1977

Applied forecasting with an autoregressive integrated moving average (ARIMA) model

Gail Ann Jensen
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Economics Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/16417

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Applied forecasting with an autoregressive integrated moving average (ARIMA) model

by

Gail Ann Jensen

A Thesis Submitted to the Graduate Faculty in Partial Fulfillment of The Requirements for the Degree of

MASTER OF SCIENCE

Major: Economics

Signatures have been redacted for privacy

Iowa State University
Ames, Iowa

1977
TABLE OF CONTENTS

I. INTRODUCTION
   A. Bayesian and Orthodox Methods 3
   B. The ARIMA Model 5
   C. Bayesian Analyses 16

II. STANDARD APPROACHES TO THE MODEL 33
   A. Choice of p, d, and q 34
   B. Estimation and Verification 38
   C. Forecasting 46

III. POST-BAYESIAN APPROACH: IDENTIFYING THE TRADE-OFFS 47
   A. Choice of Methods 47
   B. Inaccuracy and Complexity Costs 52

IV. EMPIRICAL ANALYSES 60
   A. The Experiment and Results 60
   B. Summary 69

V. REFERENCES 71

VI. ACKNOWLEDGMENTS 73

VII. APPENDIX 74
   A. Preliminary Estimates of the Parameters 74
   B. Back Forecasting 75
I. INTRODUCTION

The objective of this thesis is to develop and illustrate the "Post-Bayesian" approach [7, 8] to the problem of applied forecasting with an Autoregressive Integrated Moving Average (ARIMA) model. The study demonstrates the application of a cost-benefit approach to utilizing an integrated autoregressive model empirically by comparing two competing strategies for analysis. The comparison is in terms of model and forecast accuracy and the computational costs associated with each mode of analysis.

A complete theory of the Post-Bayesian approach to statistical inference has not been developed here; neither has a completely unified theory of the approach as it applies to forecasting with an ARIMA model. Rather, the basic concept of the trade-off between the costs of model complexity and model inaccuracy is discussed as it applies to the ARIMA model.

The paper is divided into four sections and an Appendix. Section I broadly identifies the Post-Bayesian approach to econometric problems. Also, in this part of the paper the ARIMA model and the basic concepts relating to the model are presented. The section is concluded with a Bayesian analysis of stationary autoregressive, invertible moving average, stationary invertible autoregressive moving average, and ARIMA processes.

In Section II, the usual sampling theory approach to the model is sketched out. (The Appendix supplements this section.)
The third section of the presentation discusses the Post-Bayesian approach to the model as an alternative to the modes of analysis presented in Sections I and II.

Section IV illustrates the application of the Post-Bayesian approach to four time series which are hypothesized as evolving from integrated autoregressive schemes. Some suggestions for further study are also presented in this final section.

Implicit in every empirical econometric analysis is the weighing of two costs: costs due to model complexity and costs due to model inaccuracy. The time and funds available to a researcher constrain the depth of his analysis. The problem he faces is to maximize the "accuracy" of his analysis subject to a given level of complexity, where allowable complexity is a function of his available time and funds.

Another investigator with both time and money strives to achieve a target level of accuracy, \( a^* \) (possibly set by his contractor or supervisor). In an economic framework his problem is to minimize complexity costs subject to \( E(\text{accuracy}) = a^* \).

In other settings, researchers may desire to minimize the expectation of the sum of complexity and inaccuracy costs.

Until recently a formal and applicable theory did not exist which dealt with such fundamental problems in statistics and econometrics. The complexity-inaccuracy trade-off was simply alluded to in discussions of model selection and estimation. The Post-Bayesian approach to econometric problems takes formal account of this trade-off in its
development of a descriptive and prescriptive statistical theory.

Posing econometric problems in an economic cost–benefit framework brings us closer to the way people actually behave and thus, hopefully, closer to reality. It is only natural to address an econometric investigation in terms of economic criteria. For example, the benefits from employing a model will depend in part on the model's level of accuracy as decided upon by the investigator. The link between policy objectives and his methods for specifying, estimating, testing, and applying the model is secured with his use of the Post-Bayesian cost–benefit approach.

No attempt will be made here to give a detailed presentation of the foundations of this approach. The interested reader is referred to Faden and Raussur [7, 8]. It is enough to mention that the approach has been fruitfully applied to the following problems:

1) estimating (or testing) the mean of a normal distribution
2) point estimation in general
3) simple hypothesis testing
4) optimal roundoff
5) optimal prediction using the normal multiple regression model
6) optimal one stage control
7) optimal N-state control

A. Bayesian and Orthodox Methods

Econometricians have traditionally employed non-Bayesian techniques to problems of statistical estimation. Recently however, a Bayesian approach has been applied to a variety of econometric problems. (For
a fairly comprehensive collection, see Zellner [20]). Arguments have been made for and against the adoption of Bayesian techniques in econometric inference. Since this is not the place to criticize or justify the use of either subjective informative priors or more typically, diffuse priors (e.g. Jeffreys [10]) a comparison of a sampling theory approach to a Bayesian approach will be made on the basis of computational complexity and accuracy of prediction. I feel that such a comparison directs itself to the factors an applied researcher considers when he decides among his alternative modes of analysis.

In the absence of complexity costs a researcher's ideal procedure for inference is taken to be the use of Bayes theorem where the marginal for \( \theta \) perfectly describes his prior beliefs. Also, given zero complexity costs and a loss function defined over all possible states of the world and all possible estimators, the author accepts the criterion that an estimator minimize expected risk.\(^1\) Average risk as a measure of the performance of alternative estimators is discussed and compared to other criteria by Chernoff and Moses [5, pp. 119-165].

It is well-known that the Bayes estimator (when the weighting function is taken as the prior described above) has this property, and it is unique in this sense. In fact, actual studies have shown that sampling theory

\(^1\)For the \( i \)th estimator, \( \hat{\theta}_i \), the risk function is the average loss associated with the estimator \( \hat{\theta}_i \) as a function of \( \theta \). The expected loss of the estimator is the scalar quantity obtained by taking a weighted average of the risk function over possible values of \( \theta \). The weighting function in the latter integral is designed to reflect personal probabilities in various regions of the parameter space.
methods for econometric models tend to yield inferior levels of accuracy relative to Bayesian techniques (especially when working with small samples).²

However, the use of Bayes theorem is often computationally complex relative to orthodox methods of parameter estimation.

For the econometric model to be examined in this paper, if by using a sampling theory estimator, we witness a small increase in expected risk (a proxy for inaccuracy) and a substantial decrease in complexity costs, then it may not be worthwhile to use the Bayesian prescription. After all, a large portion of statistical theory has been developed from a sampling theory point of view, and packaged computer programs are readily available for empirical application of standard techniques.

