Order Number 9314427

Evaluation of order determination procedures in ARMA models

Parkhurst, Anne Mullins, Ph.D.

The University of Nebraska - Lincoln, 1992

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

EVALUATION OF ORDER DETERMINATION PROCEDURES IN ARMA MODELS

by

Anne M. Parkhurst

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Interdepartmental Area of Engineering

Under the Supervision of Professor John L. Ballard

Lincoln, Nebraska

December, 1992

DISSERTATION TITLE

Evaluation of Order Determination Procedures in ARMA Models

BY

Anne M. Parkhurst

SUPERVISORY COMMITTEE:

APPROVED	DATE
Signature	11/20/92
John L. Ballard Typed Name	
Signature	11/20/92
Morris H. Schneider Typed Name	
Signature	11/20/92
K. P. Rajurkar Typed Name	
Signature	11/20/92
James A. DeShazer Typed Name	
Signature	
Typed Name	
Signature	
Typed Name	

GRADUATE COLLEGE
UNIVERSITY OF NEBRASKA

EVALUATION OF ORDER DETERMINATION PROCEDURES IN ARMA MODELS

Anne M. Parkhurst, Ph.D.

University of Nebraska, 1992

Adviser: John L. Ballard

Autoregression-moving average (ARMA) models provide insight into many biological systems. One of the most difficult decisions in ARMA modeling is identifying the order of the model. Many procedures have been proposed. The purpose of this study was to compare Pandit and Wu's multi-objective test criterion with Akaike's information criterion and Schwartz's Bayesian criterion, SBC.

Several versions of the criteria were compared for 10 processes which were simulations of ARMA models ranging in order from (1,1) to (8,7). Batches of 100 realizations were generated for each model. The performance of the criteria varied depending on how closely a process complied with the assumptions, length of series and alpha levels chosen. The error variance and method of parameter estimation had no effect on the relative merits of the criteria.

The proposed strategy uses Pandit and Wu's $(2p,2p-1)$ portfolio and SBC (or F-ratio, if sample size $\leq$ 125) to identify a neighborhood for the model on the first pass. The portfolio is modified to include all models between adjacent primary models which define the neighborhood. In the second

pass, a model is selected using SBC. Three classical time series were assessed by this strategy. The results compare favorably to those in the literature.

The proposed strategy was utilized to identify a sensible heat loss model for non-laying hens. The ARMA(2,0) model which was selected for all four hens, explained 70 to 89% of the total variation. The proposed stragety was also used to identify a tympanic temperature model for steers housed in a constant thermoneutral environment. The ARMA(3,0) model was adequate for both animals in the study, although the model accounted for only 15 to 27% of the total variation.

This dissertation is divided into four sections. The first contains a literature review of time series relative to dynamic systems. The second provides a comparison of order determination procedures and introduces a new strategy. The third describes the utilization of proposed strategy in modeling biological processes. The last section discusses various aspects in the construction of simulation software.

ACKNOWLEDGEMENTS

The road to completing this study was long and winding. I could not have made it without the help and support of my professors, friends and colleagues. As I began the program, the first person to encourage me and expedite the process was Prof. Morris H. Schneider - without him there would have been no beginning.

I would like to express my sincere appreciation to my advisor, Prof. John L. Ballard whose support, counsel, and encouragement kept me going through many trials and tribulations. Without his patience there would have been no end.

Of all those who helped me through the middle, I'd especially like to thank the members of the reading committee, Kamlakar P. Rajurkar and James A. DeShazer. To Prof. Rajurkar I express gratitude for introducing me to data dependent systems, and providing valuable constructive criticism and suggestions.

To Prof. DeShazer, I am deeply indebted for the tutorial in environmental systems, for sharing his creativity and friendship, and for introducing me to other biological engineers, especially G. Leroy Hahn.

I wish to thank Dr. Hahn for sharing his enthusiasm, for his receptiveness to new ideas, for the introduction to chaos and for the many thoughtful and extremely stimulating scientific discussions.

To my loving parents - James and Josephine Mullins

and

my darling children

Lars Jimmy James Paul Mullins Parkhurst

Vera Elizabeth Josephine Anne Mullins Parkhurst

The past and the future,

May they forever be entwined

TABLE OF CONTENTS

I. BACKGROUND

5. Order Determination Procedures

6. References

II. COMPARISON OF ORDER DETERMINATION CRITERIA

1. Introduction

2. Materials and Methods

3. Results and Discussion

4. Conclusions

5. References

III. IMPLEMENTATION - Modeling Biological Processes of Farm Animals

1. Introduction

2. Dynamical Systems Approach

3. Results and Discussion

4. Conclusions

5. References

6. Appendix

IV. SIMULATION - DEVELOPMENT OF PROGRAMS

1. Introduction

2. Materials and Methods

3. Programs

4. Conclusions

5. References

6. Appendix

I. BACKGROUND

This section contains both a literature review and a review of the concepts in time series analysis that are relevant to the study of dynamic systems. Numerous journal articles are referenced as well as classical textbooks. The textbooks range from the elementary (Chatfield, 1989) to the intermediate (Box and Jenkins, 1976; Pandit and Wu, 1983) to the more advanced (Priestly, 1981).

