\iota\iota} &= \left(1 + \sum_{i=1}^{\iota-1}\phi_i^2 - \sum_{i=\tau+1}^{p^*}\phi_i^2 - \sum_{i=1}^{\tau} d_{(\iota+i)\iota}^2\right)^{\frac{1}{2}}, \\
&\qquad \text{for} \quad \iota = p^* - 1, \cdots, 2 \quad \text{and} \quad \tau = p^* - \iota, \\
d_{\iota\iota-k} &= \frac{1}{d_{\iota\iota}}\left(-\phi_k - \sum_{i=1}^{\iota-k-1}\phi_i\phi_{i+k} - \sum_{i=\tau+1}^{p^*-k}\phi_i\phi_{i+k} - \sum_{i=\iota+1}^{p^*} d_{i\iota}d_{i(\iota-k)}\right), \\
&\qquad \text{for} \quad k = 1, \cdots, \iota - 2, \\
d_{\iota 1} &= \frac{1}{d_{\iota\iota}}\left(-\phi_{\iota-1} - \phi_{m+1}\phi_{p^*} - \sum_{i=\iota+1}^{p^*} d_{i\iota}d_{i1}\right), \\
d_{11} &= \left(1 - \phi_{p^*}^2 - \sum_{i=1}^{p^*-1} d_{(i+1)1}^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Wu (1991) and Silvapulle (1991) presented an alternative derivation of the $\mathcal{D}(\theta)$ matrix. The above decomposition of $\Omega(\theta)$ helps in easy data generation from AR(p^*) processes as well as in calculating the MGL. In our experience, it also helps to optimize the MGL function without frustrating computational error messages.

In most econometric applications, an adequate representation of a model in-

volves many unknown factors about which we seek information. In the present context, this comes down to deriving optimal estimators for the parameter vector $\theta = (\phi_1, \cdots, \phi_{p^*})'$ of the models in which we are interested. The likelihood function plays an important role in all forms of statistical inference, i.e., estimation, testing and model selection. However, problems arise with the estimation of θ in the presence of nuisance parameters, particularly for small samples. The studies of Chapter 3 and 4 (also, see Billah and King, 2000b) show that the MGL based IC procedures perform better than those based on PL. Therefore, in this chapter we consider MGL based estimates of θ for model selection.

5.3 Model Selection and Forecasting Using Existing IC Procedures

As already discussed, in many practical problems, the order of an autoregressive series is unknown and must be estimated using the available data. Once the model is selected, it is of interest to evaluate its forecasting performance. Suppose we have a data series, $y_1, y_2, \cdots, y_n$, where n is the size of the series. The series is divided into two segments. Let us assume that the number of observations in the first segment is n^* and that in the second segment is H. The values of n^* and H are determined by the forecaster. The first n^* observations are used to estimate the model and the next H observations are held back to evaluate forecasting performance of various models. Let $\widehat{\phi}_i$, $i = 1, \cdots, p^*$, be the MGL estimates of the ϕ_i parameters. Then, a reasonable h-step ahead forecast from the regression model with AR(p^*) errors would be:

$$y_t(h) = c_{t+h} + \sum_{i=1}^{p^*} \widehat{\phi}_i \widehat{u}_{(t+h-i)}, \quad t = 1, \cdots, n, \tag{5.3.1}$$

where $\hat{c}_{t+h} = X'_{t+h}\hat{\beta}$, $\hat{\beta}$ is the GLS estimate of β and $\hat{u}_{(t+h-i)}$ is given by

$$\hat{u}_{(t+h-i)} = \sum_{j=1}^{p^*} \hat{\phi}_j \hat{u}_{(t+h-i-j)}, \text{ if } h-i>0 \quad (5.3.2)$$

$$= y_{t+h-i} - c_{t+h-i}, \text{ if } h-i \leq 0. \quad (5.3.3)$$

The unknown parameter values of the variance covariance matrix in the GLS estimators are obtained by maximum MGL estimation methods of the underlying model.

To evaluate the forecast accuracy of a model, a loss function associated with the forecast error needs to be specified. In the literature, different methods have been proposed for forecast accuracy evaluation, such as mean absolute error (MAE), MAPE, MSE, root mean square error (RMSE), symmetric mean absolute percentage error (SMAPE). Discussions of some of these criteria can be found in Makridakis et al. (1982) and Makridakis and Hibon (2000). Among these measures, the most widely used forecast evaluation criteria in applied and theoretical research is MSE and MAPE. However, MAPE requires non-zero observations, and hence, may not be appropriate for simulated data. Therefore, in this chapter, we use the MSE as our measure of forecast accuracy.

Let us assume that we have N data generating models, $M_1, \cdots, M_N$. We also consider $M_1, \cdots, M_N$ as the plausible models. For each data series generated from the ith model, we fit each of the N models and select a model by using IC. Note that the selected model may not be the data generating model. Even when the true order of the generating process is known, a lower order model may give better forecasts (see e.g., Larimore and Mehra, 1985). The selected model is then used to forecast future observations. Let us assume that from each model we have R simulated data series. We also assume that at time t, $e_{i'it}(h) = y_{it(t+h)} - y_{i't}(h)$, $i, i' = 1, \cdots, N$, is the h-step ahead forecast error from the i'th selected model when

data are generated from the ith model at the ℓth replication, where $y_{i\ell t}(h)$ is the h-step ahead forecast for observed value $y_{i\ell(t+h)}$. The selected model for each data generating process is then used for making $H-h+1$ h-step ahead forecasts without re-estimating the model. Thus, the forecast MSE when the data are coming from the ith model at the ℓth replication is given by

$$\text{MSE}_{i\ell}(h) = \frac{1}{H-h+1}\sum_{t=n^*+h}^{n^*+H} e^2_{i\ell t}(h). \tag{5.3.4}$$

For simplicity, we replace $\text{MSE}_{i\ell}(h)$ and $e^2_{i\ell t}(h)$ by $MSE_{i\ell}$ and $e^2_{i\ell t}$, respectively, in the following sections of this chapter.

The selection of a forecast model given a true model depends on the true parameter values θ_i being considered in data generation. Following the arguments for estimating APCS of Billah and King (2000b) (also, see Chapter 4), this problem can be solved by calculating average MSE (AMSE), which needs a weighting function for various values of θ_i. As mentioned by Billah and King (2000b), this is similar to a prior density function used in Bayesian statistics. Bayesians use prior distributions and we can do the same by considering a weighting density function for the values of θ_i. Let $\zeta(\theta_i)$ be a weighting distribution of the parameter of interest θ_i. Thus, when data are generated from the ith model at the ℓth replication, the AMSE can be obtained by

$$\text{AMSE}_i = \int \text{MSE}_{i\ell}\zeta(\theta_i)d\theta_i. \tag{5.3.5}$$

However, in general, the direct application of (5.3.5) to obtain AMSE_i is not easy. Therefore, to estimate (5.3.5), we consider Monte Carlo integration by drawing θ_i randomly from $\zeta(\theta_i)$ and using this to generate R simulated sets of data. Then, the estimate of (5.3.5) is obtained as follows:

$$\widehat{\text{AMSE}}_i = \frac{1}{R}\sum_{\ell=1}^{R}\text{MSE}_{i\ell}, \quad i=1,\cdots,N. \tag{5.3.6}$$

The forecast MSE_{it} in the above equation can be estimated using (5.3.4).

Equation (5.3.6) can be used for estimating the overall or aggregate AMSE (OAMSE). This is obtained by pooling N different AMSE_i in (5.3.6) with equal weighting. Thus, the estimated OAMSE is given by

$$\widehat{\text{OAMSE}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\text{AMSE}}_i. \tag{5.3.7}$$

How well different IC procedures perform in ultimately producing good forecasts is judged using OAMSE. The criterion having the smallest OAMSE is considered best in the sense that it produces the best post-sample forecasts on average. As we will see from the results of the simulation study, none of the criteria clearly perform better than the others. Hence, similar to PEM of Chapter 4, we feel that there is a need for a new approach which will perform better in every aspect. In the next section, we discuss the theory of optimal penalty estimation for selecting a forecasting model for small samples.

5.4 Small Sample Theory of Optimal Penalties for Minimization of Mean Square Forecast Error

In this section, we introduce the theory for developing simulation based model selection procedures which will estimate optimal penalty values so that OAMSE is minimized. The motivation for developing these new procedures arises from the fact that no single model selection procedure (IC procedure) stands out as the best for producing forecasts for all data series. These sentiments have been echoed by Engle and Brown (1986), Ray and Crato (1994), and Crato and Ray (1996), among others. As in model selection (see e.g., Billah and King, 1998b; Billah and King, 2000b and Grose and King, 1994), the forecasting performance of a criterion

varies with changes of sample size, number of parameters in the model, forecasting horizon, design matrix, and data series being forecast. However, the existing IC based selection procedures are different from one another through their penalty function, which depends on the sample size and the number of parameters in the model. In contrast, if this function is estimated numerically, we expect that it will include other factors (which are ignored in standard IC), at least partially.

Let us assume that $p_1, \cdots, p_N$ denote the penalties corresponding to models $M_1 \cdots, M_N$, respectively. The penalty values will be estimated so that the OAMSE is minimized. We also assume that $e_{i'\ell\ell}|M_i, \theta_{i\ell}, p_1, \cdots, p_N$ denotes the forecast error for the i'th selected model at the ℓth iteration when M_i is the true model (DGP) and $\text{MSE}_{i\ell}|M_i, \theta_{i\ell}, p_1, \cdots, p_N$ is the corresponding mean square error for given $p_1, \cdots, p_N$, where $\theta_{i\ell}$ is the value of θ_i at the ℓth iteration. Thus, the estimated $\text{MSE}_{i\ell}|M_i, \theta_{i\ell}, p_1, \cdots, p_N$ is given by

$$\widehat{\text{MSE}}_{i\ell}(p_1 \cdots, p_N) = \frac{1}{H-h+1} \sum_{t=n^*+h}^{n^*+H} \left(e^2_{i'\ell t}|M_i, \theta_{i\ell}, p_1, \cdots, p_N\right). \qquad (5.4.1)$$

This can be estimated for M_i and θ_i for any given set of p_i, $i = 1, \cdots, N$. If the parameter values are generated randomly from the uniform weighting function $\zeta(\theta_i)$, each drawing of θ_i produces a random value for (5.4.1). Therefore, for a given set of p_i, $i = 1, \cdots, N$, the AMSE_i with respect to θ_i is given by

$$\text{AMSE}_i(p_1, \cdots, p_N) = \int [\text{MSE}_{i\ell}|M_i, \theta_{i\ell}, p_1, \cdots, p_N]\, \zeta(\theta_i) d\theta_i. \qquad (5.4.2)$$

The value of APCS_i varies with changes in the penalty values $p_1, \cdots, p_N$. For any given penalty values, equation (5.4.2) can be estimated by using Monte Carlo integration for R simulated data series for different θ_i, which are generated randomly from the prior distribution $\zeta(\theta_i)$. Through the estimation of optimal penalties, (5.4.1) and hence, (5.4.2) can be estimated easily. Thus, the estimated AMSE_i is

given by

$$\widehat{\text{AMSE}}_i(p_1,\cdots,p_N) = \frac{1}{R}\sum_{\ell=1}^{R}\widehat{\text{MSE}}_{i\ell}(M_i,\theta_{i\ell},p_1,\cdots,p_N), \tag{5.4.3}$$

where $\theta_{i\ell}$ is the value of θ_i generated from $\xi(\theta_i)$ at the ℓth replication.

For the penalty set $p_1,\cdots,p_N$, the OAMSE is given by

$$\widehat{\text{OAMSE}}(p_1,\cdots,p_N) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\text{AMSE}}_i(p_1,\cdots,p_N). \tag{5.4.4}$$

The proposed IC procedure will find the penalties $p_1,\cdots,p_N$ so that (5.4.4) is minimized. This can be done by extending simulation based methods such as PEM-GS and PEM-SA, which were discussed in Chapter 4 in the context of maximizing OAPCS. In the context of forecasting model selection, the extended methods are discussed in the following section.

5.5 Penalty Optimization Methods

In this section, we outline the extension of PEM-GS and PEM-SA (discussed in Chapter 4) in the context of estimating optimal penalties for model selection so that the selected model's OAMSE is minimized. The extended methods are as follows.

5.5.1 PEM-GS Algorithm for Forecasting

As discussed in Section 4.4 of Chapter 4, the PEM-GS algorithm generates a number of penalty sets and the optimum penalty set is the one that yields the highest OAPCS. A detailed description of how the algorithm can be implemented for selecting the forecast model from a group of competing models is as follows. Without loss of generality we assume $p_1 = 0$.

Steps for Estimating Optimal Penalties:

1. Decide on values for b_{il} and b_{iu} such that $b_{il} \leq p_i \leq b_{iu}$, for $i = 2, \cdots, N$, where b_{il} and b_{iu} are preselected lower and upper limits for penalty p_i.

2. Calculate $D_i = b_{iu} - b_{il}$ for $i = 2, \cdots, N$. Let N^* be defined as the total number of all possible combinations from p_i, $i = 2, \cdots, N$, where $N^* = \prod_{i=2}^{N} \left(\frac{D_i}{\delta} + 1\right)$, with δ being an incremental value of b_{il} chosen between 0 and 0.1.

3. For each set of penalties, estimate the OAMSE.

4. Choose $\min_{\ell=1,\cdots,N^*}$(OAMSE) and the corresponding penalty set with elements p_i^*, $i = 2, \cdots, N$.

5. Repeat steps 2 to 4 by redefining p_i with $b_{il}^* \leq p_i \leq b_{iu}^*$, where $b_{il}^* = p_i^* - \delta$, $b_{iu}^* = p_i^* + \delta$ and by using a new incremental value $\delta^* (< \delta)$.

6. Continue step 5 until successive minimum OAMSEs (of step 4) are approximately the same.

As we will see in Section 5.7, PEM-GS works well in selecting the AR forecasting model. However, as mentioned in Chapter 4, the only drawback of this approach is that it requires a large amount of computational time, particularly if the number of models in the plausible group is very large. In the following subsection, we therefore, modify PEM-SA (which was successfully implemented in Chapter 4 to maximize OAPCS) for minimizing OAMSE.

5.5.2 PEM-SA Algorithm for Forecasting

In Chapter 4, the PEM-SA was discussed in the context of estimating optimal penalties so that the OAPCS is maximized. The same method with a slight modification (only in step 4 of the PEM-SA algorithm outlined in Chapter 4) can be used to minimize OAMSE. In the context of a minimization problem, the modification,

particularly in step 4 of the PEM-SA algorithm in Chapter 4, is briefly discussed as follows.

Let $f = f(p_0)$ be the value of the objective function (5.4.4) evaluated at the initial value p_0 for the penalty vector $p = (p_2, \cdots, p_N)'$.

Using step 4 of the PEM-SA algorithm of Chapter 4, a new penalty vector p' is generated by replacing its ith element by p_i. The objective function is evaluated at this new point and is denoted by $f' = f(p')$. If $f' \leq f$, then the new penalty is accepted and p' is stored as p and f' as f. If $f' > f$, the Metropolis criterion decides on acceptance or rejection of the penalty with acceptance probability

$$p_r = \exp\left(\frac{f' - f}{T}\right),$$

where T is the temperature that has been discussed in Section 4.4.2. In fact, a pseudo random number p_u is generated in the range [0,1] and is compared with p_r. If $p_r > p_u$, the new point is accepted, otherwise it is rejected.

Indeed, this method is the same as the PEM-SA algorithm of Chapter 4, but with the above change. We hope this new procedure will perform more efficiently than the regular IC procedures and select models that give the best forecasts on average.

5.6 Design of the Monte Carlo Study

A Monte Carlo study was conducted to compare the performance of various existing IC procedures in selecting the order of AR models for forecasting. The main aim of this study was to investigate how the PEM-SA algorithm works when compared with the existing IC procedures. Four error generating processes, namely, AR(1), AR(2), AR(3) and AR(4) error models were used as the competing error models in the context of the linear regression model (5.2.1). The parameter values were

generated randomly from a weighting distribution, which is essential for estimating AMSE and also for OAMSE. From (5.2.2), for AR(1) processes, $\theta = \rho$ and θ can be generated from a uniform prior $\zeta(\rho)$ on [-1,1]. For AR(2) disturbances, $\theta = (\phi_1, \phi_2)'$, where $\phi_1 = \rho_1 + \rho_2$ and $\phi_2 = \rho_1\rho_2$, while for AR(3) disturbances, $\theta = (\phi_1, \phi_2, \phi_3)'$ with $\phi_1 = \rho_1 + \rho_2 + \rho_3$, $\phi_2 = \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3$ and $\phi_3 = -\rho_1\rho_2\rho_3$. For AR(4) processes, $\theta = (\phi_1, \phi_2, \phi_3, \phi_4)'$, where $\phi_1 = \rho_1 + \rho_2 + \rho_3 + \rho_4$, $\phi_2 = -(\rho_1\rho_2 + \rho_1\rho_3 + \rho_1\rho_4 + \rho_2\rho_3 + \rho_2\rho_4 + \rho_3\rho_4)$, $\phi_3 = \rho_1\rho_2\rho_3 + \rho_1\rho_2\rho_4 + \rho_1\rho_3\rho_4 + \rho_2\rho_3\rho_4$ and $\phi_4 = -\rho_1\rho_2\rho_3\rho_4$. In general, for an AR($p^*$) process, ρ_i, $i = 1, \cdots, p^*$, are roots of (5.2.3) which lie inside the unit circle. Therefore, ρ_i, $i = 1, \cdots, p^*$, can be generated satisfying the condition $|\rho_i| < 1, i = 1, \cdots, p^*$, which also identifies ϕ_i, $i = 1, \cdots, p^*$. For each AR(p^*) process, ρ_i, $i = 1, \cdots, p^*$, were generated randomly from p^* independent uniform distributions each on [-1,1].