B. The ARIMA Model

1. Stationarity

In analyzing a time series, \((z_1, z_2, \ldots, z_N)\), we will regard it as the realization of a stochastic process, that is, as the observations of a statistical phenomenon evolving through time according to probabilistic laws. The stochastic process is stationary³ if and only if the joint


³This definition refers to "strict stationarity". Two other types of stationarity appear in the literature. Both are referred to as "weak stationarity of order n". One requires that the first n central moments of \(p(z_t, \ldots, z_{t+k})\) depend only on time differences. Another requires that the joint distribution of neighboring observations are identical up to order n, that is, marginals involving 1,2,\ldots, up to n adjacent observations are identical. Unless otherwise specified, stationarity will refer to strict stationarity.
distribution of any set of observations is unaffected by shifting all the times of observation by any integer amount m. Denoting the joint probability density function of \( z_t, \ldots, z_{t+k} \) by \( p(z_t, \ldots, z_{t+k}) \), the stochastic process is stationary if and only if

\[
p(z_t, \ldots, z_{t+k}) = p(z_{t+m}, \ldots, z_{t+k+m})
\]

for any point in time \( t \) and pair of integers, \( k \) and \( m \). A time series is homogeneous of degree \( d \) if the \( d \)th differences of the series are stationary, but not the \( d \)th minus first differences.

In practice, most economic time series are nonstationary since they usually display a time trend. However, as a useful working hypothesis, it is frequently assumed that \((z_1, z_2, \ldots, z_N)\) are observations from a homogeneous stochastic process.

For a nonstationary nonhomogeneous series, by taking the natural logarithms of the series homogeneity can sometimes be induced.

---

4 Here and throughout the thesis the symbol \( p \) will denote probability density functions generally and not a specific one. The argument of the function \( p \) will identify the particular density function being considered.

5 Defining the backward shift operator, \( B \), so that \( Bz_t = z_{t-1} \) and \( B^m z_t = z_{t-m} \), the \( d \)th differences of the series \( z_t \), \( t = 0, \pm 1, \pm 2, \ldots \) are defined as \( w_t \), \( t = 0, \pm 1, \pm 2, \ldots \) Where

\[
w_t = (1-B)^d z_t = \sum_{k=0}^{d} \binom{d}{k} (-1)^{d-k} B^{d-k} z_t.
\]
2. **Usefulness and flexibility of the model**

The ARIMA model is extremely useful and flexible for purposes of forecasting. It is flexible since it incorporates both purely autoregressive processes and purely moving average processes. Also, it allows for nonstationary homogeneity in either pure scheme. Its usefulness depends in part on its flexibility but primarily on other factors. It appears to be an appropriate choice of model when economic theory or intuition offer little in terms of explanatory variables available. Even if plausible structural models can be considered, the time and effort they require in estimation and analysis will be large relative to implementing a time series model such as the ARIMA.

Also, if a structural model is to be used for forecasting it is frequently necessary to estimate explanatory variables as well. When this is done the total forecast error for the endogenous variable may be so large that the forecast is unacceptable. This problem is somewhat alleviated by using the ARIMA model.

Even for those times when a regression model may be desirable, by using some form of an ARIMA model on the residuals, the predictive power of the model may be greatly enhanced.\(^6\)

Nelson [13] uses an ARIMA model to forecast 14 endogenous variables

---

\(^6\) Empirical examples of this combined modeling approach can be found in Pindyck and Rubinfeld, [14, Chapter 17].
of the Federal Reserve Board-MIT-Penn (FMP) quarterly model of the U.S. economy. His sample period is 1956-01 through 1966-04. His postsample predictions (one quarter ahead) using the ARIMA model have smallest mean square errors most frequently when compared to either the FMP forecasts or composite FMP-ARIMA forecasts.

Although on the average, the FMP predictions for the sample period had smaller mean square errors than the ARIMA predictions, he found that the coefficient \((1-\beta)\) of the composite linear model was significant at the 5 percent level for 10 out of the 14 cases. This clearly suggests that ARIMA models contribute to predictive accuracy. The composite postsample forecast came closer to the actual value in 12 out of the 14 cases when compared to the FMP forecast.

Naylor et al. [12] estimated ARIMA models for 4 endogenous variables of the Wharton quarterly model of the U.S. economy. Their sample period was 1963-02 through 1967-04. One step ahead forecasts for the 4 endogenous variables throughout the sample period were computed for both the ARIMA and Wharton models.

For each variable, the average absolute error of the ARIMA forecasts was considerably smaller (almost half in size) than the average absolute

\[ A_t = \beta (FMP)_t + (1-\beta) (ARIMA)_t + \epsilon_t, \]

over the sample period where:

- \(A_t\) = actual value for time \(t\)
- \((FMP)_t\) = FMP prediction for time \(t\)
- \((ARIMA)_t\) = ARIMA prediction for time \(t\)
- \(\epsilon_t\) = composite prediction error assumed NID \((0, \sigma^2)\).
error of the Wharton Simulation forecasts.

The success of the model as a forecasting tool is also demonstrated in Box and Jenkins [3, 4], Nelson [13], and Pindyck and Rubinfeld [14]. Naylor and Seaks also provide a list of other investigators who have used ARIMA models for forecasting with much success.

3. Notation

For notational convenience define the following:

The variance of $z_t$
$$\gamma_0 = E[z_t - E(z_t)]^2$$

The auto covariance at lag $j$ for a stationary process $z_t$, $t=0, \pm 1, \pm 2, \ldots$ is defined as
$$\gamma_j = E[z_t - E(z_t)][z_{t+j} - E(z_{t+j})]$$

The autocorrelation at lag $j$
$$\rho_j = \rho_{-j} = \gamma_j / \gamma_0$$

The autocorrelation function is the sequence $\rho_0, \rho_1, \ldots$

The difference operator, $\Delta$
$$\Delta = 1 - B \quad \text{and} \quad \Delta^d = (1 - B)^d$$

And the following polynomials in the backward shift operator, $B$
We write \( f(x, y) \preceq g(x, y) \) if there is a function \( h(y) \) such that 
\[ f(x, y) = h(y)g(x, y); \] 
\( \preceq \) alone indicates that \( h \) is a constant.

The probability density function written as \( p(x|y) \) is the density function of \( x \) conditional on a specific value of \( y \).