1. Stochastic Processes

A stochastic process is a statistical phenomenon that evolves according to the laws of probability. The term *stochastic* is derived from the Greek word pertaining to chance or conjecture. Stochastic (or random) processes are useful for modeling time series. The phrase "*time series*" has become a generic term for records of events that fluctuate over time or space. Some examples of time series are: mechanical vibrations, the length of a queue, population growth, and air temperature.

Statistically, a stochastic process is a collection of random variables, ordered in time. Usually time is measured at discrete intervals, t=0,1,2,... and only one observation is available for each point in time or space. Nevertheless, the observed time series is regarded as one sample from a collection of time series. The infinite population of a

collection of time series is called the *ensemble*, while a single time series is called a *realization*. Hence, time series analysis may be regarded as a way of evaluating parameters of the probability model which generated the observed time series.

1.a. AUTOCORRELATION AND AUTOCOVARIANCE

Autocorrelation measures the association between pairs of observations. In the conventional situation, the correlation between two variables is

$$\rho = \frac{COV(X,Y)}{\sqrt{var(X)}\,\sqrt{var(Y)}}$$

The association depends primarily on the covariance. The variances merely serve to standardize the result. The covariance is estimated by

$$COV(X,Y) = \frac{1}{N-1}\sum_{t=1}^{N}(X_t-\overline{X})(Y_t-\overline{Y})$$

This differs from the present context where there is only one variable and we want to measure the association between observations separated by k units of time,

$$(X_1, X_{k+1}),\ (X_2, X_{k+2}), \cdots,\ (X_{N-k}, X_N)$$

Here, there are only N-k pairs instead of N pairs of observations. Furthermore, each random variable, X_t, is assumed to have the same mean, μ, so there is no point in

using separate estimates for the mean. As usual, the best estimate is $\bar{X}$ which is based on all the observations. Thus, the estimated autocovariance at lag k is

$$\gamma_k^* = \frac{1}{N-k}\sum_{t=1}^{N-k} (X_t-\overline{X})(X_{t+k}-\overline{X}) \qquad (1.a.1)$$

Although this estimate has a relatively small bias, most computer packages use an alternate estimate (Priestly, 1981, p.323)

$$\hat{\gamma}_k = \frac{1}{N}\sum_{t=1}^{N-k} (X_t-\overline{X})(X_{t+k}-\overline{X}) \qquad (1.a.2)$$

The estimate, $\hat{\gamma}_{k,}$ is preferred for a couple of reasons. One reason is the assertion (Parzen,1961) that $\hat{\gamma}_k$ is usually more precise, i.e., has a smaller mean square error. Thus, when k becomes large relative to N, the increase in precision more than compensates for the larger bias. (When k is small relative to N, $\hat{\gamma}$ is still preferred, since there is little difference in precision between the two estimators.)

However, another reason $\hat{\gamma}$ is used is we are interested in the function γ_k for all values of k from 1 to N-1. This function is positive semi-definite, as is $\hat{\gamma}_k$. The function γ_k^* is not necessarily positive semi-definite. Positive semi-definiteness means the roots of the autocorrelation matrix are ≥ 0. (Strictly positive definite implies the autocorrelations lie inside the interval [-1,1]). This has important implications for construction of spectra.

1.b. STATIONARITY

In time series analysis, it is usual to confine our attention to processes which are in statistical equilibrium. Such processes are called *stationary*, i.e., their statistical properties do not change over time. As time approaches infinity, the probability distribution of the X_t approximates a limit called the equilibrium distribution which does NOT depend on the initial conditions. Random processes which do not possess this property are called *nonstationary* or *evolutionary* processes, because the statistical properties evolve over time.

There are several ways to describe a stochastic process. One way is to specify the joint probability distribution. The behavior of a time series variable is determined by a probability distribution. The variables associated with different points in the time series are assumed to be statistically dependent. So for two time points, the joint behavior of X_t and X_s is given by the joint bivariate distribution. For k time points, the joint behavior of X_1, X_2, ..., X_k is given by the multivariate joint probability distribution. This method of describing stochastic processes is complicated and not usually attempted in practice (Chatfield, 1989, p.28).

Another method is to specify the full infinite sequence of moments. This method is simpler and useful, especially