Pseudo random numbers were generated using the GAUSS function RNDNS which generates standard normal variates. For estimation and model selection, the normality assumption of the error term was maintained. The parameters of the models were estimated by using maximum MGL. However, the closed form MGL estimator for the parameters is not available for all of the above models. To overcome this difficulty, the GAUSS constrained optimization technique was used to estimate the parameters. For each model, $n = 300$ data values were generated. Among these n, the first n^* were used to estimate and select an appropriate model and the remaining values treated as future observations to calculate the corresponding AMSE for forecast horizons 1 to 4. The experiment was conducted using 2000 replications.

The necessary steps of the experiment are as follows:

1. Using appropriate $\zeta(\theta)$ distributions, at each replication, draw ρ, $(\phi_1, \phi_2)'$, $(\phi_1, \phi_2, \phi_3)'$ and $(\phi_1, \phi_2, \phi_3, \phi_4)'$ values.

2. Generate four samples of size n from each of the AR(1), AR(2), AR(3) and AR(4) error models.

3. For each of the four models, use the first n^* observations to calculate the maximized log-likelihood for each set of data series generated in step 2.

4. Repeat the above steps 2000 times (the number of replications used in this chapter) which will give 32,000 maximized log-likelihoods.

5. Use the maximized log-likelihoods of step 4 and select a forecast model from the competing group by a specific existing IC procedure for each DGP in step 2. Thus, at each replication, four models are selected for four different DGPs. Finally, for each true model, 2000 models are selected. Each of these models are used to calculate h-step ahead forecast MSE using equation (5.3.4). Then, the AMSE as well as the OAMSE of the selected models can be estimated from equations (5.3.6) and (5.3.7), respectively.

6. Using PEM-GS and PEM-SA, which have been discussed in Section 5.5, estimate penalties p_2, p_3 and p_4 (without loss of generality, p_1 can be set to zero). Use these penalties for estimating the AMSE and OAMSE from equations (5.4.3) and (5.4.4), respectively.

The following five design matrices with $n^* = 20, 30$ and 50 were used.

$X1 : n^* \times 2$. A constant dummy and time trend.

$X2 : n^* \times 3$. A constant dummy, the first n^* observations of Durbin and Watson's (1951, p.159) consumption of spirits example.

$X3 : n^* \times 4$. The quarterly Australian consumer price index commencing 1959(1) and the same index lagged one quarter, two quarters and three quarters.

$X4 : n^* \times 5$. A constant dummy, quarterly Australian private capital movements and quarterly Australian Government capital movements.

$X5 : n^* \times 6$. A full set of quarterly seasonal dummy variables plus quarterly seasonally adjusted Australian household disposable income and private final consumption expenditure commencing 1959(4).

The number of observations in the design matrices $X2 - X5$ are less than n. Therefore, to make the design matrices workable for forecasting, we construct a modified design matrix of order $n \times k$ $(n > n^*)$ by using random drawings from the rows of $X(n^* \times k)$. The first n^* observations were used for estimation (model selection) and the next $H = n - n^*$ observations for forecasting.

5.7 Results and Discussion

An important component of time series analysis is forecasting and a selection criterion's forecasting performance needs to be assessed independently of other characteristics of the criterion. Tables 5.1 to 5.5 indicate the AMSE and OAMSE for all selection criteria used in this chapter, including the PEM-SA and PEM-GS. First, we discuss the results in terms of OAMSE. According to the results presented in Tables 5.1 to 5.5, it is evident that for sample size 20, among the existing IC procedures, BIC performs better than the other criteria in selecting the best forecast model for relatively longer forecasting horizons. In general, AIC performs well for one-step-ahead forecasts, but as the forecasting horizon increases, AIC loses its ability to pick on average the best out-of-sample forecast model. Overall, GCV appears to perform moderately well and does relatively better for longer forecasting horizons. MCp performs worst, while HQ works slightly better than MCp. Similar to AIC, HQ does not work well for longer forecasting horizons.

The relative overall performance of BIC declines as the sample size increases.

For example, for design matrix $X3$ with sample size 50, BIC performs better than the other criteria only for four-step-ahead forecasts. In large samples, the selection ability of HQ improves compared to its small sample performance. In contrast, any advantage from using AIC for sample size 20 diminishes as the sample size increases. At $h = 1$, GCV is found to work better for some cases, but its performance deteriorates for longer horizons. Although MCp does better for a few situations for one-step-ahead forecasts, its overall performance remains unchanged compared to its relative performance for sample size 20.

From the results presented in Tables 5.1 to 5.5, irrespective of forecasting horizon and sample size, clearly the new procedures (PEM-SA and PEM-GS) consistently perform better than all the existing IC procedures considered in this cahpter for selecting the best forecast model. This is because the new procedures use data dependent penalties. For example, when the sample size is 20 and $h = 4$, the relative OAMSE values of PEM-GS are, on average 3.07-12.88% larger than those of the existing IC procedures. For larger sample sizes or smaller forecasting horizons, the new procedures dominate all the existing IC procedures comparatively less heavily. This confirms that the new procedures provide much better forecasts in small samples and for longer horizons.

If the forecast results are examined in terms of AMSE, it is evident that for all forecast horizons, the criteria with larger penalties (e.g., BIC) do considerably better than the others when the regression errors are generated from an AR(1) process. The relative performance of other existing criteria varies over design matrices and sample sizes. Interestingly, the new procedures dominate all the existing IC procedures for higher order ($p^* > 1$) data generating processes. However, among the existing IC procedures, generally, no one criterion performs better than the others for all sample sizes and design matrices, particularly for shorter forecasting

horizons. This confirms that the ability of the new procedures in selecting the best forecast model increases if the regression errors come from higher order AR models. When the regression errors are generated from an AR(4) process, the relative AMSE values of PEM-GS are, on average 2.63-18.81% larger than those from the existing IC procedures for sample size 20 and forecasting horizon $h = 4$.

Exceptions to the above generalizations were found for some processes, but the essential conclusion of the study is that the existing IC procedures are not generally equivalent in terms of forecasting performance for small sample observations, and in terms of OAMSE, PEM is a better approach for selecting forecasting models. The performances of PEM-SA and PEM-GS are found to be very similar, although on some occasions the latter performs slightly better than PEM-SA at a high computational cost.

The estimated penalties for PEM-SA are also presented in Table 5.6. This table shows that the estimated penalty values for different AR error models change with change of sample size, design matrix and forecast horizon. However, the penalty values for all of the existing IC procedures are not affected by changing the design matrix and forecast horizon. These results clearly show the importance of the PEM based model selection approach for selecting forecasting models, and hence, we recommend the use of PEM-SA in practice. PEM-GS could also be used, but it requires high amount of computational time.

5.8 Conclusions

In terms of OAMSE, the results of the experiment show that among the existing IC procedures, the recommendation of AIC as a criterion for model selection is moderately well founded for shorter forecasting horizons. However, for longer horizons, the model selected by BIC gives much better forecasts on average than the

model selected by any other existing IC procedures. In general, the criteria HQ, MCp and GCV are not as good as BIC. When the performances of the selection criteria are examined in terms of AMSE, the different criteria are found useful for different forecasting horizons. We observe that the criterion which works best for, say, shorter horizons, may not be so well for longer horizons. This means that the criterion for choosing a model needs to be matched to the given application. In particular, when h-step-ahead ($h > 1$) forecasts are required, we recommend a criterion that minimizes h-steps-ahead ($h > 1$) forecast errors. In contrast, the proposed PEM-GS and PEM-SA consistently dominate all the existing IC procedures considered in this chapter in all situations with very few exceptions. This finding suggests that the model selection criteria based on PEM-GS and PEM-SA are better than any existing IC procedures for choosing a model based on within-sample fit in order to obtain better out-of-sample forecasts. However, we recommend the use of PEM-SA, because it requires much less computation time compared to PEM-GS and performs nearly as well.

Table 5.1: Estimated AMSE and OAMSE (where indicated) for design matrix $X1$ with sample sizes $n = 20$, 30 and 50.

DGP	n	h	AIC	BIC	HQ	MCp	GCV	PEM-GS	PEM-SA
AR(1)	20	1	1.342	1.333	1.341	1.342	1.341	1.344	1.341
AR(2)			1.491	1.534	1.493	1.494	1.493	1.435	1.438
AR(3)			1.566	1.644	1.581	1.580	1.581	1.506	1.513
AR(4)			1.983	2.206	1.977	1.970	1.984	1.797	1.788
OAMSE			1.595	1.680	1.598	1.596	1.600	1.520	1.500
AR(1)	20	2	1.910	1.908	1.909	1.909	1.909	1.911	1.921
AR(2)			2.357	2.402	2.362	2.363	2.360	2.357	2.366
AR(3)			3.061	3.075	3.084	3.089	3.084	3.061	3.043
AR(4)			3.555	3.667	3.509	3.505	3.508	3.467	3.477
OAMSE			2.721	2.763	2.716	2.717	2.715	2.699	2.702
AR(1)	20	3	2.362	2.361	2.361	2.361	2.361	2.369	2.383
AR(2)			3.606	3.632	3.610	3.616	3.608	3.636	3.642
AR(3)			5.994	5.927	6.007	6.033	6.007	5.913	5.902
AR(4)			8.531	8.674	8.472	8.474	8.472	8.357	8.350
OAMSE			5.123	5.148	5.113	5.121	5.112	5.069	5.070
AR(1)	20	4	2.672	2.670	2.671	2.671	2.672	2.682	2.666
AR(2)			4.673	4.694	4.675	4.686	4.672	4.700	4.712
AR(3)			9.699	9.499	9.710	9.770	9.710	9.540	9.534
AR(4)			15.063	15.060	14.936	15.025	14.921	14.763	14762
OAMSE			8.027	7.981	7.998	8.038	7.994	7.921	7.919
AR(1)	30	1	1.231	1.228	1.228	1.229	1.228	1.271	1.261
AR(2)			1.244	1.299	1.244	1.244	1.243	1.231	1.222
AR(3)			1.344	1.349	1.354	1.344	1.349	1.290	1.271
AR(4)			1.370	1.527	1.48	1.368	1.481	1.318	1.321
OAMSE			1.297	1.351	1.329	1.296	1.326	1.277	1.269
AR(1)	30	2	1.764	1.759	1.755	1.757	1.756	1.786	1.798
AR(2)			2.037	2.053	2.037	2.038	2.037	2.028	2.001
AR(3)			2.981	2.691	2.985	2.982	2.988	2.633	2.621
AR(4)			3.012	3.072	3.062	3.015	3.032	3.021	3.042
OAMSE			2.449	2.394	2.459	2.448	2.453	2.367	2.365
AR(1)	30	3	2.183	2.173	2.169	2.172	2.169	2.169	2.177
AR(2)			3.107	3.082	3.106	3.108	3.107	3.084	3.109
AR(3)			6.338	5.198	6.308	6.340	6.352	5.229	5.231
AR(4)			7.264	7.400	7.400	7.275	7.372	7.269	7.258
OAMSE			4.723	4.463	4.746	4.724	4.750	4.438	4.444
AR(1)	30	4	2.499	2.487	2.482	2.486	2.483	2.480	2.493
AR(2)			4.043	4.005	4.043	4.046	4.045	3.975	3.955
AR(3)			11.026	8.429	10.921	11.030	11.051	8.035	8.044
AR(4)			13.345	13.303	13.329	13.363	13.304	13.070	13.104
OAMSE			7.728	7.056	7.694	7.731	7.721	6.891	6.899
AR(1)	50	1	1.076	1.070	1.074	1.076	1.075	1.081	1.101
AR(2)			1.085	1.088	1.084	1.085	1.085	1.084	1.087
AR(3)			1.152	1.142	1.155	1.152	1.153	1.143	1.38
AR(4)			1.185	1.200	1.188	1.186	1.186	1.167	1.156
OAMSE			1.125	1.125	1.125	1.125	1.125	1.119	1.127
AR(1)	50	2	1.440	1.432	1.437	1.440	1.439	1.437	1.444
AR(2)			1.841	1.840	1.839	1.841	1.840	1.833	1.852
AR(3)			2.563	2.373	2.565	2.563	2.563	2.363	2.252
AR(4)			2.797	2.772	2.781	2.797	2.796	2.753	2.749
OAMSE			2.160	2.104	2.155	2.160	2.160	2.096	2.074
AR(1)	50	3	1.687	1.677	1.684	1.686	1.686	1.673	1.684
AR(2)			2.759	2.752	2.756	2.759	2.758	2.745	2.755
AR(3)			5.277	4.570	5.274	5.278	5.276	4.553	4.501
AR(4)			6.643	6.509	6.577	6.643	6.639	6.495	6.496
OAMSE			4.091	3.877	4.073	4.092	4.090	3.866	3.859
AR(1)	50	4	1.864	1.851	1.862	1.864	1.864	1.842	1.833
AR(2)			3.634	3.622	3.631	3.634	3.633	3.615	3.666
AR(3)			8.965	7.344	8.959	8.966	8.964	7.274	7.267
AR(4)			12.289	11.997	12.156	12.289	12.282	11.787	11.791
OAMSE			6.688	6.203	6.652	6.689	6.686	6.129	6.139

Table 5.2: Estimated AMSE and OAMSE (where indicated) for design matrix $X2$ with sample sizes $n = 20$, 30 and 50.

DGP	n	h	AIC	BIC	HQ	MCp	GCV	PEM-GS	PEM-SA
AR(1)	20	1	1.460	1.454	1.458	1.459	1.458	1.493	1.501
AR(2)			1.672	1.683	1.676	1.675	1.676	1.621	1.633
AR(3)			1.809	1.849	1.816	1.810	1.814	1.719	1.705
AR(4)			2.564	2.633	2.567	2.572	2.567	2.505	2.513
OAMSE			1.876	1.905	1.879	1.879	1.879	1.835	1.838
AR(1)	20	2	1.872	1.865	1.867	1.871	1.868	1.885	1.901
AR(2)			2.421	2.414	2.422	2.426	2.424	2.398	2.409
AR(3)			3.176	3.194	3.178	3.177	3.171	3.168	3.188
AR(4)			4.480	4.501	4.470	4.509	4.467	4.329	4.322
OAMSE			2.987	2.993	2.984	2.996	2.982	2.945	2.955
AR(1)	20	3	2.159	2.150	2.153	2.159	2.153	2.156	2.165
AR(2)			3.422	3.375	3.424	3.430	3.428	3.367	3.373
AR(3)			5.780	5.768	5.775	5.781	5.770	5.748	5.735
AR(4)			9.681	9.640	9.649	9.812	9.636	9.418	9.419
OAMSE			5.261	5.233	5.250	5.295	5.247	5.172	5.173
AR(1)	20	4	2.349	2.337	2.340	2.348	2.340	2.336	2.302
AR(2)			4.290	4.190	4.290	4.299	4.296	4.273	4.253
AR(3)			8.544	8.477	8.526	8.544	8.510	7.865	7.855
AR(4)			16.170	15.960	16.110	16.580	16.083	15.540	15.543
OAMSE			7.838	7.741	7.816	7.943	7.807	7.504	7.488
AR(1)	30	1	1.332	1.325	1.330	1.332	1.331	1.342	1.334
AR(2)			1.330	1.353	1.332	1.331	1.329	1.325	1.333
AR(3)			1.410	1.437	1.421	1.411	1.413	1.371	1.371
AR(4)			1.821	1.868	1.830	1.821	1.822	1.767	1.752
OAMSE			1.473	1.496	1.478	1.474	1.474	1.451	4.448
AR(1)	30	2	1.755	1.747	1.753	1.755	1.755	1.762	1.762
AR(2)			2.104	2.123	2.101	2.105	2.101	2.071	2.070
AR(2)			2.584	2.610	2.616	2.585	2.587	2.587	2.601
AR(3)			3.846	3.823	3.842	3.851	3.847	3.603	3.592
OAMSE			2.572	2.576	2.578	2.574	2.572	2.506	2.506
AR(1)	30	3	2.040	2.032	2.039	2.040	2.040	2.043	2.046
AR(2)			3.101	3.110	3.094	3.103	3.099	3.032	3.043
AR(3)			4.705	4.704	4.772	4.708	4.707	4.675	4.688
AR(4)			8.444	8.296	8.413	8.459	8.440	7.569	7.578
OAMSE			4.573	4.535	4.579	4.578	4.571	4.330	4.338
AR(1)	30	4	2.230	2.221	2.228	2.230	2.230	2.226	2.232
AR(2)			4.032	4.042	4.022	4.034	4.029	3.910	3.901
AR(3)			7.132	7.063	7.240	7.138	7.134	6.942	6.932
AR(4)			14.735	14.321	14.650	14.767	14.715	12.736	12.733
OAMSE			7.032	6.912	7.035	7.042	7.027	6.454	6.450
AR(1)	50	1	1.128	1.122	1.125	1.128	1.127	1.134	1.134
AR(2)			1.154	1.159	1.154	1.154	1.154	1.149	1.151
AR(3)			1.196	1.212	1.196	1.197	1.195	1.185	1.192
AR(4)			1.278	1.302	1.286	1.278	1.279	1.263	1.221
OAMSE			1.189	1.199	1.190	1.189	1.189	1.183	1.175
AR(1)	50	2	1.490	1.480	1.484	1.490	1.489	1.486	1.478
AR(2)			1.861	1.859	1.855	1.863	1.861	1.856	1.866
AR(3)			2.290	2.298	2.268	2.290	2.286	2.268	2.252
AR(4)			2.836	2.837	2.832	2.837	2.835	2.816	2.836
OAMSE			2.119	2.119	2.110	2.120	2.118	2.107	2.108
AR(1)	50	3	1.738	1.725	1.729	1.738	1.737	1.730	1.743
AR(2)			2.669	2.651	2.654	2.675	2.669	2.647	2.654
AR(3)			4.151	4.117	4.078	4.152	4.141	4.094	4.032
AR(4)			6.291	6.235	6.258	6.293	6.285	6.196	6.210
OAMSE			3.712	3.682	3.680	3.714	3.708	3.667	3.660
AR(1)	50	4	1.920	1.904	1.908	1.920	1.920	1.906	1.915
AR(2)			3.411	3.378	3.387	3.424	3.410	3.368	3.377
AR(2)			6.332	6.268	6.202	6.333	6.319	6.193	6.243
AR(4)			11.232	11.070	11.153	11.235	11.221	10.921	10.913
OAMSE			5.724	5.655	5.663	5.728	5.718	5.597	5.612

Table 5.3: Estimated AMSE and OAMSE (where indicated) for design matrix $X3$ with sample sizes $n = 20$, 30 and 50.