4. The model

The Autoregressive Integrated Moving Average (ARIMA) model of order \( (p, d, q) \) can be represented as

\[ \zeta(B)z_t = \delta + \theta(B)a_t \]

or

\[ \phi(B)\omega_t = \delta + \theta(B)a_t \]

\[ \omega_t = \Delta^d z_t \]

It is assumed \( \omega_t \) is a stationary time series, that is, \( z_t \) is homogeneous of degree \( d \). The \( a_t \) are a stationary time series distributed
with common mean $0$ and finite variance $\sigma^2$. For statistical analysis it is usually assumed that the $a_t$ are normally and independently distributed.

$\zeta(B)$ is called the autoregressive operator, $\phi(B)$, the autoregressive operator of the stationary series, $w_t$, and $\theta(B)$, the moving average operator, if $q=0$ then $w_t$ is simply autoregressive (AR) of order $p$, that is, $z_t$ is an integrated autoregressive (ARI) process of order $(p, d)$. If $p=0$ then $w_t$ follows a moving average (MA) process, or in terms of the original time series, $z_t$ is an integrated moving average (IMA) process of order $(d, q)$.

Typically the stationary series, $w_t$, is first analyzed and forecasted. $z_t$ is forecasted by summing or integrating over the stationary series in accordance with $d$, the order of differencing.

There are three forms for the general model; each is useful or illuminating in a particular way. Without loss of generality, we will omit $\delta$ from the specification. Provided $w_t$ is interpreted as the deviation from the mean, $E(w_t)$, the omission is justified. The current value of $z_t$ can be expressed as:

(a) a weighted sum of previous values of the $z$'s and current and previous values of the $a$'s.

(b) a weighted sum of current and previous values of the $a$'s only

(c) a weighted sum of previous values of the $z$'s and the current shock $a_t$.

The difference equation form of the model, (a) may be written as
\[ z_t = \zeta_1 z_{t-1} + \zeta_2 z_{t-2} + \ldots + \zeta_{p+d} z_{t-d-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \]

If we allow the additional assumption that the z's are stationary, then we can identify this form as the one used for a Bayesian analysis of the model. This form is also most frequently implemented for forecasting when a standard approach to the model is used.

The random shock form of the model, (b), can be written as

\[ z_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots = \psi(B)a_t \]

The weights \( \psi_k \) are functions of the \( \zeta_i \) and \( \theta_j \) coefficients in the difference equation form of the model. Operating on both sides of the above equation with \( \zeta(B) \) yields

\[ \zeta(B)z_t = \zeta(B)\psi(B)a_t \]

However,

\[ \zeta(B)z_t = \theta(B)a_t \]

so that

\[ \zeta(B)\psi(B) = \theta(B) \]

By equating coefficients of \( B \), term by term, the \( \psi_k \) weights can be determined, although the simplest way to determine the \( \psi_k \)'s is by successive substitution for \( z_{t-1}, z_{t-2}, \) and so forth, using the difference equation form, (a). The random shock form of the process is often used with \( \sigma^2 \) to obtain the variance and autocovariances of \( z_t \).

It is this form also that is used to obtain estimates of the variance of
sampling theory forecasts of more than one time period ahead.

The inverted form of the model may be written as

\[ z_t = \sum_{i=1}^{\infty} z_{t-i} + a_t \]

or as

\[ \Pi(B)z_t = a_t \]

For an ARMA process, it is usually required that the series \( \Pi(x) \) converges for \( |x| < 1 \) (\( x \) complex). Since \( \Pi(B) = \theta^{-1}(B)\zeta(B) \), this is equivalent to requiring that \( \theta^{-1}(x) \) converges for \( |x| < 1 \) (x complex) since \( \zeta(x) \) is a polynomial. If this condition is satisfied, the process is said to be invertible.

Since \( \zeta(B)z_t = \theta(B)a_t \)

substitution for \( a_t \) in terms of \( z_t \) yields

\[ \zeta(B)z_t = \theta(B)\Pi(B)z_t \]

or that

\[ \zeta(B) = \theta(B)\Pi(B) \]

Again, by equating coefficients of \( B \) in the left and right hand series we can derive the \( \Pi_j \). If \( d > 1 \) and the process is invertible, it is easily proven that \( \sum_{j=1}^{\infty} \Pi_j = 1 \).\(^8\)

---

\(^8\)Proof: for \( d > 1 \), \( \zeta(B) = \phi(B)(1-B)^d = 0 \) when \( B=1 \). By the invertibility condition, the roots of \( \theta(B) \) lie outside the unit circle, i.e. \( \theta(1) \neq 0 \). Hence \( \zeta(B) = \Pi(B)\theta(B) + \Pi(1) = 0 \), \( \therefore \sum_{j=1}^{\infty} \Pi_j = 1 \). The connection between the invertibility condition and the roots of \( \theta(B) \) is discussed in the next section.
5. Assumptions for stationarity and invertibility

For convenience, let \( z_t \) be interpreted as the deviation from its mean, \( E(z_t) \). Suppose \( z_t \) is generated by the ARMA \((p,q)\) process given by

\[
\phi(B)z_t = \theta(B)a_t
\]

Operating on both sides with \( \phi^{-1}(B) \) yields

\[
z_t = \phi^{-1}(B)\theta(B)a_t = \Psi(B)a_t
\]

For \( z_t \) to follow a normal stationary process it is necessary that \( E(z_t) \) and the autocovariance matrix of any set of \( k \) neighboring observations are fixed through time, and finite. (By finite autocovariance matrix is meant each element is finite.) For \( \text{Var}(z_t) \) finite, we must have that the power series \( \Psi(B) \), above, evaluated at \( B=1 \) converges. \( \theta(1) \) being a polynomial with constant term \( \neq 0 \), \( \Psi(1) \) will converge if and only if \( \phi^{-1}(1) \) converges. And, if the radius of convergence, \( R \), of the power series \( \phi^{-1}(B) \) is greater than unity, we are assured that \( \phi^{-1}(1) \) converges.

Now, letting \( \lambda_1, \lambda_2, \ldots, \lambda_p \) be the roots of \( \phi(B) \) we have

\[
\phi^{-1}(B) = [(1-B/\lambda_1)(1-B/\lambda_2) \ldots (1-B/\lambda_p)]^{-1}
\]

\[
= (1-B/\lambda_1)^{-1}(1-B/\lambda_2)^{-1} \ldots (1-B/\lambda_p)^{-1}
\]

If \( R_i \) \((i = 1, 2, \ldots, p)\) represents the radius of convergence of the \( i \)th power series,

\[
(1-B/\lambda_i)^{-1} = 1 + B/\lambda_i + B^2/\lambda_i^2 + B^3/\lambda_i^3 + \ldots
\]
it is true that the radius of convergence, $R$, of $\phi^{-1}(B)$ will equal the minimum of $[R_1, R_2, \ldots, R_p]$. Hence $R>1$ implies $\min[R_1, R_2, \ldots, R_p] > 1$.

But since $R_i = |\lambda_i|$, we must have that $\min[|\lambda_1|, |\lambda_2|, \ldots, |\lambda_p|]$ is greater than unity. Put another way, if the roots of the characteristic equation,

$$\phi(B) = 0$$

lie outside the unit circle, $\phi^{-1}(B)$ has a radius of convergence $R>1$.\footnote{A detailed proof that this is a strict condition for stationarity is presented in Grenander and Rosenblatt [9].}

In the context of the present problem, this means $z_t$ will have finite variance as required.\footnote{By Schwarz's inequality, $\text{Var}(z_t)$ finite guarantees all auto-covariances will be finite.}

The condition that the roots of the characteristic equation

$$\phi(B) = 0$$

lie outside the unit circle can be expressed as a set of restrictions (necessary and sufficient) on the parameters $\phi_1, \phi_2, \ldots, \phi_p$.\footnote{For reference, see either Samuelson [16], or Chipman [6].}

These restrictions define a stability region in $\mathbb{R}^p$.

Note that all purely moving average processes are automatically stationary since $\theta_1 + \theta_2 + \cdots + \theta_q < \infty$.

The invertibility condition in the present content requires that

$$\Pi(B) = \psi^{-1}(B) = \theta^{-1}(B)\phi(B)$$
converge when $B = 1$. As before, since $\phi(1)$ is finite, $\Pi(1)$ converges if and only if $\phi^{-1}(1)$ converges, and we can be assured this will occur if the radius of convergence, $R$, of the power series $\theta^{-1}(B)$ is greater than unity. Using the results of Grenander and Rosenblatt [9], this occurs if and only if the roots of the characteristic equation

$$\theta(B) = 0$$

lie outside the unit circle. Again, this restriction on the roots can be translated into a set of necessary and sufficient conditions imposed on $\theta_1, \theta_2, \ldots, \theta_q$ which define an invertible region in the parameter space, $R^q$.

This section is concluded with a Bayesian analysis of:

1) Stationary autoregressive processes
2) Invertible moving average processes
3) Stationary, invertible autoregressive moving average processes
4) ARIMA processes

C. Bayesian Analyses

1. Stationary autoregressive processes

We represent the autoregressive process of order $p$ as

$$w_t = \delta + \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} + a_t, \quad t=1,2,\ldots,T \hspace{1cm} (1)$$

and assume the following:
(a) \( w_t \) is a stationary time series

(b) \( p(a_{p+1}, a_{p+2}, \ldots, a_T | w_1, \ldots, w_p, \sigma, \phi) \sim N(0, \sigma^2 I) \)

where

\[ \phi' = (\phi_1, \phi_2, \ldots, \phi_p) \]

Using Bayes' theorem we will derive the posterior for \( \sigma \) and \( \phi \) in the following way:

\[
p(\sigma, \delta, \phi | w_1, w_2, \ldots, w_T) \propto p(\sigma, \delta, \phi, w_1, w_2, \ldots, w_T)
\]

\[
= p(w_1, w_2, \ldots, w_T | \sigma, \delta, \phi) p(\sigma, \delta, \phi)
\]

\[
= p(w_{p+1}, \ldots, w_T | w_1, w_2, \ldots, w_p, \sigma, \delta, \phi) p(w_1, w_2, \ldots, w_p | \sigma, \delta, \phi) p(\sigma, \delta, \phi)
\]

Obtaining \( p(w_{p+1}, w_{p+2}, \ldots, w_T | w_1, w_2, \ldots, w_p, \sigma, \delta, \phi) \) is straightforward. The inverse transformation is given by

\[
a_t = w_t - \delta - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p} \quad t = p+1, p+2, \ldots, T
\]

The Jacobian of the inverse transformation, \( (a_{p+1}, a_{p+2}, \ldots, a_T) \) in terms of \( (w_{p+1}, w_{p+2}, \ldots, w_T) \) is triangular with determinant equal to unity. Specifically, the Jacobian is

\[
\begin{bmatrix}
a_{p+1} & a_{p+2} & \cdots & a_T \\
w_{p+1} & 1 & -\phi_1 & -\phi_2 & \cdots & -\phi_p & 0 \\
w_{p+2} & 1 & -\phi_1 & \cdots & -\phi_p & 0 & 0 \\
& \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
w_T & 0 & \cdots & \cdots & \cdots & \cdots & 1
\end{bmatrix}
\]
Since $p(a_{p+1}, \ldots, a_T | w_1, \ldots, w_p, \sigma, \delta, \phi) \sim N(0, \sigma^2 I)$ we have

$$p(w_{p+1}, \ldots, w_T | w_1, \ldots, w_p, \sigma, \delta, \phi)$$

$$\propto \frac{1}{\sigma^{T-p}} \exp\left[-\frac{1}{2} \sum_{t=p+1}^{T} \left( w_t - \delta - \phi_1 w_{t-1} - \ldots - \phi_p w_{t-p} \right)^2 \right]$$

To obtain $p(w_1, \ldots, w_p / \sigma, \delta, \phi)$ we use the assumption that the $w$'s evolved from a stationary process. Thus each $w$ has the same mean.

From (1),

$$Ew_t = \delta + \phi_1 Ew_{t-1} + \phi_2 Ew_{t-2} + \ldots + \phi_p Ew_{t-p} + E\alpha_t$$

Solving for $Ew_t$ yields

$$Ew_t = \delta / (1 - \phi_1 - \phi_2 - \ldots - \phi_p)$$

For easier manipulation, we will scale the process to obtain the covariances. Interpreting $w_t$ as the deviation from its mean,

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} + \alpha_t$$

(2)

For $k = 1, 2, \ldots, p$, multiplying (2) by $w_{t-k}$ yields $w_{t-k}w_t = \phi_1 w_{t-k}w_{t-1} + \phi_2 w_{t-k}w_{t-2} + \ldots + \phi_p w_{t-k}w_{t-p} + w_{t-k}\alpha_t$.