DGP	n	h	AIC	BIC	HQ	MCp	GCV	PEM-GS	PEM-SA
AR(1)	20	1	2.576	2.505	2.570	2.575	2.571	2.598	2.603
AR(2)			2.989	2.984	2.974	3.006	2.991	2.969	2.974
AR(3)			3.193	3.211	3.182	3.188	3.181	3.169	3.157
AR(4)			4.025	4.119	4.063	4.034	4.061	3.991	3.994
OAMSE			3.196	3.204	3.197	3.201	3.201	3.182	3.182
AR(1)	20	2	2.744	2.668	2.738	2.742	2.740	2.670	2.684
AR(2)			3.477	3.491	3.473	3.497	3.477	3.465	3.472
AR(3)			4.438	4.343	4.414	4.451	4.410	4.314	4.298
AR(4)			5.919	5.921	5.983	5.942	5.972	5.766	5.772
OAMSE			4.145	4.106	4.152	4.158	4.149	4.054	4.057
AR(1)	20	3	3.035	2.964	3.028	3.034	3.030	2.952	2.965
AR(2)			4.413	4.425	4.409	4.439	4.414	4.379	4.379
AR(3)			7.283	7.008	7.234	7.326	7.227	6.985	6.985
AR(4)			11.542	11.225	11.627	11.616	11.550	10.772	10.763
OAMSE			6.568	6.405	6.575	6.603	6.555	6.272	6.273
AR(1)	20	4	3.164	3.097	3.158	3.163	3.161	3.079	3.092
AR(2)			4.928	4.967	4.928	4.958	4.929	4.927	4.941
AR(3)			10.182	9.669	10.103	10.295	10.096	9.552	9.523
AR(4)			8.591	17.635	18.678	18.771	18.409	16.430	16.445
OAMSE			9.216	8.842	9.217	9.297	9.149	8.497	8.500
AR(1)	30	1	1.650	1.631	1.642	1.647	1.645	1.682	1.672
AR(2)			1.843	1.840	1.843	1.842	1.842	1.842	1.841
AR(3)			1.891	1.928	1.901	1.893	1.891	1.868	1.872
AR(4)			2.232	2.261	2.233	2.238	2.233	2.182	2.207
OAMSE			1.904	1.915	1.905	1.905	1.903	1.893	1.898
AR(1)	30	2	1.939	1.918	1.931	1.936	1.934	1.960	1.951
AR(2)			2.558	2.543	2.550	2.556	2.556	2.523	2.514
AR(3)			3.060	3.077	3.062	3.062	3.052	3.031	3.051
AR(4)			4.044	4.001	3.996	4.057	4.005	3.924	3.911
OAMSE			2.901	2.885	2.885	2.903	2.887	2.859	2.857
AR(1)	30	3	2.188	2.171	2.180	2.186	2.184	2.203	2.203
AR(2)			3.602	3.586	3.591	3.599	3.600	3.539	3.512
AR(3)			5.244	5.250	5.231	5.247	5.224	5.153	5.183
AR(4)			8.187	7.968	7.981	8.205	7.992	7.853	7.855
OAMSE			4.805	4.744	4.746	4.809	4.750	4.687	4.688
AR(1)	30	4	2.334	2.317	2.326	2.331	2.330	2.322	2.319
AR(2)			4.551	4.533	4.538	4.546	4.548	4.429	4.428
AR(3)			7.713	7.675	7.675	7.716	7.685	7.450	7.472
AR(4)			13.801	13.247	13.309	13.832	13.322	12.990	13.011
OAMSE			7.100	6.943	6.962	7.107	6.971	6.799	6.808
AR(1)	50	1	1.226	1.219	1.223	1.226	1.226	1.236	1.236
AR(2)			1.271	1.273	1.272	1.271	1.270	1.264	1.281
AR(3)			1.332	1.344	1.335	1.333	1.333	1.322	1.313
AR(4)			1.393	1.417	1.396	1.394	1.393	1.381	1.379
OAMSE			1.306	1.313	1.306	1.306	1.306	1.301	1.303
AR(1)	50	2	1.547	1.538	1.543	1.547	1.547	1.556	1.543
AR(2)			1.929	1.926	1.928	1.929	1.928	1.913	1.923
AR(3)			2.403	2.402	2.395	2.403	2.403	2.377	2.375
AR(4)			2.939	2.953	2.938	2.941	2.942	2.931	2.929
OAMSE			2.204	2.205	2.201	2.205	2.205	2.194	2.193
AR(1)	50	3	1.766	1.756	1.762	1.766	1.765	1.766	1.765
AR(2)			2.687	2.677	2.684	2.687	2.686	2.675	2.663
AR(3)			4.192	4.146	4.142	4.193	4.189	4.132	4.126
AR(4)			6.310	6.319	6.305	6.315	6.315	6.273	6.274
OAMSE			3.739	3.725	3.723	3.740	3.739	3.711	3.707
AR(1)	50	4	1.918	1.907	1.913	1.918	1.917	1.913	1.902
AR(2)			3.375	3.355	3.368	3.374	3.372	3.351	3.367
AR(3)			6.256	6.157	6.153	6.258	6.245	6.188	6.188
AR(4)			11.179	11.179	11.170	11.189	11.186	10.997	10.999
OAMSE			5.682	5.649	5.651	5.685	5.680	5.612	5.613

Table 5.4: Estimated AMSE and OAMSE (where indicated) for design matrix $X4$ with sample sizes $n = 20$, 30 and 50.

DGP	n	h	AIC	BIC	HQ	MC_p	GCV	PEM-GS	PEM-SA
AR(1)	20	1	1.900	1.879	1.897	1.901	1.894	1.962	1.962
AR(2)			2.241	2.224	2.242	2.252	2.247	2.216	2.222
AR(3)			2.532	2.554	2.543	2.540	2.537	2.454	2.476
AR(4)			3.618	3.510	3.632	3.629	3.629	3.488	3.489
OAMSE			2.573	2.542	2.578	2.581	2.577	2.530	2.537
AR(1)	20	2	1 2.286	2.259	2.281	2.286	2.278	2.255	2.254
AR(2)			3.025	2.962	3.011	3.026	3.017	2.956	2.968
AR(3)			4.111	4.051	4.124	4.135	4.113	4.068	4.073
AR(4)			6.246	5.713	6.172	6.242	6.167	5.644	5.635
OAMSE			3.917	3.748	3.897	3.922	3.894	3.731	3.733
AR(1)	20	3	2.542	2.512	2.538	2.543	2.535	2.500	2.533
AR(2)			4.004	3.921	3.985	3.996	4.002	3.894	3.900
AR(3)			7.044	6.850	7.054	7.074	7.036	6.656	6.644
AR(4)			12.150	10.952	11.873	12.117	11.869	10.205	10.219
OAMSE			6.435	6.059	6.363	6.433	6.360	5.814	5.524
AR(1)	20	4	2.718	2.682	2.714	2.719	2.711	2.661	2.646
AR(2)			4.851	4.751	4.832	4.836	4.850	4.674	4.684
AR(3)			10.274	9.838	10.268	10.305	10.237	9.373	9.367
AR(4)			19.674	17.432	18.984	19.600	18.981	15.970	15.999
OAMSE			9.379	8.676	9.200	9.365	9.195	8.171	8.174
AR(1)	30	1	1.503	1.494	1.500	1.503	1.501	1.517	1.536
AR(2)			1.486	1.491	1.492	1.489	1.493	1.481	1.475
AR(3)			1.576	1.610	1.587	1.581	1.580	1.547	1.546
AR(4)			1.724	1.770	1.726	1.709	1.720	1.693	1.701
OAMSE			1.572	1.591	1.576	1.570	1.574	1.559	1.565
AR(1)	30	2	1.908	1.903	1.905	1.908	1.907	1.907	1.904
AR(2)			2.266	2.262	2.272	2.273	2.275	2.243	2.253
AR(3)			2.769	2.805	2.772	2.791	2.763	2.766	2.784
AR(4)			3.559	3.541	3.533	3.543	3.542	3.489	3.469
OAMSE			2.626	2.628	2.621	2.629	2.622	2.601	2.603
AR(1)	30	3	2.159	2.157	2.157	2.160	2.159	2.158	2.153
AR(2)			3.233	3.223	3.241	3.245	3.244	3.182	3.173
AR(3)			4.866	4.896	4.851	4.932	4.845	4.814	4.832
AR(4)			7.637	7.505	7.556	7.617	7.600	7.395	7.410
OAMSE			4.474	4.446	4.451	4.488	4.462	4.387	4.392
AR(1)	30	4	2.337	2.336	2.335	2.337	2.336	2.316	2.337
AR(2)			4.156	4.141	4.168	4.172	4.171	4.056	4.035
AR(3)			7.297	7.297	7.246	7.418	7.242	6.904	6.910
AR(4)			13.352	12.991	13.167	13.354	13.291	12.681	12.679
OAMSE			6.785	6.690	6.729	6.820	6.760	6.489	6.490
AR(1)	50	1	1.165	1.157	1.161	1.165	1.165	1.172	1.178
AR(2)			1.193	1.200	1.195	1.193	1.193	1.194	1.209
AR(3)			1.242	1.254	1.243	1.243	1.242	1.230	1.221
AR(4)			1.277	1.299	1.285	1.277	1.277	1.266	1.263
OAMSE			1.219	1.228	1.221	1.220	1.219	1.215	1.218
AR(1)	50	2	1.524	1.511	1.519	1.524	1.522	1.521	1.510
AR(2)			1.898	1.901	1.895	1.898	1.897	1.895	1.892
AR(3)			2.342	2.345	2.331	2.344	2.337	2.319	2.320
AR(4)			2.849	2.846	2.842	2.850	2.844	2.829	2.820
OAMSE			2.153	2.151	2.147	2.154	2.150	2.141	2.136
AR(1)	50	3	1.768	1.751	1.762	1.769	1.767	1.763	1.74
AR(2)			2.706	2.701	2.696	2.707	2.705	2.695	2.696
AR(3)			4.187	4.153	4.148	4.194	4.173	4.126	4.110
AR(4)			6.314	6.254	6.269	6.318	6.299	6.229	6.231
OAMSE			3.744	3.715	3.719	3.747	3.736	3.703	3.694
AR(1)	50	4	1.949	1.926	1.941	1.949	1.945	1.942	1.943
AR(2)			3.460	3.444	3.437	3.460	3.458	3.443	3.443
AR(3)			6.379	6.298	6.306	6.391	6.353	6.239	6.231
AR(4)			11.265	11.102	11.154	11.270	11.230	11.009	11.018
OAMSE			5.763	5.693	5.709	5.768	5.746	5.658	5.659

Table 5.5: Estimated AMSE and OAMSE (where indicated) for design matrix $X5$ with sample sizes $n = 20$, 30 and 50.

DGP	n	h	AIC	BIC	HQ	MCp	GCV	PEM-GS	PEM-SA
AR(1)	20	1	1.973	1.948	1.969	1.973	1.969	2.069	2.072
AR(2)			2.262	2.317	2.268	2.253	2.266	2.211	2.222
AR(3)			2.437	2.496	2.445	2.446	2.446	2.412	2.421
AR(4)			43.413	3.591	3.441	3.423	3.414	3.331	3.332
OAMSE			2.521	2.588	2.531	2.526	2.524	2.506	2.512
AR(1)	20	2	2.318	2.294	2.314	2.318	2.314	2.291	2.301
AR(2)			2.972	2.969	2.974	3.973	2.977	2.945	2.943
AR(3)			3.882	3.804	3.865	3.890	3.868	3.798	3.798
AR(4)			5.573	5.479	5.580	5.708	5.547	5.456	5.458
OAMSE			3.686	3.636	3.683	3.722	3.675	3.622	3.625
AR(1)	20	3	2.588	2.562	2.586	2.590	2.585	2.552	2.563
AR(2)			4.011	4.011	4.011	4.010	4.009	4.006	4.018
AR(3)			6.726	6.388	6.694	6.740	6.701	6.316	6.309
AR(4)			10.987	10.838	10.998	11.477	10.950	10.440	10.450
OAMSE			6.078	5.950	6.072	6.204	6.061	5.829	5.835
AR(1)	20	4	2.744	2.726	2.744	2.747	2.743	2.675	2.666
AR(2)			4.829	4.813	4.827	4.828	4.825	4.706	4.706
AR(3)			9.732	9.031	9.672	9.751	9.677	8.909	8.917
AR(4)			17.700	17.388	17.723	18.802	17.654	16.182	16.180
OAMSE			8.751	8.489	8.742	9.032	8.725	8.118	8.117
AR(1)	30	1	1.480	1.466	1.474	1.479	1.477	1.507	1.511
AR(2)			1.573	1.581	1.574	1.574	1.571	1.521	1.502
AR(3)			1.599	1.632	1.607	1.601	1.599	1.555	1.566
AR(4)			1.789	1.900	1.796	1.793	1.790	1.751	1.751
OAMSE			1.611	1.645	1.613	1.612	1.609	1.583	1.582
AR(1)	30	2	1.894	1.877	1.886	1.893	1.893	1.894	1.899
AR(2)			2.295	2.279	2.284	2.296	2.286	2.270	2.281
AR(3)			2.810	2.801	2.805	2.813	2.807	2.780	2.777
AR(4)			3.636	3.612	3.637	3.647	3.632	3.565	3.564
OAMSE			2.659	2.642	2.653	2.662	2.654	2.627	2.630
AR(1)	30	3	2.178	2.160	2.172	2.178	2.178	2.209	2.201
AR(2)			3.303	3.242	3.264	3.305	3.272	3.215	3.233
AR(3)			4.972	4.899	4.944	4.977	4.963	4.836	4.821
AR(4)			7.722	7.626	7.706	7.747	7.702	7.551	7.542
OAMSE			4.544	4.482	4.522	4.552	4.529	4.453	4.449
AR(1)	30	4	2.376	2.354	2.368	2.375	2.376	2.341	2.343
AR(2)			4.249	4.136	4.172	4.251	4.189	4.051	4.073
AR(3)			7.473	7.293	7.415	7.481	7.456	7.044	7.031
AR(4)			13.398	13.005	13.356	13.446	13.353	12.969	12.954
OAMSE			6.874	6.697	6.828	6.888	6.844	6.601	6.600
AR(1)	50	1	1.198	1.191	1.194	1.198	1.198	1.204	1.211
AR(2)			1.213	1.220	1.215	1.214	1.213	1.211	1.224
AR(3)			1.273	1.289	1.277	1.273	1.273	1.259	1.248
AR(4)			1.331	1.357	1.340	1.330	1.331	1.310	1.310
OAMSE			1.254	1.254	1.257	1.254	1.254	1.246	1.248
AR(1)	50	2	1.553	1.545	1.548	1.553	1.553	1.561	1.563
AR(2)			1.911	1.916	1.912	1.912	1.909	1.908	1.910
AR(3)			2.378	2.376	2.377	2.379	2.378	2.350	2.346
AR(4)			2.917	2.921	2.911	2.917	2.917	2.876	2.876
OAMSE			2.189	2.189	2.187	2.191	2.189	2.174	2.174
AR(1)	50	3	1.798	1.788	1.791	1.797	1.797	1.795	1.799
AR(2)			2.706	2.705	2.704	2.712	2.703	2.712	2.712
AR(3)			4.235	4.187	4.219	4.241	4.234	4.173	4.172
AR(4)			6.415	6.373	6.375	6.417	6.415	6.296	6.298
OAMSE			3.789	3.764	3.772	3.792	3.787	3.744	3.745
AR(1)	50	4	1.979	1.969	1.972	1.979	1.979	1.971	1.972
AR(2)			3.444	3.437	3.439	3.456	3.437	3.431	3.433
AR(3)			6.422	6.316	6.393	6.434	6.420	6.338	6.334
AR(4)			11.399	11.268	11.297	11.406	11.399	11.120	11.134
OAMSE			5.811	5.747	5.775	5.819	5.808	5.715	5.718