Taking expectations we have

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \ldots + \phi_p \gamma_p + \sigma^2$$

(3)

12 The result $E(w_1 \alpha_t) = \sigma^2$ is easily seen if $w_t$ is first written in the random shock form before expectations are taken.
where
\[ \gamma_{-j} = \gamma_j \]

[Note \( E(w_{t-k} a_t^i) = 0 \) for \( k > 0 \) since \( w_{t-k} \) involves shocks \( a_j \) up to time \( t-k \) which are uncorrelated with \( a_t \).] For displacements \( k \) greater than \( p \), the covariances are determined from

\[ \gamma_{k} = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \ldots + \phi_p \gamma_{k-p} \]

Dividing the \( p \) equations in (4) by \( \gamma_0 \) given in (3) yields the system (known as the Yule-Walker equations)

\[
\begin{align*}
\rho_1 &= \phi_1 + \phi_2 \rho_1 + \ldots + \phi_p \rho_{p-1} \\
\vdots & \quad \vdots \\
\rho_p &= \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \ldots + \phi_p 
\end{align*}
\]

Solving this system for \( \rho_1, \rho_2, \ldots, \rho_p \) in terms of \( \phi_1, \phi_2, \ldots, \phi_p \), we can readily set up the Variance-Covariance matrix of \( w_1, \ldots, w_p \) given by

\[
\Omega(\sigma, \phi) = \gamma_0 \begin{bmatrix}
1 & \rho_1 & \rho_2 & \ldots & \rho_{p-1} \\
1 & 1 & \rho_2 & \ldots & \rho_{p-2} \\
1 & 1 & 1 & \ddots & \vdots \\
& & & \text{symm.} & 1 & \rho_1 \\
& & & & 1 & 1 
\end{bmatrix}
\]
Assuming \( p(w_1, \ldots, w_p | \sigma, \delta, \phi) \sim \text{Normal} \), we have

\[
p(w_1, \ldots, w_p | \sigma, \delta, \phi) \propto |\Omega(\sigma, \delta, \phi)|^{-1/2} \exp\left[-\frac{1}{2}(w_1, \ldots, w_p)\Omega^{-1}(\sigma, \delta, \phi)(w_1, \ldots, w_p)'ight]
\]

For the prior density of \( \sigma, \delta, \) and \( \phi \), a particular prior (uniform in \( \delta \), \( \phi \), and \( \log \sigma \)) is chosen for \( \phi \) compatible with a stationary process, namely,

\[
p(\sigma, \delta, \phi) \propto 1/\sigma \text{ for all } \phi \text{ such that the roots of the characteristic equation } \phi(B) = 0 \text{ lie outside the unit circle}
\]

0 otherwise

Collecting results,

\[
p(\sigma, \delta, \phi | w_1, w_2, \ldots, w_T) \propto p(\sigma, \delta, \phi, w_1, w_2, \ldots, w_T)
\]

\[
\propto |\Omega(\sigma, \delta, \phi)|^{-1/2} \exp\left[-\frac{1}{2}(w_1, \ldots, w_T)\Omega^{-1}(\sigma, \delta, \phi)(w_1, w_2, \ldots, w_T)'ight]
\]

\[
+ (w_1, w_2, \ldots, w_p)\Omega^{-1}(\sigma, \delta, \phi)(w_1, w_2, \ldots, w_p)'
\]

for \( \phi \) compatible with a stationary process, 0 otherwise.

To make a one period ahead Bayesian point forecast, \( \hat{w}_{T+1}^{\prime} \), we assume that our loss, \( L \), in missing the true value, \( w_{T+1} \), is quadratic, that is

\[
L = c(w_{T+1} - \hat{w}_{T+1}^{\prime})^2, \quad (c>0)
\]

The posterior expectation of the quadratic loss function conditional on \( w_1, w_2, \ldots, w_T \) with \( \delta, \phi_1, \ldots, \phi_p \), treated here as random variables is
\[ E(L|w_1, \ldots, w_T) = E \left( \left( w_{T+1} - \hat{w}_{T+1} \right)^2 \right| w_1, \ldots, w_T \] 

\[ = cE \left( \left( w_{T+1} - E(w_{T+1}|w_1, \ldots, w_T) \right)^2 \right| w_1, \ldots, w_T \] 

\[ - \left( \hat{w}_{T+1} - E(w_{T+1}|w_1, \ldots, w_T) \right)^2 \right| w_1, \ldots, w_T \] 

\[ = cE \left( \left( w_{T+1} - E(w_{T+1}|w_1, w_2, \ldots, w_T) \right)^2 \right| w_1, \ldots, w_T \] 

\[ + c \left( \hat{w}_{T+1} - E(w_{T+1}|w_1, \ldots, w_T) \right)^2 \] 

since the cross product term disappears.

Only the second term in the last expression involves \( \hat{w}_{T+1} \). It is nonstochastic and will be minimized if we take

\[ \hat{w}_{T+1} = E(w_{T+1}|w_1, w_2, \ldots, w_T) \]

Since \( w_{T+1} = \delta + \phi_1 w_T + \phi_2 w_{T-1} + \ldots + \phi_p w_{T-p+1} + \epsilon_{T+1} \), we have

\[ E(w_{T+1}|w_1, w_2, \ldots, w_T) = E\delta + E\phi_1 w_T + E\phi_2 w_{T-1} + \ldots + E\phi_p w_{T-p+1} \]

where \( E\delta \) and \( E\phi_1 \) are the posterior expectations of \( \delta \) and \( \phi_1 \) (i=1,2,\ldots,p).

The derivation of Bayesian forecasts \( \ell \) periods ahead (\( \ell > 1 \)) given the quadratic loss function is considerably more complicated if they are unconditional on \( w_{T+1}, w_{T+2}, \ldots, w_{T+\ell-1} \). As an illustration, the optimal point prediction for \( w_{T+2}, \hat{w}_{T+2} \), is

\[ E(w_{T+2}|w_1, \ldots, w_T) = E\delta + E[\phi_1 (\delta + \phi_1 w_T + \phi_2 w_{T-1} + \ldots + \phi_p w_{T-p+1})] \] 

\[ + E\phi_2 w_T + \ldots + E\phi_p w_{T-p+2} \]

The evaluation of the second term on the right hand side of the equation involves the numerical integration of \( p \) integrals, each two fold. However, if we substitute the obtained forecast, \( \hat{w}_{T+1} \), for \( w_{T+1} \),
\[ \hat{\omega}_{T+2} \] becomes simply

\[ \hat{\omega}_{T+2} = E\delta + E\phi_1 \hat{\omega}_{T+1} + E\phi_2 \omega_T + \ldots + E\phi_p \omega_{T-p+2}. \]

Using this method, the forecast for \( \omega_{T+3} \) would be

\[ \hat{\omega}_{T+3} = E\delta + E\phi_1 \hat{\omega}_{T+2} + E\phi_2 \hat{\omega}_{T+1} + E\phi_3 \omega_T + \ldots + E\phi_p \omega_{T-p+3} \]

Forecasts for greater lead times can easily be completed by this step-by-step recursive calculation procedure.