Table 5.6: Estimated penalties by PEM-SA for design matrices $X1$, $X2$ and $X3$, and different forecast horizons.

n	Design Matrix	Forecast Horizon	p_1	p_2	p_3	p_4
20	$X1$	1	0.0	0.201	0.201	1.702
		2	0.0	0.903	1.912	2.835
		3	0.0	1.451	2.542	3.990
		4	0.0	1.451	2.411	3.990
30		1	0.0	0.000	0.201	0.106
		2	0.0	0.501	1.891	3.271
		3	0.0	1.161	2.417	3.321
		4	0.0	1.508	5.101	7.541
50		1	0.0	0.000	0.702	2.261
		2	0.0	0.312	2.143	4.511
		3	0.0	0.312	3.891	6.203
		4	0.0	0.312	5.443	7.802
20	$X2$	1	0.0	0.000	0.702	2.205
		2	0.0	0.518	2.205	4.419
		3	0.0	1.007	2.708	4.917
		4	0.0	1.711	4.000	7.101
30		1	0.0	0.281	0.319	2.904
		2	0.0	0.457	3.115	6.637
		3	0.0	0.405	4.631	8.181
		4	0.0	1.000	5.210	8.713
50		1	0.0	0.610	1.100	2.614
		2	0.0	1.301	1.886	4.771
		3	0.0	1.300	3.401	7.152
		4	0.0	1.600	5.081	8.700
20	$X3$	1	0.0	0.551	1.708	2.709
		2	0.0	1.440	3.919	7.621
		3	0.0	1.700	5.104	8.809
		4	0.0	1.700	6.081	9.700
30		1	0.0	0.110	1.389	2.516
		2	0.0	0.110	2.861	5.507
		3	0.0	0.223	3.100	6.501
		4	0.0	0.863	4.332	8.100
50		1	0.0	0.336	0.708	2.631
		2	0.0	0.411	1.990	4.624
		3	0.0	1.000	2.513	6.324
		4	0.0	1.818	2.103	6.995
20	$X4$	1	0.0	0.100	1.613	2.894
		2	0.0	1.531	4.102	5.990
		3	0.0	2.204	4.835	6.613
		4	0.0	2.221	6.209	9.237
30		1	0.0	0.000	1.219	3.205
		2	0.0	0.806	3.421	6.781
		3	0.0	0.900	3.636	6.610
		4	0.0	2.901	5.603	8.714
50		1	0.0	0.609	1.222	3.541
		2	0.0	1.000	2.457	5.603
		3	0.0	1.000	3.418	5.900
		4	0.0	1.000	3.418	8.261
20	$X5$	1	0.0	0.103	0.916	1.553
		2	0.0	1.523	3.334	5.801
		3	0.0	1.442	4.810	7.892
		4	0.0	2.230	5.872	7.882
30		1	0.0	0.000	1.322	2.612
		2	0.0	0.701	2.883	6.995
		3	0.0	0.000	3.002	6.117
		4	0.0	2.332	5.891	8.993
50		1	0.0	1.205	1.211	1.259
		2	0.0	0.722	1.206	4.340
		3	0.0	1.202	1.737	4.891
		4	0.0	1.600	2.023	7.313

Chapter 6

Exponential Smoothing Model Selection Using ISM

6.1 Introduction

In Chapter 4 and 5, different existing IC procedures and PEM were used for model selection as well as for model selection and forecasting, based on simulated data. Using simulated data might raise the criticism that the real time series data are not so well behaved as the simulated time series. This happens even when random errors and outliers are included in the simulated series. However, the use of real time series data are free from the above criticism. This chapter, therefore, aims to develop a new individual selection method (ISM), based on PEM proposed in Chapter 5, for selecting forecast models for real life time series data.

In this chapter, we consider a forecast model selection problem which we face in many industrial applications including production planning, production scheduling and inventory control. In these applications, hundreds or even thousands of series need to be forecast on a routine basis. The forecaster may either select one appropriate model for all series under consideration, or may build up a method that will select the appropriate model for each series from a group of competing models. The former is known as *aggregate* selection method and the latter as an *individual*

selection method (ISM) (see Fildes, 1989). In this chapter, we develop an ISM, based on PEM.

The motivation of developing an ISM arises from the results of previous studies which showed that no single forecasting procedure or model stands out as the best for producing forecasts for all series in a given data set (see Makridakis and Hibon, 1979; Makridakis et al., 1982, Makridakis et al., 1993; Fildes et al, 1998; Makridakis and Hibon, 2000). Further, the results of Chapters 4 and 5 show that no one existing IC procedure (which selects a model for each data series) uniformly performs better than the others, but the proposed PEM performs better than the existing IC procedures considered in these chapters. Therefore, we consider a PEM based ISM which will simulate the penalty values from the thousands of roughly similar time series that need to be forecast on a regular basis.

Accurate prediction of future events is the ultimate objective of any forecasting method. In the last few decades, many methods ranging from judgemental or intuitive to highly structured and complex quantitative procedures have become available for forecasting. Between these two, there are a very large number of possibilities. These procedures differ in their cost, their accuracy, their complexity and their underlying philosophies. However, because of a lack of information about these differences, objective selection between these methods was extremely difficult until the papers of Makridakis and Hibon (1979) and Makridakis et al. (1982). The purpose of these papers was to assess different forecasting methodologies on the basis of their post-sample predicting accuracy. Reid (1969, 1972) and Newbold and Granger (1974) also compared a large number of series to determine their post-sample accuracy. A shortcoming is that these early studies considered a limited number of forecasting methods in their comparisons.

The major finding of Makridakis and Hibon (1979) and Makridakis et al. (1982)

is that simple methods (e.g., exponential smoothing, damped trend) perform better than those that are more statistically sophisticated. Another finding is that the performances of various methods depend upon the forecast horizons. Further, from the above studies it is evident that the accuracy of the various methods depends on the series being forecasted. Recent studies of Makridakis et al. (1993), Fildes et al. (1998) and Makridakis and Hibon (2000) also support these findings. Therefore, using only one method for all series might be questionable. As there are many competitive methods available, for each series there is considerable choice for selecting a single method from a plausible group. Such a choice may improve forecasting accuracy. From the theoretical standpoint as well as in practice, the implications of making the right choice are very important. In many situations, considerable savings can be made even through small improvements in forecasting accuracy. With this in mind, we consider IC methods for selecting forecasting models which, to the best of our knowledge, have not been applied to the M3 competition data of Makridakis and Hibon (2000).

For seasonal data, three models, namely, the simple exponential smoothing or the local level model (SM1), Holt's exponential smoothing or the local trend model (SM2) and Holt-Winters' exponential smoothing or seasonal model (SM3) are used to form the plausible group. The SM3 model in the above set is replaced by the damped trend model (SM4) for annual data. For comparison, six existing IC procedures, namely, AIC, BIC, HQ, MCp, GCV and FPE are considered. A combination of forecasts (COM), which is a simple average of methods in the plausible group, is also included in the comparison (see Makridakis et al., 1982). Makridakis et al. (1982) and Makridakis and Hibon (2000) showed that the COM method performs better than individual methods considered in the combination. Further, on the basis of PEM (either of PEM-GS and PEM-SA), an ISM, which is the main con-

tribution of this chapter, is proposed for model selection for forecasting for the M3 competition data, where the true model is unknown.

The plan of this chapter is as follows. The point forecasts of the models under consideration are presented in Section 6.2. The COM method for forecasting is outlined in Section 6.3. Section 6.4 includes the description of an ISM for penalty estimation for selecting a forecast model. The M3 competition data are discussed in Section 6.5. The estimation method and design of experiment are discussed in Section 6.6. The results of the study are discussed in Section 6.7 and the final section contains some concluding remarks.

6.2 The Models and Their Point Forecasts

As mentioned earlier, in this chapter we use the SM1, SM2, SM3 and SM4 models as the forecast methods. These models were chosen because they are easy to understand, non-linear in nature and very familiar to practitioners. Further, these models produce better forecasts for real life time series data (see Makridakis et al., 1982; Makridakis et al., 1993; Fildes et al., 1998, Makridakis and Hibon, 2000). The point forecasts of these models are given as follows.

The state space form of the above models, as discussed in Section 3.2 of Chapter 3, is given by

$$y_t = x_t'\beta_{t-1} + e_t, \tag{6.2.1}$$

$$\beta_t = T^*\beta_{t-1} + \alpha e_t, \tag{6.2.2}$$

$$\beta_0 = \beta_0^*, \tag{6.2.3}$$

where x_t, β_t, T^*, α and β_0^* are as defined in Section 3.2 of Chapter 3. Equations (6.2.1) and (6.2.2) are known as the measurement and transition equations, respec-

tively. At the end of period n, the transition equation can be written as follows:

$$\beta_n = T^* \beta_{n-1} + \alpha e_n. \tag{6.2.4}$$

Now, recursive substitutions ($h-1$ times) for β_n in the right hand side of (6.2.4) yields the equation:

$$\beta_n = T^{*h} \beta_{n-h} + \alpha \sum_{j=0}^{h-1} T^* e_{n-j}, \tag{6.2.5}$$

$$\text{or,} \quad \beta_{n+h} = T^{*h} \beta_n + \alpha \sum_{j=0}^{h-1} T^* e_{n+h-j}. \tag{6.2.6}$$

From (6.2.1) and (6.2.6), the h-step ahead forecast $y_n(h)$ is given by

$$y_n(h) = x_t' T^{*h-1} \beta_n, \tag{6.2.7}$$

where x_t is independent of time.

Substituting appropriate values of x_t, T^* and β_n for the SM1, SM2, SM3 and SM4 models, the point forecasts for these models are given, respectively, by

$$y_n(h) = l_n, \tag{6.2.8}$$

$$y_n(h) = l_n + hb_n, \tag{6.2.9}$$

$$y_n(h) = l_n + hb_n + c_{n+h-s}, \quad c_{n+h} = c_{n+h-s}, \tag{6.2.10}$$

$$y_n(h) = l_n + b_n \sum_{i=0}^{h-1} \phi^i, \tag{6.2.11}$$

where s is the length of seasonality, c_t is the seasonal index at time t and l_n and b_n respectively, are the level and trend at period n.

As mentioned in Chapter 5, MSE and MAPE are the most popular and widely used forecast evaluation criteria. For real life data, the calculated forecast MSE often becomes very large, and hence, may be difficult to manage and also interpret (see Makridakis et al., 1982). Further, MAPE, or equivalently symmetric mean absolute percentage error (SMAPE) was used as the forecast evaluation criterion

in all of the previous studies on the M (Makridakis et al., 1982), M2 (Makridakis et al., 1993) and M3 competition data as well as for the sub-sample of the M competition data (Makridakis et al., 1982). Therefore, to avoid large errors and to enable comparison of the findings of our study with those of the previous studies, in this chapter, we use MAPE as the forecast evaluation criterion.

6.3 Combining Forecasts (COM)

In this section, we outline the COM method in the context of the SM1, SM2 and SM3 models. Once the seed values (for example, for the SM2 model, the seed values are l_0 and b_0) and the smoothing parameters of the models are estimated, the point forecasts corresponding to the SM1, SM2 and SM3 models are generated using the formulae given in (6.2.8) to (6.2.10), respectively. Let us assume that $y_n^{SM1}(h)$, $y_n^{SM2}(h)$ and $y_n^{SM3}(h)$ denote the point forecasts for the SM1, SM2 and SM3 models, respectively. Based on the forecast methods (models) under consideration, there are two types of COM methods, namely, the simple average of methods and the weighted average of methods (see Makridakis et al., 1982). The weights are calculated on the basis of the sample covariance matrix of percentage error for all forecasting methods under consideration for the model fitted to each series (for more details, see Makridakis et al., 1982). Between these two, the former is simple, easy to understand and provides better forecasts than does the latter method (see Makridakis et al. 1982). Therefore, in this study, we consider the simple average of forecasts, which is given as follows:

$$y_n(h) = \frac{1}{3}\left(y_n^{SM1}(h) + y_n^{SM2}(h) + y_n^{SM3}(h)\right). \tag{6.3.1}$$

From the study of Makridakis et al. (1982), this COM method was found to perform better than the individual methods used in the combination. This method can be

generalized easily for any number of models under consideration.

6.4 Individual Selection Method (ISM) Via PEM

In this section, we develop an ISM using PEM-GS, where PEM-GS was discussed in Chapter 5. As mentioned in Chapter 5, the required computational time for PEM-GS was very high, particularly for a refined search. However, the computational cost is less of an issue when using PEM-GS to estimate penalties for ISM. This is because the structure of the current problem is very simple, compared to that in Chapter 5. For example, the number of maximized log-likelihoods that were penalized in the previous chapter is N^2R, while this number reduces to NR when dealing with real data, where N is the number of competing models and R is the total number of series. Further, a simple grid can find a penalty set that selects the model with the smallest forecast error for most of the series.

ISM is a two stage method because it requires estimation of the penalties first, and then we use these estimated penalties to select models for forecasting future observations. This method is appropriate for a particular set of time series. This means, different types of time series data such as annual, quarterly and monthly data will be treated separately by ISM. The selection criteria of ISM is the forecast accuracy measure such as MAPE rather than the selection percentage or probability of correct selection of a model. In general, for real life data, the true model is not known. Therefore, for a particular data series, we select an appropriate model from a group of competing models. Detailed description of this method is as follows.

In ISM we treat annual, quarterly and monthly data separately. Let us assume that we are dealing with a particular class of series containing R series and each series has n_i observations, $i = 1, 2, \cdots, R$ (the number of observations may vary from one series to another). Initially, each series is divided into two segments, as

was done by Makridakis and Hibon (1979), Makridakis et al. (1982), Makridakis et al. (1993), Fildes et al. (1998) and Makridakis and Hibon (2000). Let n_i^* be the number of observations in the first segment and H be the number of observations held back in the second segment. The value of H is determined by the forecaster, according to the number of future forecasts to be made. The first n_i^* observations are used to estimate the competing models and their corresponding maximized log-likelihoods. The rest of H observations are used for forecast evaluation. The models are estimated by maximizing MGL, and hence, maximized log-likelihoods (log of ICL) are calculated.

Let us assume that for a particular type of data set, $p_1, p_2, \cdots, p_N$ are penalties to be estimated for the $M_1, M_2, \cdots, M_N$ models, respectively. This means, the penalty values for quarterly, monthly and annual data will be estimated by running ISM for each data set separately. Without loss of generality, we assume $p_1 = 0$. Then, the steps involved in the ISM for a particular class of R series are as follows:

Stage One:

Step 1: By using the first $\breve{n}_i^* = n_i^* - 1$ observations of the first segment, all the competing models are estimated and the maximized log-likelihoods are calculated. For all models, a one-step ahead forecast is made at origin $\breve{n}_i^*$. Then, for each competing model we calculate $\text{APE}_i(1) = \left(|y_{\breve{n}_i^*+1} - y_{\breve{n}_i^*}(1)|\right) / y_{\breve{n}_i^*+1}$, where $\text{APE}_i(1)$ is the one-step ahead forecast absolute percentage error (APE) for the ith series. Thus, for each competing model, the maximized log-likelihood and APE are calculated for every series under consideration.

Step 2: Following Step 1 and Step 2 of PEM-GS in Section 5.5, N^* penalty sets are generated. Let us denote these penalty sets by $p_1^j, \cdots, p_N^j$, $j = 1, \cdots, N^*$. Then, the MAPE for the jth penalty set is calculated by

$$\text{MAPE}_j(1)(p_1^j, \cdots, p_N^j) = \frac{1}{R} \sum_{i=1}^{R} \text{APE}_i(1)(p_1^j, \cdots, p_N^j),$$

where $\text{APE}_i(1)$ corresponds to the selected model for the ith series. Then, we choose $\min_{j=1,\cdots,N^*}(\text{MAPE}_j(1))$ and the corresponding penalty set with elements $p_1, \cdots, p_N$. Similar to AIC, in this case, we have a single penalty set for a particular type of data set under consideration. The penalty values are recorded.

Stage Two:

Step 3: For each series, the competing models are fitted for the first n_i^* observations and the corresponding maximized log-likelihoods are calculated and are recorded (we have R sets of maximized log-likelihood for R data series).

Step 4: The maximized log-likelihoods (recorded in step 3) are penalized by the penalties estimated (recorded) in step 2, and for each series, a model is selected from the competing models. Each selected model is then used for making h-step ahead forecasts at origin n_i^*, $h = 1, \cdots, H$. Thus, APE is calculated for all forecast horizons for each data series under consideration. Finally, for comparison with the existing IC procedures, for each forecast horizon h, an average (for APE) is taken across all R series which gives MAPE for the corresponding h, i.e.,

$$\text{MAPE}(h)(p_1, \cdots, p_N) = \frac{1}{R}\sum_{i=1}^{R} \text{APE}_i(h)(p_1, \cdots, p_N), \ h = 1, \cdots, H.$$

6.5 Data Description

In this chapter, we discuss the M3 competition data of Makridakis and Hibon (2000). This data set is a large and diverse set of time series collected from various sources. This data set has been extensively used in many experiments for forecast accuracy measures and also for comparing the accuracy of univariate forecasting procedures. The popularity of this data set is because it has become a sort of benchmark. It is rare that such a data set is readily available. Therefore, in this chapter, we decided to apply the ISM as well as the existing IC procedures to this

Table 6.1: Major categories of the M3 competition data.

Data Type	Frequency of Observations				
	Annual	Quarterly	Monthly	Other	Total
Micro	146	204	474	4	828
Industry	102	83	334		519
Macro	83	336	312		731
Finance	58	76	145	29	308
Demographic	245	57	111		413
Other	11		52	141	204
Total	645	756	1428	174	3003

data set for selecting a model from a group of competing exponential smoothing models.

The M3 competition includes 3003 series which were selected on a quota basis to include different types of time series data, such as macro, micro, industry, etc., and different time periods such as annual, quarterly and monthly. The minimum number of observations for annual, quarterly, monthly and other series was set at 14, 16, 48 and 60 respectively. The reason for such a restriction was to ensure enough data to develop adequate forecasting models. The classification of 3003 series is given in Table 6.1.

Let us assume that n_i, $i = 1, 2, \cdots, R$, be the number of observations in each series. Following Makridakis and Hibon (1979), Makridakis et al. (1982), Makridakis et al. (1993) and Fildes et al. (1998), Makridakis and Hibon (2000) divided each series into two segments. For monthly series, the models were estimated from the first $n_i - 18$ observations and the post-sample forecasts were made at origin $n_i - 18$ for forecasting horizons 1 to 18. Forecast generation was independent of post-sample observations. Similarly, for quarterly and annual series, the models were estimated from the first $n_i - 8$ and $n_i - 6$ observations, respectively. Then, forecasts were made at origins $n_i - 8$ and $n_i - 6$, respectively, and the post-sample

forecasts were compiled by horizon.

6.6 Estimation Method and Design of Experiment

By using various existing IC procedures, we selected a forecast model from a set of competing models for each series under consideration. The ISM was also used for model selection and forecasting. As we know, the IC based model selection procedures are based on the estimated maximized log-likelihood function of the corresponding data series. Therefore, all the models in the plausible group need to be estimated using a suitable estimation method. Among others, the CL method proposed by Ord et al. (1997) is the only likelihood approach in the literature that can be used directly to estimate exponential smoothing models. Also, after some suitable transformations of the original exponential smoothing models, the MGL method can be used to estimate these models, which was discussed in Chapter 3.

In Chapter 3, we compared the CL and ICL based six versions of widely used existing IC procedures. The simulation results of this chapter show that gains can be made by using the IC procedures, based on the ICL method (which in turn is based on MGL estimates). Further, from Chapter 4, it is evident that MGL based IC procedures perform better than those based on the PL. In this chapter, we therefore used ICL based IC procedures. For some models, the time taken for estimation of parameters is very high, particularly for the Holt-Winters' model with monthly data. In this case, there are 13 initial values (states) and 3 smoothing parameters to estimate, and searching for optimal values in a space of 16 dimensions is very time consuming. Therefore, the OLS method as discussed in Chapter 3 was used to estimate the seed state vector (initial values). Then, using the GAUSS (see Aptech, 1996) constrained optimization algorithm, the smoothing parameters were

estimated, holding the seed state vector at the OLS values. Also, the seed state vector was re-estimated by using the optimized value of the smoothing parameters.

For optimization, the exponential smoothing parameters were restricted to lie within the following intervals:

$$\begin{aligned} &\text{level parameter:} && 0.0 \leq \alpha_1 \leq 0.99, \\ &\text{trend parameter:} && 0.0 \leq \alpha_2 \leq 0.99, \\ &\text{seasonal parameter:} && 0.0 \leq \alpha_3 \leq 0.99, \\ &\text{damped parameter:} && 0.0 \leq \phi \leq 0.99. \end{aligned}$$

The seed state vector (see Chapter 3) was also restricted so that the seasonal indices of the Holt-Winters' model add to zero. Once the likelihood is optimized, the models can be selected by using the existing IC procedures and the ISM. Then, the post sample forecasts for quarterly, monthly and annual data were made at origins $n_i - 8$, $n_i - 18$ and $n_i - 6$, respectively (Makridakis and Hibon, 1979; Makridakis et al., 1982; Makridakis et al., 1993; Fildes et al., 1998; Makridakis and Hibon, 2000).

6.7 Results and Discussion

The results from the M3 competition data are presented in Tables 6.2 to 6.6 and Figures 6.1 to 6.4. Table 6.2 and Figure 6.1 show the selection performance of various existing IC procedures and the ISM. Tables 6.3 to 6.5 include the MAPEs, and the graphical presentation of these MAPEs are shown in Figure 6.2. The differences (ISM−Best Existing IC) between selection percentages are shown in Figure 6.3. The estimated penalties for the M3 data are presented in Table 6.6 and in Figure 6.4. Detailed discussion of the results is as follows.

The selection percentage of various models by the existing IC procedures for different types of series are presented in Table 6.2 and Figure 6.1. Table 6.2 and

Table 6.2: Selection percentages of different models by various IC procedures and ISM for the M3 competition data series.

Data Type	Models	AIC	BIC	HQ	MCp	GCV	FPE	ISM
Quarterly	SM1	11.24	18.65	13.89	11.11	12.04	11.51	35.05
	SM2	21.16	26.85	23.28	19.58	22.62	21.43	25.26
	SM3	67.59	54.50	62.83	69.31	65.34	67.06	39.68
Monthly	SM1	1.33	5.74	2.94	1.26	1.40	1.33	52.10
	SM2	5.60	3.43	1.68	4.90	5.60	5.60	16.81
	SM3	98.11	90.83	95.38	98.25	98.04	98.11	31.09
Annual	SM1	17.46	24.05	18.07	16.85	19.66	17.83	32.17
	SM2	29.79	37.48	31.01	27.84	35.53	30.28	1.10
	SM4	52.75	38.46	50.92	55.31	44.81	51.89	65.93

Figures 6.1(a) and 6.1(b) show that for quarterly and monthly data, the SM3 model is selected most often, followed by the SM2 model, and the SM1 model is chosen least often. Exceptions are found for monthly series where BIC and HQ select the SM2 model least often. Compared to the existing IC procedures, the ISM selects the single exponential smoothing model more often. For example, as shown in Table 6.2, for quarterly and monthly data, BIC selects the SM1 model 18.65% and 5.74% of the time respectively. On the other hand, for these data, the ISM selects the SM1 model 35.05% and 52.10% of the time respectively. Table 6.2 and Figure 6.1(c) show that for annual data, all the existing IC procedures and the ISM selects the SM4 model more often than any other model in the competitive group. Again, the ISM selects the SM4 model more often than the existing IC procedures. The results from the ISM are quite consistent with the findings of Makridakis and Hibon (2000), where, in general, simple methods such as Brown's single and Gardner's damped trend exponential smoothing were found to perform as well as, or better, than some statistically sophisticated ones.

For the quarterly M3 data, as indicated by Table 6.3, it is clear that among

Table 6.3: MAPE for forecast horizons 1 to 8 of models selected by various IC procedures, ISM and the COM method for the quarterly M3 competition data series.

h	AIC	BIC	HQ	MCp	GCV	FPE	ISM	COM
1	5.85	5.78	5.87	5.87	5.84	5.86	5.77	7.35
2	7.62	7.53	7.63	7.59	7.62	7.62	7.12	7.96
3	10.02	9.79	10.00	10.00	9.98	10.02	9.74	11.12
4	9.44	9.26	9.45	9.44	9.40	9.44	8.95	10.06
5	13.12	12.96	13.19	13.12	13.10	13.12	11.41	13.35
6	13.83	13.54	13.84	13.80	13.77	13.83	11.56	13.25
7	15.77	15.36	15.74	15.75	15.66	15.77	13.17	15.45
8	27.04	26.55	26.97	27.03	26.93	27.05	24.38	23.56

the existing IC procedures, in terms of MAPE, BIC performs better for all forecast horizons. All the existing IC procedures used in this chapter are better than the COM method for the forecast horizons 1 to 5. The MAPE of BIC, the ISM and the COM methods are also presented in Figure 6.2(a) which shows that the ISM performs better than the COM method for forecast horizons 1 to 7, and better than BIC for all horizons. For the quarterly M3 competition data, compared to the existing IC procedures, the performance of the ISM improves as the forecast horizon increases.

The MAPE for the monthly M3 competition data are presented in Table 6.4, which shows that in general, the MAPE of BIC is slightly lower than that for other existing IC procedures at each forecast horizon. Indeed, there is no significant difference between the MAPE of various IC procedures. Figure 6.2(b) and Table 6.4 show that the COM method performs better than the existing IC procedures for all forecast horizons longer than 2. Clearly, irrespective of forecast horizon, the ISM performs better than the COM method and the existing IC procedures. As the forecast horizon increases, the ISM improves relative to the existing IC procedures and the COM method.

Table 6.4: MAPE for forecast horizons 1 to 18 of models selected by various IC procedures, ISM and the COM method for the monthly M3 competition data series.

h	AIC	BIC	HQ	MCp	GCV	FPE	ISM	COM
1	15.54	15.45	15.48	15.54	15.55	15.54	12.86	15.58
2	12.76	15.95	15.81	15.93	15.95	15.95	12.76	16.14
3	16.49	16.37	16.41	16.49	16.48	16.49	13.24	15.88
4	20.10	20.00	20.07	20.10	20.09	20.10	16.21	18.70
5	19.77	19.32	19.34	19.77	19.76	19.77	16.07	18.50
6	20.08	19.77	19.95	20.08	20.06	20.08	16.19	19.22
7	20.23	20.13	20.13	20.23	20.22	20.23	16.05	19.95
8	23.66	23.53	23.60	23.66	23.65	23.66	19.18	22.36
9	25.45	25.26	25.28	25.45	25.44	25.45	21.36	24.14
10	22.61	22.49	22.44	22.61	22.58	22.61	17.83	21.22
11	28.63	28.41	28.53	28.64	28.61	28.63	24.22	26.16
12	25.67	25.50	25.48	25.67	25.64	25.67	20.53	24.34
13	34.91	34.76	34.72	34.91	34.87	34.91	26.62	31.47
14	30.59	30.36	30.36	30.42	30.59	30.56	22.27	26.76
15	30.23	29.86	30.06	30.23	30.20	30.23	21.69	26.62
16	28.34	28.06	28.15	28.34	28.35	28.43	19.55	25.46
17	29.10	28.76	28.82	29.10	29.06	29.10	19.76	25.68
18	31.23	31.26	31.17	31.23	31.19	31.23	22.62	27.64

Table 6.5 and Figure 6.2(c) show the MAPE for the annual M3 competition data. Among the existing IC procedures, in general, MCp performs better in terms of MAPE. Surprisingly, the COM method is better than all existing IC procedures. However, irrespective of forecast horizon, the performance of ISM is better than the COM method, and hence, better than all existing IC procedures. Starting from $h = 2$, as the forecast horizon increases, the difference between MAPE of the ISM and the existing IC procedures (for example, MCp) increases (see Figure 6.2(c)).

Further, for the M3 competition data, the differences between the selection performances of the ISM and the best existing IC procedure (for quarterly and monthly data, BIC is better and for annual data, MCp is better) are presented in Figure 6.3. This figure shows the differences of all the existing IC from the ISM

Table 6.5: MAPE for forecast horizons 1 to 6 of models selected by various IC procedures, ISM and the COM method for the annual M3 competition data series.

h	AIC	BIC	HQ	MCp	GCV	FPE	ISM	COM
1	19.36	19.40	19.47	19.31	19.33	19.29	16.03	18.38
2	19.05	19.07	19.08	18.89	19.13	19.01	17.59	18.67
3	19.67	19.70	19.64	19.52	19.63	19.59	17.28	18.85
4	26.20	27.04	26.99	26.92	27.05	26.17	23.82	26.03
5	26.01	26.09	25.99	25.75	26.11	25.97	22.20	24.85
6	28.55	28.71	28.45	28.27	28.78	28.59	24.53	27.35

Table 6.6: Estimated penalties by ISM for the M3 competition data series.

Data Type	Estimated Penalties for ISM		
	p_1	p_2	p_3
Quarterly	0.0	4.500	22.250
Monthly	0.0	3.750	6.875
Annual	0.0	6.425	1.925

for all forecast horizons. These differences are very high (negatively), particularly for the SM3 (for quarterly and monthly data) and SM2 (for annual data) models. This means that depending on the type of data, the existing IC procedures have a tendency to select the SM3 and SM2 models more often than the ISM.

The penalty values for AIC and those estimated by the ISM for the M3 competition data are presented in Figure 6.4 and Table 6.6, which show that the estimated penalties for larger models are much higher than those of AIC for quarterly and monthly data. This is also true for other existing IC procedures. This leads the ISM to select simple models more frequently than the existing IC procedures. In fact, the existing IC procedures cannot penalize sufficiently the log-likelihood function of larger models (for example, seasonal model SM3), and hence, have a tendency to select bigger models. Therefore, the ISM provides better forecasts than the existing IC procedures as it is able to penalize the maximized log-likelihood functions

more efficiently than the existing IC procedures. For annual data, compared to the existing IC procedures, the ISM penalizes the SM2 model more heavily, but it penalizes the SM4 model slightly less. However, this does not occur for any of the existing IC procedures.

6.8 Conclusions

We have proposed an ISM for selecting forecast models for real life time series data and have applied it to the M3 competition data, which we subdivided in to three sets of data, namely, annual, quarterly and monthly data. The forecasting accuracy of the ISM, the COM method and the existing IC procedures have been compared. Compared to the existing IC procedures, for quarterly and monthly data, typically, BIC performs better and for annual data, overall, MCp performs better. In general, the COM method performs better than the existing IC procedures with an exception for quarterly data, where the latters perform better than the COM method for $h \leq 4$. Irrespective of forecast horizon, in general, the ISM performs better than the existing IC procedures and the COM method for all forecast horizons. The ISM penalizes the larger models more efficiently than the existing IC procedures, and hence, selects appropriate models more often than that of the existing IC procedures. This performance of the ISM is in agreement with the previous study on the M3 competition time series data, where the simple exponential smoothing model and the damped trend model were found to perform as well as or better than Holt's linear trend model and the Holt-winters' seasonal model, particularly for longer forecast horizons. Although, the division of the M3 data set into annual, quarterly and monthly data was fairly rough, yet the ISM produced better results compared to the existing IC procedures and the COM method. We are optimistic that it might work even better for typical industrial applications, where thousands

of very similar data series need to be forecast on a regular basis.

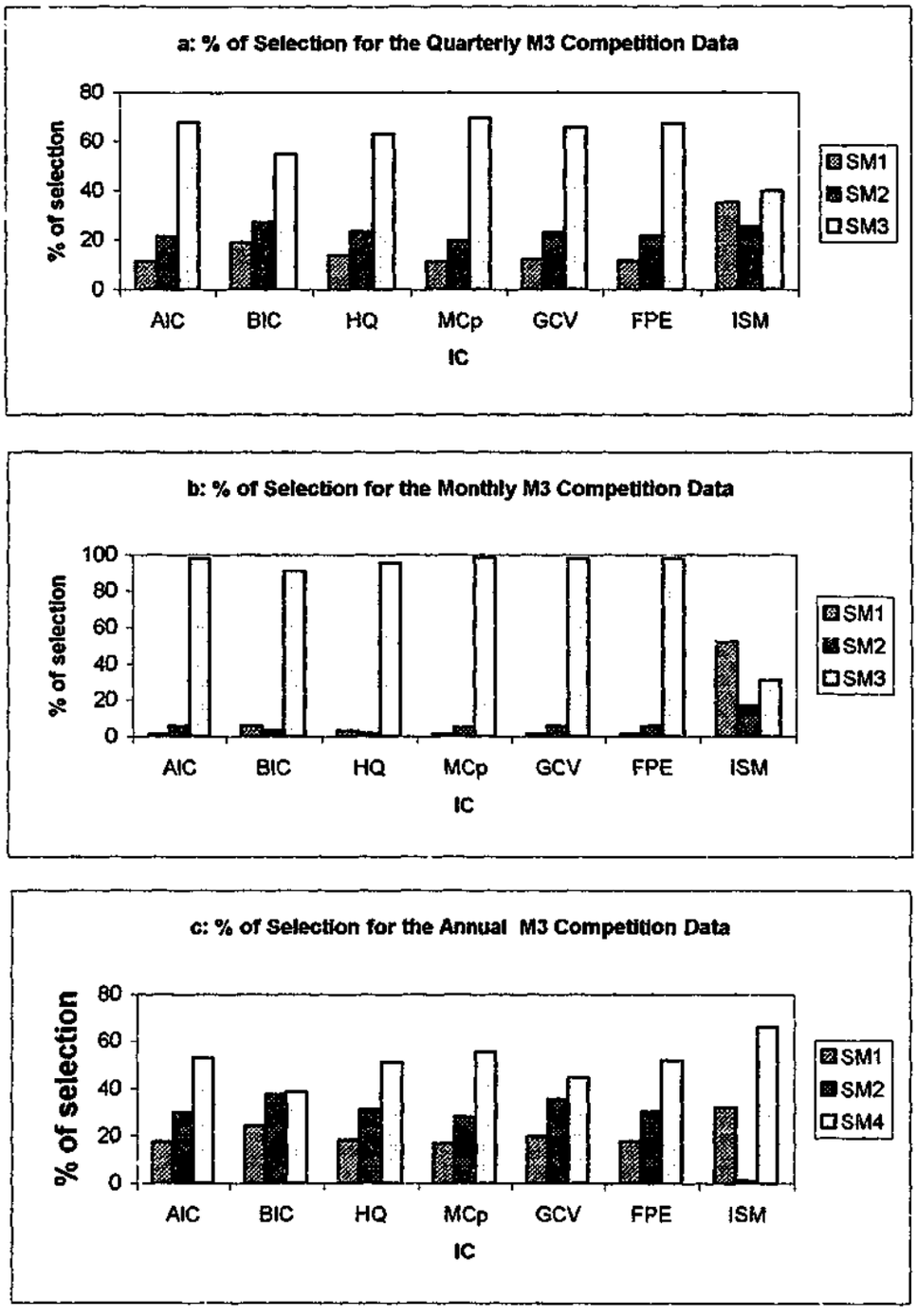

Figure 6.1: Selection percentages for different models by the ISM and various existing IC procedures across all forecast horizons for the M3 competition data.