In general, under a quadratic loss function, the optimal point estimate for \( \omega_{T+\ell} \) conditional only on \( \omega_1, \ldots, \omega_T \) will be very complex computationally relative to the estimate computed in the step-by-step manner, that is, when

\[ \omega_{T+1} = \hat{\omega}_{T+1}, \omega_{T+2} = \hat{\omega}_{T+2}, \ldots, \omega_{T+\ell-1} = \hat{\omega}_{T+\ell-1}. \]

The latter forecast can be thought of an approximation to the more complex unconditional point forecast.

In the empirical portion of this paper, the two types of forecasts are compared, both in terms of computational costs and accuracy of prediction.

2. Invertible moving average processes

We represent the moving average process of order \( q \) as

\[ \omega_t = \delta + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}, \quad t = 1, 2, \ldots, T \tag{5} \]

and assume the following:
(a) the process is invertible

(b) \( p(a_{1-q}, \ldots, a_T | \sigma, \delta, \theta) \sim N(0, \sigma^2 I) \)

where

\[ \theta' = (\theta_1, \theta_2, \ldots, \theta_q) \]

The system in (5) can be written

\[
\begin{bmatrix}
    w_1 \\
    w_2 \\
    \vdots \\
    w_T
\end{bmatrix} = \begin{bmatrix}
    \delta \\
    \delta \\
    \vdots \\
    \delta
\end{bmatrix} + \begin{bmatrix}
    -\theta_q & -\theta_{q-1} & \cdots & -\theta_1 & 1 \\
    -\theta_q & -\theta_{q-1} & \cdots & -\theta_1 & 0 \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    -\theta_q & -\theta_{q-1} & \cdots & -\theta_1 & 0
\end{bmatrix} \begin{bmatrix}
    a_{1-q} \\
    a_{2-q} \\
    \vdots \\
    a_T
\end{bmatrix}
\]

or in matrix form, \( w = \delta + Aa \) [A(T,T+q)] where \( w' = (w_1, \ldots, w_T) \), \( \delta = (\delta, \ldots, \delta) \), \( a' = (a_{1-q}, \ldots, q_T) \), and A is the matrix above.

Since \( w \) is a linear transformation of \( a \) where \( p(a | \sigma, \delta, \theta) \sim N(0, \sigma^2 I) \),

\[ p(w | \sigma, \delta, \theta) \sim \text{Normal with} \]

\[ Ew = \delta + AEa = \delta \]

and

\[ E[(w-\delta)(w-\delta)'] = A\sigma^2 A' = AA'\sigma^2 \]

so that

\[ p(w_1, \ldots, w_T | \sigma, \delta, \theta) \propto |AA'|^{-1/2} \exp[-\frac{1}{2\sigma^2} (w-\delta)'(AA')^{-1}(w-\delta)] \]

The covariance matrix of \( w_1, w_2, \ldots, w_T \).
\[ AA' \sigma^2 = \begin{bmatrix}
\gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_q & 0 \\
\gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_q \\
\gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_q \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \gamma_0 \\
\gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_q
\end{bmatrix}_{T \times T} \]

where

\[ \gamma_0 = \sigma^2 (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2) \]

\[ \gamma_k = \sigma^2 (-\theta_k + \theta_1 \theta_{k+1} + \cdots + \theta_{q-k} \theta_q) \quad k = 1, 2, \ldots, q \]

The covariance for lags \( k > q \) is 0 since \( w_t \) involves only disturbances up to \( t-q \).

For the prior density of \( \sigma \) and \( \theta \), a particular prior is chosen for \( \theta \) compatible with an invertible process, namely \( p(\sigma, \delta, \theta) \propto 1/\sigma \) for all \( \theta \) such that the roots of the characteristic equation \( \theta(B) = 0 \) lie outside the unit circle, 0 otherwise.

Combining the prior and likelihood via Bayes theorem gives us

\[
p(\sigma, \delta, \theta | w_1, w_2, \ldots, w_T) \propto \frac{|AA'|^{1/2}}{\sigma^{T+1}} \exp\left[-\frac{1}{2\sigma^2} (w-\delta)'(AA')^{-1}(w-\delta)\right]
\]

for \( \theta \) compatible with an invertible process, 0 otherwise.

Since \( AA' \) does not involve \( \sigma \), this density is an inverted gamma distribution in \( \sigma \). Integrating out \( \sigma \) yields
Forecasting one period ahead with a purely moving average model is not as straightforward as it was for the autoregressive model since the \(a_t\) are unobservable. Derivation of the predictive probability density function for \(w_{T+\ell}\) (\(\ell \geq 1\)) is the most direct route for obtaining a forecast. Letting \(\Delta = (\sigma, \delta, \theta)\) we have

\[
p(w_{T+\ell} \mid w_1, w_2, \ldots, w_T) = \int_{\mathbb{R}^{q+2}} \frac{p(w_{T+\ell}, \Delta \mid w_1, w_2, \ldots, w_T) p(\Delta)}{\int_{\mathbb{R}^{q+2}} [p(w_1, w_2, \ldots, w_T \mid \Delta) p(\Delta)] d\Delta} d\Delta
\]

The integrand in the denominator has already been derived. It is proportional to the joint distribution of \(\sigma, \delta, \theta,\) and \(w_1, \ldots, w_T\) (proportional since \(p(\Delta)\) is improper). For \(\ell \geq 2\) the derivation of the integrand in the numerator is fairly complex. But even with its functional form in hand, the computation of a single mean square error forecast requires \(2q+5\) integrations, most of which would be numerically preformed. (The integral in the denominator of the last expression may possibly not converge. In this case \(p(w_{T+\ell}, \Delta \mid w_1, w_2, \ldots w_T)\) should be directly integrated with respect to \(\Delta\).)
3. **Stationary invertible autoregressive moving average processes**

We represent the autoregressive moving average process of order \((p,q)\) as

\[
\begin{align*}
    w_t &= \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} + u_t \\
    u_t &= \delta + a_t - \theta_1 q_{t-1} - \theta_2 q_{t-2} - \ldots - \theta_{q-t+1} q_{t-q} \quad t = 1, 2, \ldots, T
\end{align*}
\]  

and assume the following:

(a) the process is stationary

(b) the process is invertible

(c) \(p(a_{1-q}, \ldots, a_T | a, \delta, \phi, \theta) \sim N(0, \sigma^2 I)\)

where

\[
\phi' = (\phi_1, \phi_2, \ldots, \phi_p) \quad \text{and} \quad \theta' = (\theta_1, \theta_2, \ldots, \theta_q)
\]

To obtain the likelihood function,

\[
p(w_1, \ldots, w_T | a, \delta, \phi, \theta)
\]

we will first obtain

\[
p(u_1, \ldots, u_T | a, \delta, \phi, \theta).
\]

Since the \(w\)'s are stationary and are a linear transformation of the \(u\)'s, we can then derive the density for the \(w\)'s.

The system in (7) can be written as

\[
u = \delta + Aa
\]

where
\[ \delta' = (\delta, \delta, \ldots, \delta) \text{ is } T \text{ dimensional} \]
\[ u' = (u_1, \ldots, u_T) \]
\[ a' = (a_{1-q}, a_{2-q}, \ldots, a_T) \]

and

\[ A = \begin{bmatrix}
-\theta & 0 & \cdots & 0 \\
0 & -\theta & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & -\theta
\end{bmatrix}_{T\times(T+q)} \]

Note that \( u \) takes the place of \( w \) in the analysis of purely moving average processes.

As before,

\[ Eu = \delta + \Lambda \varepsilon a = \delta \]

and

\[ E[(u-\delta)(u-\delta)'] = \Lambda \sigma^2 A' = AA' \sigma^2 \]

As a linear transformation of multinormal random variables, the multinormal density for \( u_1, u_2, \ldots, u_T \) is given by

\[ p(u_1, \ldots, u_T | \sigma, \delta, \theta, \phi) \]

\[ \propto \left| \frac{AA'}{T} \right|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (u-\delta)' (AA')^{-1} (u-\delta) \right\} \]

To obtain \( p(w_1, \ldots, w_T | \delta, \sigma, \phi, \theta) \) we use the assumption that the \( w \)'s
evolved from a normal stationary process. Taking expectations in (6) we have

\[ \text{E}w = \text{E}w(\phi_1 + \phi_2 + \ldots + \phi_p) + \delta \]

or

\[ \text{E}w = \delta/(1 - \phi_1 - \phi_2 - \ldots - \phi_p) \]

To derive the variance and autocovariances of \( w_1, \ldots, w_T \), we will write the model as

\[ w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} + a_t - \phi_1 a_{t-1} - \phi_2 a_{t-2} - \ldots - \phi_p a_{t-p} \]

and interpret \( w_t \) as the deviation from its mean.

Multiplying (8) through by \( w_t \) and taking expectations we have

\[ \gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \ldots + \phi_p \gamma_p + \sigma^2 - \theta_1 \gamma_{wa}(-1) - \ldots - \theta_q \gamma_{wa}(-q) \]

Multiplying through by \( w_{t-k} \) \((k>0)\) in (8) and taking expectations, we find that the autocovariances satisfy the difference equation

\[ \gamma_k = \phi_1 \gamma_{k-1} + \ldots + \phi_p \gamma_{k-p} + [\gamma_{wa}(k) - \theta_1 \gamma_{wa}(k-1) - \ldots - \theta_q \gamma_{wa}(k-q)] \]

The expression in brackets on the right hand side is the covariance between \( w_{t-k} \) and \( u_t \). \( \gamma_{wa}(k) = \text{E}(w_{t-k}a_t) \) and \( \gamma_{wa}(k) = 0 \) for \( k>0 \) since \( w_{t-k} \) involves only those shocks which have occurred up to time \( t-k \). For \( k \geq q+1 \) the bracketed expression is zero since the argument of each \( \gamma_{wa} \) is greater than zero.

Thus for \( k>q+1 \)
These covariances at lags greater than $q$ follow the pattern determined by the autoregressive portion of the process.

Equation (8) and the first $p+q$ equations of (9) comprise a system involving the $p+q+1$ parameters $\sigma^2, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ and $\gamma_0, \gamma_1, \ldots, \gamma_{p+q}$.

The system can be solved simultaneously and $\gamma_0, \ldots, \gamma_{p+q}$ expressed in terms of $\sigma^2, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$. $\gamma_{p+q+1}, \ldots, \gamma_{t-1}$ can be expressed in terms of $\sigma^2, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ via the difference equation

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \ldots + \phi_p \gamma_{k-p}$$

Upon solving, the variance-covariance matrix, $\Omega$, of $w_1, \ldots, w_T$ can be set up

$$\Omega(\sigma, \phi, \theta) = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & & \\ \vdots & \ddots & \ddots & \\ \gamma_{T-1} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}$$

Since $w_t$ follows a normal process, we have

$$p(w_1, \ldots, w_T | \sigma, \theta, \phi, \delta) = \alpha |\Omega(\sigma, \phi, \theta)|^{-1/2} \exp \left[ -\frac{1}{2} (w-\mu)'[\Omega(\sigma, \phi, \theta)]^{-1/2}(w-\mu) \right]$$

where $w' = (w_1, w_2, \ldots, w_T)$ and $\mu$ is the $T$ dimensional vector of which each element is $\delta/(1-\phi_1-\phi_2-\ldots-\phi_p)$.

For a marginal over $\delta, \phi, \theta$, and $\sigma$, the following prior, uniform in
\[ \delta, \phi, \theta, \text{ and log } \sigma, \text{ is chosen.} \]

\[ p(\delta, \phi, \theta, \sigma) = 1/\sigma \text{ for } \phi \text{ and } \theta \text{ such that the roots of the two characteristic equations, } \phi(B) = 0 \text{ and } \theta(B) = 0 \text{ lie outside the unit circle} \]

\[ 0 \text{ otherwise} \]

Applying Bayes theorem, we arrive at the posterior for \( \sigma, \delta, \phi, \) and \( \theta \) given by

\[ p(\sigma, \delta, \phi, \theta | w_1, \ldots, w_T) \propto \frac{1}{\sigma} |\Omega(\sigma, \phi, \theta)|^{-1/2} \exp \left\{ -\frac{1}{2}(w-\mu)'[\Omega(\sigma, \phi, \theta)]^{-1}(w-\mu) \right\} \]

times \( (w-\mu) \) for \( \phi \) and \( \theta \) compatible with a stationary, invertible process.