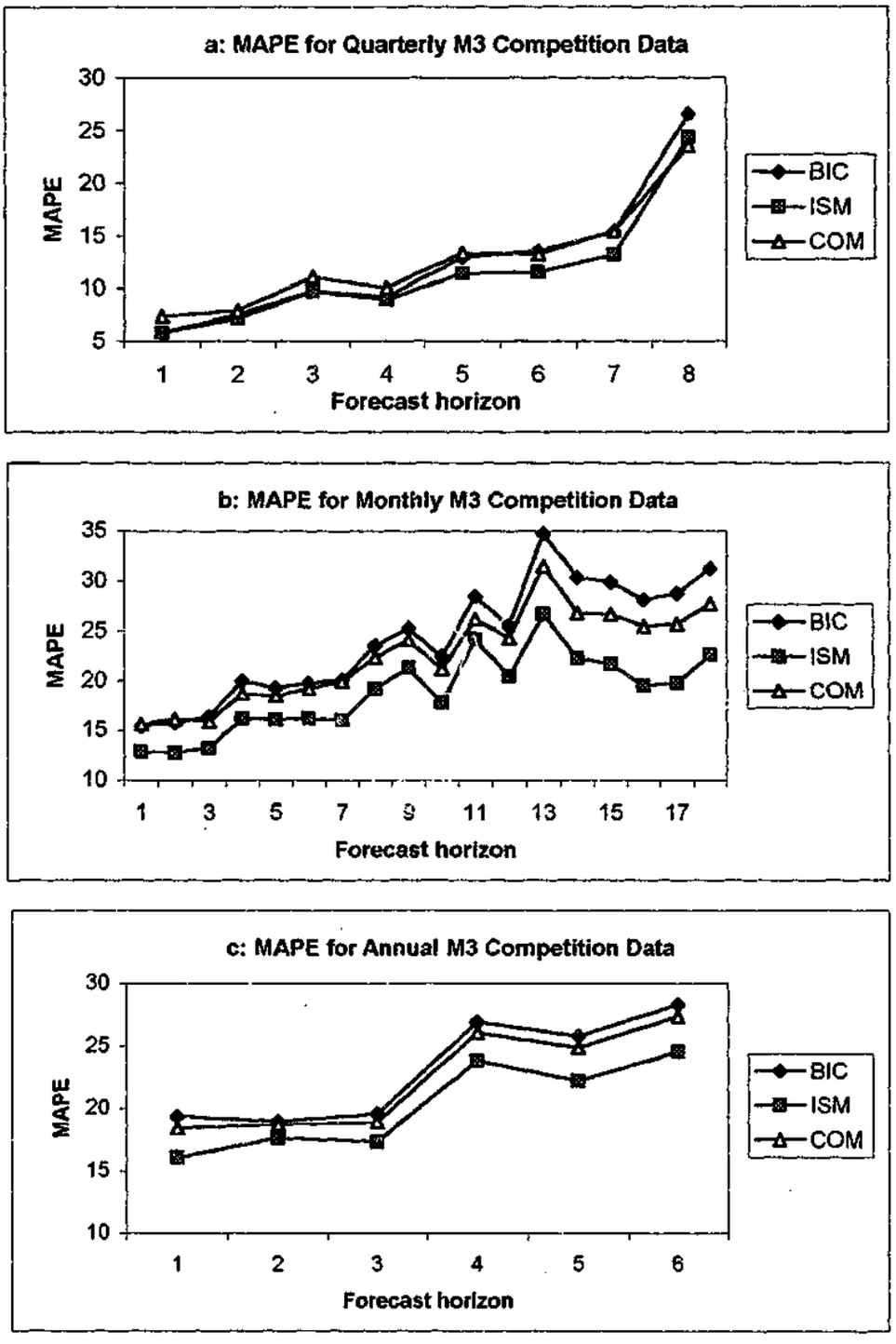

Figure 6.2: Comparison between MAPE of BIC, the ISM and the COM method across different forecast horizons for the M3 competition data.

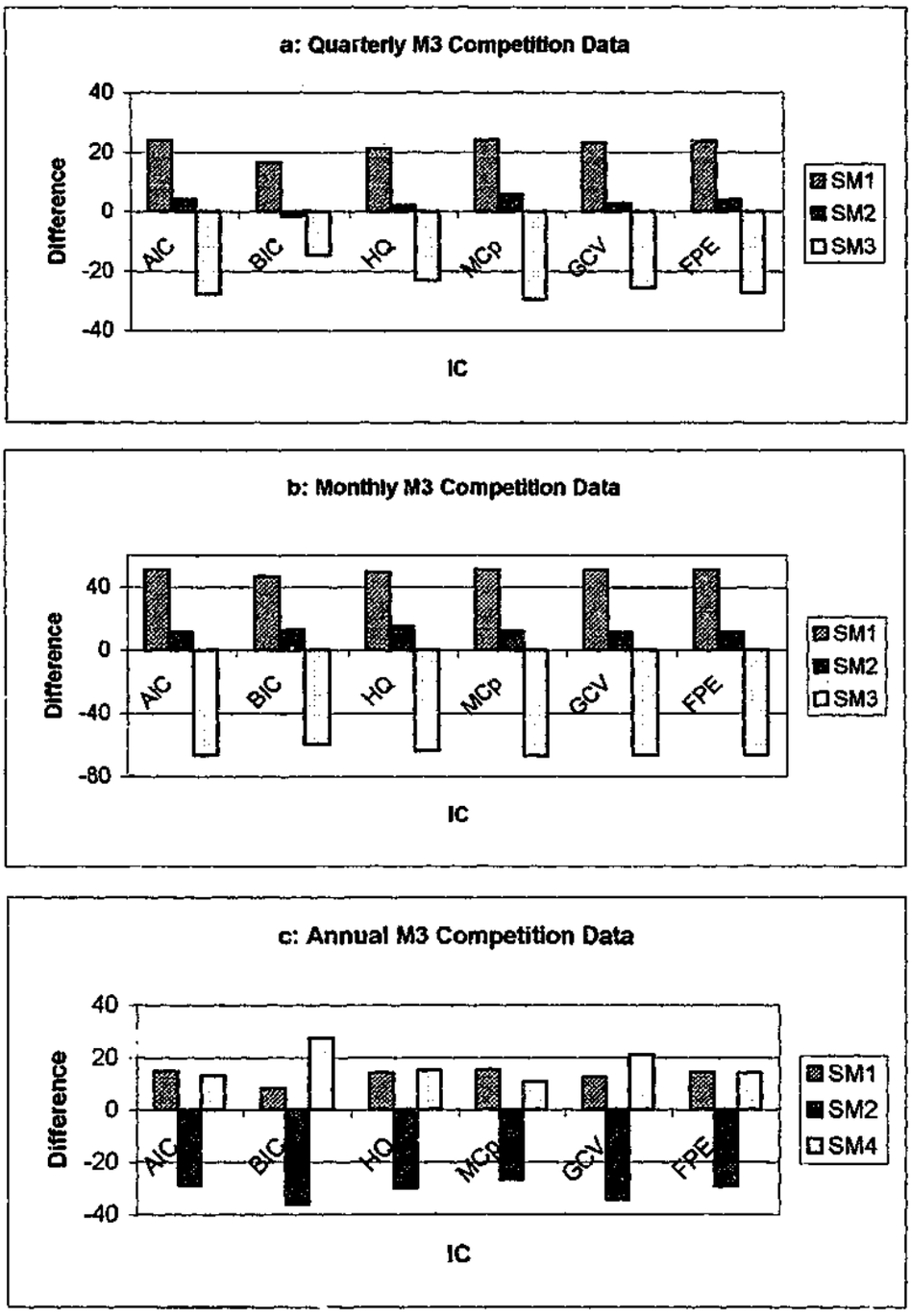

Figure 6.3: Difference (ISM minus BIC) of selection percentages for different models across all forecast horizons for the M3 competition data.

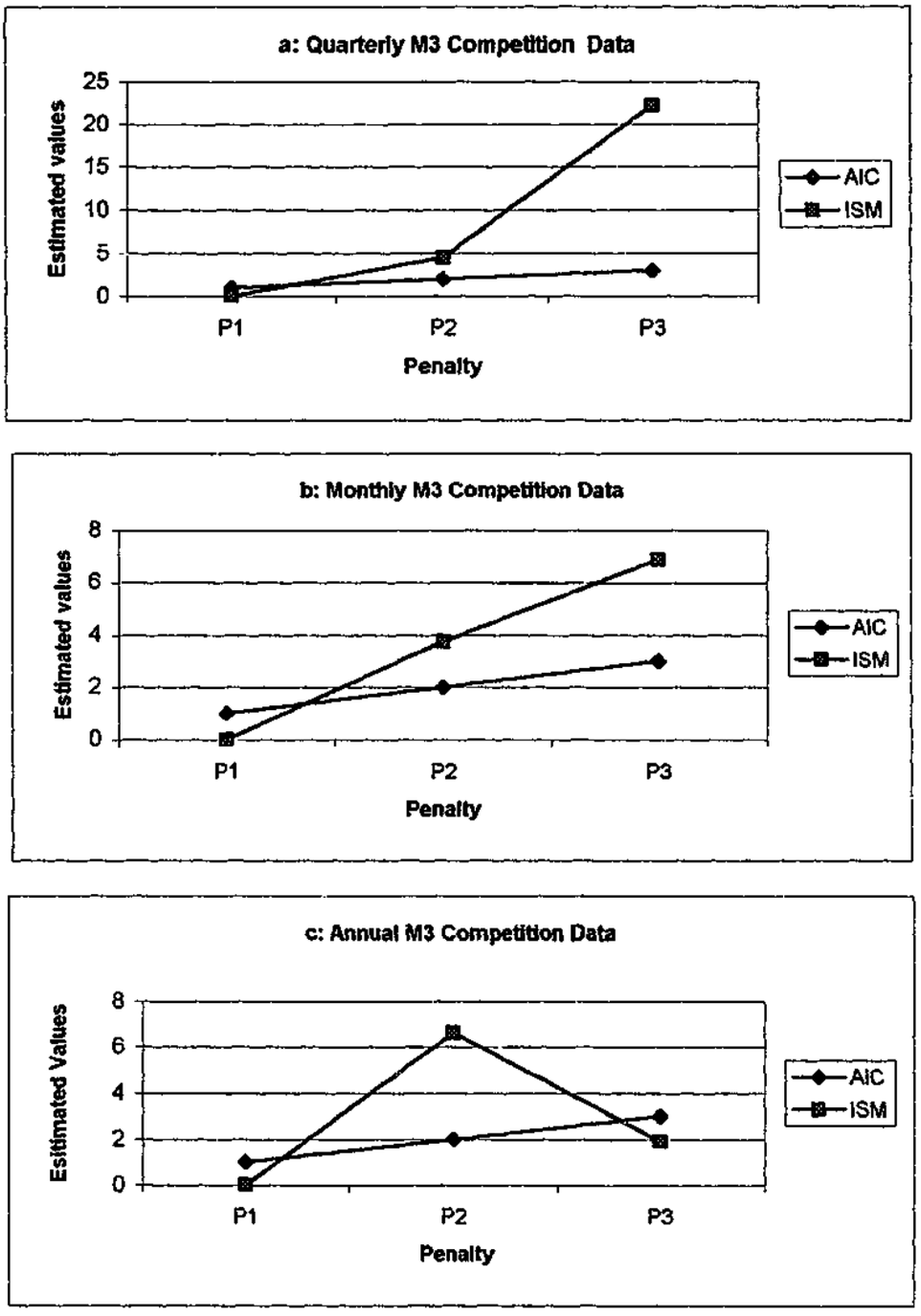

Figure 6.4: Comparison of the estimated (by ISM) and AIC penalty values for different models for the M3 competition data.

Chapter 7

Conclusions

This thesis has investigated several important issues on IC based small sample model selection for exponential smoothing models, as well as regression models with ARMA error processes. The first aim was to introduce CL based IC procedures for selecting between exponential smoothing models and also to propose ICL (based on MGL) based IC procedures, which to the best of our knowledge, have not been considered before. The second aim was to develop a new model selection procedure called PEM (PEM-GS and PEM-SA), which estimates the optimal penalty values so that the OAPCS is maximized for the particular sample size and the set of models under consideration. We also compared the performances of PL and MGL based IC procedures. Another aim was to extend the PEM based model selection procedures to a new loss function reflecting forecast accuracy rather than average probabilities of correct selection. Finally, this thesis attempted to develop an ISM for selecting between exponential smoothing models for real life collections of time series data. The following is a detailed discussion of the findings of this thesis.

The literature survey of Chapter 2 showed that there are not many comprehensive studies that have evaluated the relatively small sample performance of various existing IC model selection procedures. The majority of the research in this area has been related to asymptotic properties and Monte Carlo studies have mainly

been used to illustrate the asymptotic results. This led to the identification of the need to develop a small sample model selection procedure. Further, Chapter 2 included a survey of various exponential smoothing methods and their corresponding state space models, and promoted the application of IC based model selection procedures in the context of exponential smoothing models.

Chapter 3 showed that the development of CL methods to exponential smoothing models allows the application of IC procedures to model selection for these models. This led to a study of selection probabilities of exponential smoothing models for distinguishing the best form of penalty function among some popular and widely used existing IC procedures. Some preliminary simulations showed that the selection performances of the existing IC procedures are determined by the choice of exponential smoothing parameters. Therefore, we introduced APCS by generating exponential smoothing parameters from a weighting distribution rather similar to a prior distribution. Then, OAPCS was used as a measure for ranking the relative performances of the existing IC procedures. We found that of the existing IC procedures, BIC performs best, followed by HQ and MCp is the worst. Further, we proposed the ICL method, and compared the performances of CL and ICL based IC procedures. Our results showed that the ICL based IC procedures perform better than those based on CL.

In Chapter 4, a model selection problem for the linear regression model with different ARMA error processes was considered. We proposed new model selection procedures, namely, PEM-GS and PEM-SA, which estimate optimal penalties so that OAPCS is maximized. We found that the performances of different existing IC procedures depend on a number of factors such as sample size and design matrix, and no one procedure performs best. The new approaches (PEM-GS and PEM-SA) perform better than the existing IC procedures used in this chapter. However, in

some cases, PEM-GS was found to perform slightly better than PEM-SA, but with a cost of high computational time. We also compared the performances of PL and MGL based model selection procedures. The simulation results demonstrated that the MGL based IC procedures perform better than those based on PL.

In Chapter 5, we introduced model selection using forecast MSE as the selection loss function rather than average probabilities of correct selection. In this chapter, we developed a small sample theory for OAMSE, and extended the new model selection procedures (PEM-GS and PEM-SA) proposed in Chapter 4 to the new loss function. These procedures estimate the penalty values so that OAMSE is minimized. The results of this study showed that for longer forecast horizons, in general, BIC performs better than other existing IC procedures in selecting the appropriate forecast model. Typically, AIC was found to perform moderately well for one-step ahead forecasts. Indeed, the forecasting performances of various existing IC procedures are affected by sample size, design matrix and forecast horizon. On the other hand, irrespective of sample size, design matrix and forecasting horizon, the proposed PEM-GS and PEM-SA consistently perform better than the existing IC procedures used in this chapter, particularly for small sample sizes and larger forecast horizons. In general, PEM-GS performs slightly better than PEM-SA, but the former takes a much longer computational time compared to PEM-SA.

A new ISM was proposed in Chapter 6 for selecting exponential smoothing models for forecasting real life time series data. We applied the existing IC procedures and ISM for selecting models for the M3 competition data of Makridakis and Hibon (2000). The time series were divided into three groups, namely, annual, quarterly and monthly, and ISM was applied to each group separately. The performances of ISM, the COM method and various existing IC procedures were compared by using MAPE as the forecast accuracy measure. For quarterly and monthly data,

the existing IC procedures select the Holt-Winters additive seasonal model most often, followed by Holt's trend model, and the simple exponential smoothing model was selected the least often. However, compared to the existing IC procedures, ISM selects the simple exponential smoothing model more often. In terms of MAPE, among the existing IC procedures, typically, BIC performs best. The COM method performs better than the existing IC procedures for forecast horizons $h > 5$ (for quarterly data) and $h > 2$ (for monthly data). Irrespective of forecast horizon, in general, ISM performs better than the existing IC procedures and the COM method. For annual data, ISM selects the damped trend and simple exponential smoothing model more often than the existing IC procedures. In terms of MAPE among the existing IC procedures, MCp performs best, and the COM method performs better than MCp for forecast horizons $h > 3$. ISM performs better than the COM method as well as the existing IC procedures. We recommend the application of ISM in situations where forecasts are required on a routine basis for a large number of similar time series.

We summarize the findings of this thesis as follows. Firstly, our results showed that the IC method can be successfully applied to selecting exponential smoothing models from a group of competing models, and ICL based IC procedures perform better than those based on CL. Secondly, in small samples, with respect to OAPCS, the proposed PEM-GS and PEM-SA based model selection methods consistently perform better than the existing IC procedures, and the MGL based procedures typically perform better than those based on PL. Thirdly, PEM-GS and PEM-SA were extended for selecting models on the basis of the model's forecasting accuracy. The results showed that these new methods work better than those existing IC procedures used in our study. Finally, with respect to MAPE, the proposed ISM performs better than the existing IC procedures as well as the COM method

when applied to the M3 competition data. Overall, the model selection techniques proposed in this thesis were found to be better than the currently available IC methods. Therefore, we recommend the use of these new procedures (particularly PEM-SA and ISM) in conjunction with maximized MGL when faced with choosing between different exponential smoothing models as well as regression model with various ARMA disturbance processes.

Our results also raise several interesting questions which invite future research on model selection problems. Some of these are summarized as follows:

- Our proposed small sample model selection procedures can be applied to a number of other model selection problems which have not been considered in this thesis such as heteroscedasticity and error component regression error models.

- In Chapter 2, we discussed 24 versions of different kinds of exponential smoothing models (additive and multiplicative error models). Selecting models from this group of 24 competing models is an option worth exploring. This can be done for generated as well as for the M, M2 and M3 competition data.

- The M, M2 and M3 competition data can be deseasonalized, and the various IC procedures outlined in this thesis, can be applied to the problem of selecting between the simple exponential smoothing model, Holt's linear trend model and the damped trend model. The forecast MAPEs obtained can be compared with those from the COM method, robust trend method and theta method. Note that the study of Makridakis and Hibon (2000) showed that the robust trend and theta methods perform very well when compared to exponential smoothing methods.

- The proposed ISM can be extended to estimating a separate penalty set for

each forecast horizon. This extended method may give better forecasts compared to its present form, but it requires a higher level of computational effort, and the extended ISM cannot be applied to series with relatively small numbers of observations.

- The robustness properties of our proposed PEM can be investigated in the face of departures from normality for non-normal error distributions. The errors can be generated using Ramberg and Schmeiser's (1972) algorithm. More details about simulating non-normal error distributions can be found in Ramberg et al. (1979) and Lye and Martin (1993).

- Model selection can be based on the accuracy of estimation of a key parameter in a collection of models. For example, consider the following regression model with ARMA disturbance processes:

$$y = X\beta + Z\gamma + u, \tag{7.0.1}$$

where β is the parameter of particular interest, γ is a $k \times 1$ vector of nuisance parameters, X is a regressor vector, Z is a regressor matrix and u is an ARMA(p^*, q^*) process. Then, the loss function MSE$(\widehat{\beta})$ can be used as the criteria for determining the parameters for selecting between different models with different error processes in the context of the above regression model.

- Further, models can also be selected through the hypothesis testing approach, where the power of the test on β in (7.0.1) is used as the selection criteria. This approach may not work for all model selection problems, however it is worthy of further investigation.

References

Abraham, B. and Ledolter, J. (1986). Forecast functions implied by autoregressive integrated moving average models and other related forecast procedures, *International Statistical Review*, 54, 51-66.

Adam, E.E. and Ebert, R.J. (1976). A comparison of human and statistical forecasting, *AIITE Transactions*, 8, 120-127.

Akaike, H. (1970). Statistical predictor identification, *Annals of Institute of Statistical Mathematics*, 22, 203-217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in B.N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*, Akademiai Kiado: Budapest, 267-281.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control,* AC-19, 716-723.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting, *Biometrika*, 66, 237-242.

Akaike, H. (1981). Likelihood of a model and information criteria, *Journal of Econometrics,* 16, 3-14.

Ali, M., Torn, A. and Viitanen, S. (1997). A direct search simulated annealing algorithm for optimization involving continuous variables, TUCS Technical Report No. 97, Turku Centre for Computer Science, Finland.

Allen, D.M. (1971). The prediction sum of squares as a criterion for selecting predictor variables, Technical Report No. 23, Department of Statistics, University of Kentucky, Lexington, KY.

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, 16, 125-127.

Amemiya, T. (1972). Lecture notes on econometrics, mimeo, Department of Economics, Stanford University, Stanford, CA.

Amemiya, T. (1980). Selection of regressors, *International Economic Review*, 21, 331-354.

Aptech Systems, Inc. (1996). *GAUSS Constrained Optimization*, Maple Valley, WA.

Ara, I. (1995). Marginal likelihood based tests of regression disturbances, unpublished Ph.D. thesis, Department of Econometrics and Business Statistics, Monash University, Australia.

Ara, I. and King, M.L. (1993). Marginal likelihood based tests of regression disturbances, mimeo, Monash University, Department of Econometrics and Business Statistics.

Archibald, B.C. (1990). Parameter space of the Holt-Winters' model, *International Journal of Forecasting*, 6, 199-209.

Archibald, B.C. (1994). Winters model: three versions, diagnostic checks and forecast performances, Working Paper WP-94-4, School of Business Adminstration, Dalhouse University, Halifax, Canada.

Armstrong, J.S. and Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons, *International Journal of Forecasting*, 8, 69-80.

Armstrong, J.S. and Collopy, F. (1993). Causal forces: structuring knowledge for time series extrapolation, *Journal of Forecasting*, 12, 103-115.

Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting, *International Journal of Forecasting*, 16, 521-530.

Atkinson, A.C. (1980). A note on the generalized information criterion for choice of a model, *Biometrika*, 67, 413-418.

Atkinson, A.C. (1981). Likelihood ratios, posterior odds and information criteria, *Journal of Econometrics*, 16, 15-20.

Azam, M.N. and King, M.L. (1998). Model selection with data driven penalty when there is an unknown changepoint in the data, in C.L. Skeels (ed.), *Proceedings of the Econometric Society Australasian Meeting*, CD Rom, The Australian National University, Canberra.

Bartolomei, S.M. and Sweet, A.L. (1989). A note on a comparison of exponential smoothing methods for forecasting seasonal series, *International Journal of Forecasting*, 5, 111-116.

Bethel, J.W. (1984). Asymptotic properties of information-theoretic methods for model selection, unpublished Ph.D. thesis, University of California, Davis, CA.

Billah, M.B. and King, M.L. (1998a). Model selection and optimal penalty estimation for time series models, in C.L. Skeels (ed.), *Proceedings of Econometric Society Australasian Meeting*, CD Rom, The Australian National University, Canberra.

Billah, M.B. and King, M.L. (1998b). Model selection for time series forecasting, *Proceedings of the Fourth Annual Conference*, Faculty of Business and Economics, Monash University, Clayton, Victoria, Australia.

Billah, M.B. and King, M.L. (2000a). Time series model selection and forecasting via optimal penalty estimation, *Pakistan Journal of Statistics, Saleh-Aly Special Edition*, 16, 126-145.

Billah, M.B. and King, M.L. (2000b). Using simulated annealing to estimate penalty functions for time series model selection, *Journal of Statistical Research*, 34, 75-89.

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

Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd edition, Holden Day: San Francisco.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd edition, Englewood Cliffs: NJ.

Brown, R.G. (1959). *Statistical Forecasting for Inventory Control*, McGraw Hill: New York.

Brown, R.G. (1963). *Smoothing, Forecasting and Prediction of Discrete Time Series*, Prentice-Hall: Englewood Cliffs.

Brown, R.G. (1967). *Decision Rules for Inventory Management*, Holt, Rinehart and Winston: New York.

Carnevali, P., Coletti, L. and Patarnello, S. (1985). Image processing by simulated annealing, *IBM Journal of Research and Development*, 29, 569-579.

Chatfield, C. (1978). The Holt-Winters forecasting procedure, *Applied Statistics*, 27, 264-279.

Chatfield, C. (1989). *The Analysis of Time Series*, 4th edition, Chapman and Hall: London.

Chatfield, C. and Yar, M. (1988). Holt-Winters forecasting: theory and practice, *The Statistician*, 37, 129-140.

Chatfield, C. and Yar, M. (1991). Prediction intervals for multiplicative Holt-Winters, *International Journal of Forecasting*, 7, 31-37.

Chen, C. (1993). Some robustness properties of the simple exponential smoothing predictor, *Journal of Japan Statistical Society*, 23, 201-214.

Chen, C. (1996). Some statistical properties of the Holt-Winters seasonal forecasting method, *Journal of Japan Statistical Society*, 26, 173-187.

Chen, C. (1997). Robustness properties of some forecasting methods for seasonal time series: a Monte Carlo study, *International Journal of Forecasting*, 13, 269-280.

Chen, C., Davis, R.A., Brockwell, P.J. and Bai, Z.D. (1993). Order determination for resampling methods, *Statistica Sinica*, 3, 481-500.

Clayton, M.K., Geisser, S. and Jennings, D.E. (1986). A comparison of several model selection procedures, in P. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques*, Elsevier Science: Amsterdam, 425-439.

Clemen, R. (1989). Combining forecasts: a review and annotated bibliography with discussion, *International Journal of Forecasting*, 5, 559-608.

Cogger, K.O. (1973). Time series analysis and forecasting with an absolute error criterion, *TIMS Studies in Management Science*, 12, 189-201.

Collopy, F. and Armstrong, J.S. (1992). Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations, *Management Science*, 38, 1394-1414.

Corana, A., Marchesi, M., Martini, C. and Ridella, S. (1987). Minimizing multimodal functions of continuous variables with the simulated annealing algorithm, *ACM Transactions of Mathematical Software*, 13, 262-280.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate autoregression, *Journal of the Royal Statistical Society, Series B*, 49, 1-39.

Cox, D.R. and Reid, N. (1993). A note on the calculation of adjusted profile likelihood, *Journal of the Royal Statistical Society, Series B*, 55, 467-471.

Cramer, J.S. (1986). *Econometric Applications of Maximum Likelihood Methods*, Cambridge University Press, New York.

Crato, N. and Ray, B.K. (1996), Model selection and forecasting for long-range dependent process, *Journal of Forecasting*, 15, 107-125.

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numerical Mathematics*, 13, 377-403.

DeGroot, M.H. (1970). *Optimal Statistical Decisions*, McGraw-Hill: New York.

Derwent, D. (1988). A better way to control population, *Nature*, 331, 575-578.

Duncan, D.B. and Horn, S.D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis, *Journal of the American Statistical Association*, 67, 815-821.

Durbin, J. and Watson, G.S. (1951). Testing for serial correlation in least squares regression II, *Biometrika*, 38, 159-178.

Engle, R.F. and Brown, S.J. (1986). Model selection for forecasting, *Applied Mathematics and Computation*, 20, 313-327.

Fildes, R. (1989). Evaluation of aggregate and individual forecast method selection rules, *Management Science*, 35, 1056-1065.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods (with discussion), *International Journal of Forecasting*, 8, 81-111.

Fildes, R., Hibon, M., Makridakis, S. and Meade, N. (1998). Generalizing about univariate forecasting methods: further empirical evidence, *International Journal of Forecasting*, 14, 339-358.

Fildes, R. and Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting, *International Statistical Review*, 63, 289-308.

Finch, S.J., Mendell, N.R. and Thode, H.C. (1989). Probabilistic measures of accuracy of a numerical search for a global maximum, *Journal of the American Statistical Association*, 84, 1020-1023.

Forbes, C.S., King, M.L. and Morgan, A. (1995). A small sample variable selection procedure, in C.S. Forbes, P. Kofman and T.R.L. Fry (eds.), *Proceedings of the 1995 Econometric Conference at Monash*, Monash University, Australia, 343-360.

Fox, K.J. (1995). Model selection criteria: a reference source, unpublished manuscript, Department of Economics, University of British Columbia and School of Economics, University of New South Wales.

Gardner, E.S. (1983). Automatic monitoring of forecast errors, *Journal of Forecasting*, 2, 1-21.

Gardner, E.S. (1984). Forecast control with parabolic masks, working paper, Navy Fleet Material Support Office, Mechanicsburg, PA 17055, U.S.A.

Gardner, E.S. (1985). Exponential smoothing: the state of the art, *Journal of Forecasting*, 4, 1-28.

Gardner, E.S. and McKenzie, E. (1985). Forecasting trends in time series, *Management Science*, 31, 1237-1246.

Gardner, E.S. and McKenzie, E. (1988). Model identification in exponential smoothing, *Journal of the Operational Society*, 39, 863-867.

Gardner, E.S. and McKenzie, E. (1989). Seasonal exponential smoothing with damped trends, *Management Science*, 35, 372-376.

Gauss, C.F. (1821, collected wcrks 1873). Theoria combinationis observationum erroribus minimis obnoxiae, Werke, 4, Gottingen.

Geisser, S. (1974). A predictive approach to the random effect model, *Biometrika*, 61, 101-107.

Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association*, 70, 320-328.

Geurts, M.D. and Kelly, J.P. (1986). Forecasting demand for special services, *International Journal of Forecasting*, 2, 261-272.

Geweke, J. and Meese, R. (1981). Estimating regression models of finite but unknown order, *International Economic Review*, 22, 55-70.

Gilchrist, W.G. (1976). *Statistical Forecasting*, John Wiley and Sons: London.

Goffe, W.L., Ferrier, G.D. and Rogers, J. (1994). Global optimization of statistical functions with simulated annealing, *Journal of Econometrics*, 60, 65-99.

Golub, G.H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, 21, 215-223.

Gorman, J.W. and Toman, R.J. (1966). Selection of variables for fitting equations to data, *Technometrics*, 8, 27-51.

Granger, C.J.W., King, M.L. and White, H. (1995). Comments on testing economic theories and the use of model selection criteria, *Journal of Econometrics*, 67, 173-187.

Granger, C.J.W. and Newbold, P. (1986). *Forecasting Economic Time Series*, 2nd edition, Academic Press: New York.

Groff, G.K. (1973). Empirical comparison of models for short-range forecasting, *Management Science*, 20, 22-31.

Grose, S.D. and King, M.L. (1994). The use of information criteria for model selection between models with equal numbers of parameters, mimeo, Monash University, Depar ..nt of Econometrics and Business Statistics.

Grunwald, G.K. and Hyndman, R.J. (1998). Smoothing non-Gaussian time series with autoregressive structure, *Computational Statistics and Data Analysis*, 28, 171-191.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley and Sons: New York.

Hannan, E.J. (1970). *Multiple Time Series*, John Wiley and Sons: New York.

Hannan, E.J. (1980). The estimation of the order of an ARMA process, *Annals of Statistics*, 8, 1071-1081.

Hannan, E.J. (1981). Estimating the dimension of a linear system, *Journal of Multivariate Analysis*, 11, 211-221.

Hannan, E.J. (1982). Testing for autocorrelation and Akaike's criterion, in J. Gani and E.J. Hannan (eds.), *Essays in Statistical Science*, Paper in honour of P.A.P. Moran, Applied Probability Trust, Sheffield, 403-412.

Hannan, E.J. and Kavalieris, L. (1984). A method for autoregressive-moving average estimation, *Biometrika*, 71, 273-280.

Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression, *Journal of the Royal Statistical Society, Series B*, 41, 190-195.

Hannan, E.J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order, *Biometrika*, 69, 81-94.

Harrison, P.J. (1967). Exponential smoothing and short term sales forecasting, *Management Science*, 13, 821-842.

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

Harvey, A.C. (1991). *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Harvey, A.C. and Phillips, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive moving average disturbances, *Biometrika*, 66, 49-58.

Heyde, C.C. (1974). An iterated logarithm result for autocorrelations of a stationary linear process, *Annals of Probability*, 2, 328-332.

Heyde, C.C. and Scott, D.J. (1973). Invariance principles for the laws of the iterated logarithm for martingales and processes with stationary increments, *Annals of Probability*, 1, 428-436.

Hillmer, S.C. (1985). Comments on 'Exponential smoothing: the state of the art' by E.S. Gardner, *Journal of Forecasting*, 4, 31-32.

Holmes, J.M. and Hutton, P.A. (1989). Optimal model selection when the true relationship is weak and occurs with a delay, *Economics Letters*, 30, 333-339.

Holt, C.C. (1957). *Planning Production, Inventories and Work Force*, Prentice-Hall: Englewood Cliffs, NJ.

Holt, C.C., Modigliani, F., Muth, J.F. and Simon, H.A. (1960). *Planning, Production Inventories and Work Force*, Prentice-Hall: Englewood Cliffs, NJ.

Hogarth, R.M. (1975). Cognitive processes and the assessment of subjective probability distributions, *Journal of the American Statistical Association*, 70, 271-290.

Hossain, Z. (1998). Model selection problems involving interval restricted parameters in econometric, unpublished Ph.D. thesis, Department of Econometrics and Business Statistics, Monash University, Australia.

Hossain, Z. and King, M.L. (1998). Model selection when a key parameter is constrained to be in an interval, Working Paper 15/98, Department of Econometrics and Business Statistics, Monash University, Australia.

Hughes, A.W. (1997). Improved model selection based on AIC type criteria, unpublished Ph.D. thesis, Department of Econometrics and Business Statistics, Monash University, Australia.

Hurvich, C.M. and Tsai, C.L. (1989). Regression and time series model selection in small samples, *Biometrika*, 76, 297-307.

Hurvich, C.M. and Tsai, C.L. (1990). The impact of model selection on inference in linear regression, *The American Statistician*, 44, 214-217.

Hurvich, C.M. and Tsai, C.L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models, *Biometrika*, 78, 499-509.

Hurvich, C.M. and Tsai, C.L. (1993). A corrected Akaike information criterion for vector autoregressive model selection, *Journal of Time Series Analysis*, 14, 271-279.

Hyndman, R.J., Koehler, A.B., Snyder, R.D. and Grose, S. (2000). A state space framework for automatic forecasting using exponential smoothing methods, Working Paper 9/2000, Department of Econometrics and Business Statistics, Monash University, Australia.

Jeffreys, H. (1967). *Theory of Probability*, 3rd edition, Clarendon Press: Oxford.

Jenkins, G.M. and Alavi, A.S. (1981). Some aspects of modeling and forecasting multivariate time series, *Journal of Time Series Analysis*, 2, 1-47.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lütkepohl, H. and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*, 2nd edition, John Wiley and Sons: New York.

Kalbfleish, J. and Sprott, D. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion), *Journal of the Royal Statistical Society, Series B*, 32, 175-208.

Kalman, R.F. (1960). A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, ASME Transactions, 82D, 35-45.

King, M.L. (1980). Robust tests for spherical symmetry and their application to least squares regression, *The Annals of Statistics*, 8, 1265-1271.

King, M.L. and Bose, G. (2000). Finding optimal penalties for model selection in the linear regression model, mimeo, Monash University, Department of Econometrics and Business Statistics.

King, M.L., Forbes, C.S. and Morgan, A. (1995). Improved small sample model selection procedures, paper presented at the 1995 World Congress of the Econometric Society, Tokyo, Japan.

Kirby, R.M. (1966). A comparison of short and medium range statistical forecasting methods, *Management Science*, 15, 202-210.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing, *Science*, 220, 671-680.

Koehler, A.B. and Murphree, E.S. (1988), A comparision of the Akaike and Schwarz criteria for selecting model order, *Applied Statistics*, 37, 187-195.

Koehler, A.B., Snyder, R.D. and Ord, J.K. (1999). Forecasting models and prediction intervals for the multiplicative Holt-Winters method, Working Paper 1/1999, Department of Econometrics and Business Statistics, Monash University, Australia.

Kohn, R. (1983). Consistent estimation of minimal subset dimension, *Econometrica*, 51, 367-376.

Koreisha, S.G. and Pukkila, T. (1995), A comparison between different order-determination criteria for identification of ARIMA models, *Journal of Business and Economic Statistics*, 13, 217-231.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, 22, 79-86.

Kwek, K.T. (2000). Model selection for a class of conditional heteroscedastic processes, unpublished Ph.D. thesis, Department of Econometrics and Business Statistics, Monash University, Australia.

Kwek, K.T. and King, M.L. (1997a). Model selection and optimal penalty estimation for conditional heteroscedastic processes, in P. Bardsley and V.L. Martin (eds.), *Proceedings of the Econometric Society Australasian Meeting*, University of Melbourne, Melbourne, Victoria, 143-170.

Kwek, K.T. and King, M.L. (1997b). Model selection of ARCH processes using iterated optimal penalties, *Proceedings of the Third Annual Doctoral Research Conference*, Faculty of Business and Economics, Monash University, Clayton, Victoria, 51-62.

Kwek, K.T. and King, M.L. (1998). Information criteria in conditional heteroscedastic models: a Bayesian prior approach to penalty function building, in C.L. Skeels (ed.), *Proceedings of the Econometric Society Australasian Meeting*, CD Rom, The Australian National University, Canberra.

Larimore, W.E. and Mehra, R.K. (1985). The problem of overfitting data, *Byte*, 10, 167-180.

Laskar, M.R. and King, M.L. (1998). Estimation and testing of regression disturbances based on modified likelihood functions, *Journal of Statistical Planning and Inference*, 71, 75-92.

Lawrence, M.T., Edmundson, R.H. and O'Connor, M.J. (1985). An examination of the accuracy of judgemental extrapolation of time series, *International Journal of Forecasting*, 1, 14-24.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Non-Experimental Data*, Wiley and Sons: New York.

Lehmann, E.L. (1991). *Testing Statistical Hypotheses*, 2nd edition, John Wiley and Sons: New York.

Lewandowski, R. (1979). *La Prèvision à Court Terme*, Dunod: Paris.

Ljung, G.M. and Box, G.E.P. (1979). The likelihood function of stationary autoregressive moving average models, *Biometrika*, 66, 265-270.

Lutkepohl, H. (1984). Forecasting contemporaneously aggregated vector ARMA processes, *Journal of Business and Economic Statistics*, 2, 201-214.

Lütkepohl, H. (1985). Comparision of criteria for estimating the order of a vector autoregressive process, *Journal of Time Series Analysis*, 6, 35-52.

Lye, J.N. and Martin, V.L. (1993). Robust estimation, non-normalities and generalized exponential distributions, *Journal of the American Statistical Association*, 88, 261-267.

Mabert, V.A. (1975). Statistical versus sales force-executive opinion short-range forecasts: a time series analysis case study, Krannert Graduate School, Purdue University.

Makridakis, S. (1986). The art and science of forecasting: an assessment and future directions, *International Journal of Forecasting*, 2, 15-39.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition, *Journal of Forecasting*, 1, 111-153.

Makridakis, S., Chatfield, C., Hibon, M., Mills, T., Ord, J.K. and Simmons, L.F. (1993). The M2-Competition: a real-time judgmentally based forecasting study, *International Journal of Forecasting*, 9, 5-22.

Makridakis, S. and Hibon, M. (1979). Accuracy of forecasting: an empirical investigation (with discussion), *Journal of the Royal Statistical Society, Series A*, 142, 97-145.

Makridakis, S. and Hibon, M. (1991). Exponential smoothing: the effect of initial values and loss functions on post-sample forecasting accuracy, *International Journal of Forecasting*, 7, 317-330.

Makridakis, S. and Hibon, M. (2000). The M3-Competition: results, conclusions and implications, *International Journal of Forecasting*, 16, 451-476.

Makridakis, S. and Wheelwright, S.C. (1978). *Interactive Forecasting: Univariate and Multivariate Methods*, 2nd edition, Holden-Day: San Francisco.

Makridakis, S. and Wheelwright, S.C. (1989). *Forecasting Methods for Management*, 5th edition, John Wiley and Sons: New York.

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). *Forecasting Methods and Applications*, John Wiley and Sons: New York.

Mallows, C.L. (1964). Choosing variables in a linear regression: a graphical aid, presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

Mallows, C.L. (1973). Some comments on Cp, *Technometrics*, 15, 661-676.

McClain, J.O. (1974). Dynamics of exponential smoothing with trend and seasonal terms, *Management Science*, 20, 1300-1304.

McClain, J.O. (1981). Restarting a forecasting system when demand suddenly changes, *Journal of Operations Management*, 2, 53-61.

McClain, J.O. and Thomas, L.J. (1973). Response-variance tradeoffs in adaptive forecasting, *Operations Research*, 21, 554-568.

McKenzie, E. (1976). A comparison of some standard seasonal forecasting systems, *The Statistician*, 25, 3-14.

McKenzie, E. (1984). General exponential smoothing and the equivalent ARMA process, *Journal of Forecasting*, 3, 333-334.

McKenzie, E. (1985). Comments on 'Exponential smoothing: the state of the art' by E.S. Gardner, *Journal of Forecasting*, 4, 32-36.

McKenzie, E. (1986). Renormalization of seasonals in Winters' forecasting systems: is it necessary? *Operations Research*, 34, 174-176.

McNees, S.K. (1976). An evaluation of economic forecasting, *New England Economic Reviews*, November-December issue.

Meade, N. (2000). A note on the robust trend and ARARMA methodologies used in the M3 Competition, *International Journal of Forecasting*, 16, 517-519.

Meese, R. and Geweke, J. (1984). A comparison of autoregressive univariate forecasting procedures for macroeconomic time series, *Journal of Business and Economic Statistics*, 2, 191-200.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, 21, 1087-1091.

Mills, J.A. and Prasad, K. (1992). A comparision of model selection criteria, *Econometric Reviews*, 11, 201-233.

Montgomery, D.C. and Johnson, L.A. (1976). *Forecasting and Time Series Analysis*, McGraw-Hill: New York.

Muth, J.F. (1960). Optimal properties of exponentially weighted forecasts, *Journal of the American Statistical Association*, 55, 299-306.

Newbold, P. and Granger, C.W.J. (1974). Experience with forecasting univariate time series and the combination of forecasts, *Journal of the Royal Statistical Society, Series A*, 137, 131-165.

Nishii, R. (1986). Criteria for selection of response variables and the asymptotic properties in a multivariate calibration, *Annals of Institute of Statistical Mathematics*, 38, 319-329.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified, *Journal of Multivariate Analysis*, 27, 392-403.

Ord, J.K., Koehler, A.B. and Snyder, R.D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models, *Journal of the American Statistical Association*, 92, 1621-1629.

Parzen, E. (1974). Some recent advances in time series modeling, *Transactions on Automatic Control*, AC-19, 723-730.

Parzen, E. (1979). Time series and whitening filter estimation, *TIMS Studies in Management Science*, 12, 149-165.

Pesaran, M.H. (1974). On the general problem of model selection, *Review of Economic Studies*, 41, 153-171.

Pegels, C. (1969). Exponential forecasting: some new variations, *Management Science*, 15, 311-315.

Potscher, B.M. (1991). Effects of model selection on inference, *Econometric Theory*, 7, 163-185.

Press, W.H., Flannery, B., Teukolsky, S.A. and Vetterling, W.T. (1986). *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press: New York.

Pronzato, L., Walter, E., Venot, A. and Lebruchec, J.-F. (1984). A general-purpose optimizer: implementation and applications, *Mathematics and Computers in Simulation*, 24, 412-422.

Quinn, B.G. (1980). Order determination for a multivariate autoregression, *Journal of the Royal Statistical Society, Series B*, 42, 182-185.

Rahman, M.S., Bose, G.K. and King, M.L. (1998). Improved penalty functions for information criteria based model selection, in C.L. Skeels (ed.), *Proceedings of the Econometric Society Australasian Meeting*, University of Melbourne, Melbourne, Victoria, 613-640.

Rahman, M.S. and King, M.L. (1997). Marginal likelihood based score tests of regression disturbances in the presence of nuisance parameters, *Journal of Econometrics*, 82, 63-80.

Ramberg, J.S. and Schmeiser, B.W. (1972). An approximate method for generating symmetric random variables, *Communications of the Association for Computing Machinery*, 15, 987-990.

Ramberg, J.S., Tadikamalla, P.R., Dudewicz, E.J. and Mykytka, E.F. (1979). A probability distribution and its uses in fitting data, *Technometrics*, 21, 201-215.

Ray, B.K. (1993). Long-range forecasting of IBM product revenues using a seasonal fractionally differenced ARMA Model, *International Journal of Forecasting*, 9, 255-269.

Ray, B.K. and Crato, N. (1994). Model selection and forecasting of long-range dependent processs: results of a simulation study, Center for Applied Mathematics and Statistics, Working Paper 50, Fall 1994, New Jersey Institute of Technology.

Reid, D.J. (1969). A comparative study of time series prediction techniques on economic data, Ph.D. thesis, University of Nottingham.

Reid, D.J. (1977). A survey of statistical forecasting techniques with empirical comparisons, paper presented at the *IEEE Symposium on Statistical Model Building for Prediction and Control.*

Rissanen, J. (1978). Modeling by shortest data description, *Automatica*, 14, 465-471.

Rissanen, J. (1987). Stochastic complexity and modeling, *Journal of the Royal Statistical Society, Series B*, 49, 223-239 and 252-265 (with discussions).

Rissanen, J. (1988). Stochastic complexity and the MDL principle, *Econometric Reviews*, 85-102.

Roberts, S.A. (1982). A general class of Holt-Winters type forecasting models, *Management Science*, 28, 808-820.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models, *Econometrica*, 46, 1273-1291.

Schmidt, P. (1971). Methods of choosing alternative models, Michigan State University Econometrics Workshop Paper, No. 7004.

Schmidt, P. (1974). Choosing among alternative linear regression models, *Atlantic Economic Journal*, 2, 7-13.

Schmidt, P. (1975). Choosing among alternative linear regression models: a correction and some further results, *Atlantic Economic Journal*, 3, 61-63.

Schmidt, P. and Tschernig, R. (1993). Identification of fractional ARIMA models in the presence of long memory, Discussion Paper, University of Munich, Munich, Germany.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464.

Schwepe, F. (1965). Evaluation of likelihood functions for Gaussian signals, *IEEE Transactions on Information Theory*, 11, 61-70.

Shami, R. (1997). Exponential smoothing of seasonal time series without seasonal smoothing constant, unpublished Masters thesis, Department of Econometrics and Business Statistics, Monash University, Australia.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Annals of Statistics*, 8, 147-164.

Shibata, R. (1986). Regression variables: selection of, in *Encyclopedia of Statistical Science*, 7, John Wiley and Sons: New York, 709-714.

Silvapulle, P. (1991). Non-nested testing of regression disturbances, unpublished Ph.D. thesis, Department of Econometrics and Business Statistics, Monash University, Australia.

Sneek, J.M. (1984). *Modeling Procedures for Univariate Time Series*, Free University Press: Amsterdam.

Snyder, R.D. (1985a). Recursive estimation of dynamic linear statistical models, *Journal of the Royal Statistical Society, Series B*, 47, 272-276.

Snyder, R.D. (1985b). Further development in the estimation of dynamic linear statistical models, Working Paper No. 10/85, Department of Econometrics and Business Statistics, Monash University, Australia.

Snyder, R.D. (1985c). Estimation of a dynamic linear model, Working Paper No. 15/85, Department of Econometrics and Business Statistics, Monash University, Australia Australia.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictors (with discussion), *Journal of the Royal Statistical Society, Series B*, 36, 111-147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B*, 39, 44-47.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz, *Journal of the Royal Statistical Society, Series B*, 41, 276-278.

Stone, M. (1981). Admissible selection of an accurate and parsimonious normal linear regression model, *Annals of Statistics*, 9, 475-485.

Sweet, A.L. (1983a). Computing the variance of the forecast error for the additive seasonal model, *Research Memo No. 83-5*, School of Industrial Engineering,

Purdue University, West Lafayette, IN.

Sweet, A.L. (1983b). Computing the variance of the forecast error for the multiplicative (Holt-Winters) seasonal model, *Research Memo No. 83-7*, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Szu, H. and Hartley, R. (1987). Fast simulated annealing, *Physics Letters, A*, 3, 157-162.

Taylor, S.G. (1981). Initialization of exponential smoothing forecasts, *Transactions*, 13, 199-205.

Telly, H., Liebling, Th.M. and Mocellin, A. (1987). Reconstruction of polycrystalline structures: a new application of combinatorial optimization, *Computing*, 38, 1-11.

Theil, H. (1961). *Economic Forecast and Policy*, 2nd edition, North Holland Publishing Company: Amsterdam.

Theil, H. (1971). *Principles of Econometrics*, John Wiley and Sons: London.

Tunnicliffe Wilson, G. (1989). On the use of marginal likelihood in time series model estimation, *Journal of the Royal Statistical Society, Series B*, 51, 15-27.

Van der Leeuw, J. (1994). The covariance matrix of ARMA errors in closed form, *Journal of Econometrics*, 63, 397-405.

Verbyla, A.P. (1990). A conditional derivation of residual maximum likelihood, *Australian Journal of Statistics*, 32, 227-230.

Vetterling, W.T., Press, W.H., Teukolsky, S.A. and Flannery, B.P. (1994). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, Cambridge University: New York.

Wade, R.C. (1967). A technique for initializing exponential smoothing forecasts, *Management Science*, 13, 601-602.

Wasserman, P.D. and Schwartz, T. (1988). Neural networks, part 2: what are they and why is everybody so interested in them now? *IEEE Expert*, Spring, 10-15.

Wei, C.Z. (1992). On predictive least squares principles, *Annals of Statistics*, 20, 1-42.

Winters, P.R. (1960). Forecasting sales by exponentially weighted moving averages, *Management Science*, 6, 324-342.

Wong, D.F., Leong, H.W. and Liu, C.L. (1988). *Simulated Annealing for VLSI Design*, Kluwer Academic Publishers: Boston, MA.

Wu, P.X. (1991). One-sided and partially one-sided multiparameter hypothesis testing in econometrics, unpublished Ph.D. thesis, Department of Econometrics and Business Statistics, Monash University, Australia.

Yar, M. and Chatfield, C. (1990). Prediction intervals for the Holt-Winters forecasting procedure, *International Journal of Forecasting*, 6, 127-137.

Young, A.S. (1982). The Bivar criterion for selecting regressors, *Technometrics*, 24, 181-189.