0. otherwise

Forecasts for the autoregressive moving average model can be directly obtained via the predictive probability density function given the chosen loss function. The predictive density is complex computationally. Letting \( \Delta = (\sigma, \delta, \phi, \theta) \), it is theoretically given by

\[ p(w_{t+q} | w_1, \ldots, w_T) = \int_{\mathbb{R}^{p+q+2}} p(w_{t+q+2}, \Delta | w_1, \ldots, w_T) d\Delta \]

\[ = \int_{\mathbb{R}^{p+q+2}} p(w_{t+q+2}, w_1, w_2, \ldots, w_T | \Delta) p(\Delta) d\Delta \]

\[ = \int_{\mathbb{R}^{p+q+2}} p(w_1, \ldots, w_T | \Delta) p(\Delta) d\Delta \]

Analogous to earlier results, the denominator's integrand has already been derived; it is the joint density for the parameters of the model, \( \sigma, \)
\( \delta, \phi, \theta, \) and \( w_1, \ldots, w_T \). (As noted before, if integral in the
denominator diverges \( p(w_{T+1}, \Delta|w_1, \ldots, w_T) \) should be integrated directly
with respect to \( \Delta \).)

4. **ARIMA processes**

For the ARMA \((p, q)\) process, it has been shown that

\[
p(w_1, \ldots, w_T|\phi, \theta, \delta, \sigma^2) \propto \left| \Omega(\sigma, \phi, \theta) \right|^{-1/2} \exp\left\{ -\frac{1}{2} (w-\mu)' [\Omega(\sigma, \phi, \theta)]^{-1} (w-\mu) \right\}
\]

Now, if for \( t = 1, 2, \ldots T \),

\[
w_t = (1-B)^d z_t = \sum_{k=0}^{d} \binom{d}{k} (-1)^{d-k} \binom{d-1}{k} z_{t-k}
\]

the likelihood,

\[
p(z_1, z_2, \ldots, z_T|z_0, \ldots z_{1-d}, \phi, \theta, \delta, \sigma^2)
\]

can be derived as follows.

The inverse of the transformation from \( z \) to \( w \) is given above and the
Jacobian,

\[
J = \begin{bmatrix}
w_1 & w_2 & w_3 & \cdots & w_T \\
z_1 & 1 & (-1)^{d-1} \binom{d}{1} & (-1)^{d-2} \binom{d}{2} & 0 \\
z_2 & 1 & 1 & (-1)^{d-1} \binom{d}{1} \\
z_k & 0 & \binom{d}{k} (-1)^{d-k} & (-1)^{d-1} \binom{d}{1} & 1 \\
z_T & \vdots & \vdots & \vdots & \end{bmatrix}
\]
has a determinant of 1 so that

\[ p(z_1, z_2, \ldots, z_T | z_0, \ldots, z_{1-d}, \phi, \theta, \delta, \sigma^2) \]

\[ \propto |\Omega(\sigma, \phi, \theta)|^{-1/2} \exp\left[-\frac{1}{2} (z-\mu)' [\Omega(\sigma, \phi, \theta)]^{-1} (z-\mu) \right] \]

where

\[ z' = ((1-B)^d z_1, (1-B)^d z_2, \ldots, (1-B)^d z_T) \]
II. STANDARD APPROACHES TO THE MODEL

The standard approach to implementing an ARIMA model designed for forecasting involves four basic steps. First, the appropriate model is tentatively identified, that is, the following are determined: $p$, the order of the autoregressive portion of the model; $d$, the order of differencing necessary to achieve a stationary series; and $q$, the order of the moving average portion of the model.

The second step is the estimation of the parameters of the differenced series' model. Maximum likelihood estimates are usually advocated but because the system of partial derivatives is highly nonlinear in the model's parameters, multiple local maxima and minima may exist. Since explicit theoretical solutions are difficult to obtain, computer routines designed to find a local critical point of an approximation to the likelihood function are usually used.

The third step involves model verification, that is, diagnostic hypothesis testing to check the appropriateness of the fitted model. If a model is found inadequate at this state, another plausible model is entertained and the second and third steps repeated.

Finally, when a model has been verified and found to be adequate, it is used to forecast. We will now discuss each step in more detail, examining those procedures most widely advocated.
A. Choice of $p$, $d$, and $q$

There are two pictorial techniques commonly used to informally test for the stationarity of a series or the stationarity of the $d$th differences of the series.

The first procedure involves the inspection of the graphs of the original series and plausible differences of the series. If one of the plotted series appears to fluctuate around a fixed level through time then one may suspect stationarity since for a stationary process, the mean of the series is constant through time.

The second procedure entails inspection of the estimated autocorrelation function defined as the sequence $r_0, r_1, r_2, \ldots$ where

$$r_k = \frac{\sum_{t=1}^{T-k} (z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{T} (z_t - \bar{z})^2}$$

for $k = 0, 1, 2, \ldots$, for a series $z_t$, suspected of being stationary. A failure of the estimated autocorrelation function to die out quickly is taken as an indication that the underlying process should be treated as non-stationary.\(^\text{13}\)

In practice, $d$ is usually less than or equal to 2.

Having decided on an appropriate value for $d$, one works with the differenced series, $w_t = (1-B)^d z_t$, $t = d+1, \ldots, N$ to identify the model's autoregressive component of order $p$ and moving average component.

\(^{13}\)See Box and Jenkins [4] for a discussion of why one would expect this behavior in the autocorrelation function when the process is nonstationary.
of order q.

The two statistics used to identify p and q are the sample autocorrelation function (SAF) just defined, and the sample partial autocorrelation function (SPAF). To define the SPAF, let $\hat{\phi}_{kk}$ be the estimated coefficient of $w_{t-k}$ having fitted an autoregressive process of order k by ordinary least squares to the series $w_t$. By fitting autoregressive processes of orders 1, 2, 3, and so forth, using ordinary least squares, the estimated coefficients, $\hat{\phi}_{11}, \hat{\phi}_{22}, \ldots$, comprise the SPAF.

The SPAF approximates the partial autocorrelation function defined by the sequence $\phi_{11}, \phi_{22}, \phi_{33}, \ldots$.

Where $\phi_{kk}$ is obtained by solving the Yule-Walker system

$$
\begin{bmatrix}
1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\
\rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\
\rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_{k-1} & \ldots & 1 \\
\end{bmatrix}
\begin{bmatrix}
\phi_{k1} \\
\phi_{k2} \\
\vdots \\
\phi_{kk} \\
\end{bmatrix}
\begin{bmatrix}
\rho_1 \\
\rho_2 \\
\vdots \\
\rho_k \\
\end{bmatrix}
$$

for the $\phi_{kj}$'s in terms of $\rho_1, \rho_2, \ldots, \rho_k$. For autoregressive processes of order p, $\phi_{kk}$ is nonzero only for $k \leq p$, but since the inverted form of either purely moving average or mixed schemes is infinite in extent, the partial autocorrelation function for both processes will consist of an infinite convergent series of nonzero points, $\phi_{kk}$ ($k = 1, 2, \ldots$).

Since the SAF and SPAF tend to mimic the true autocorrelation and partial autocorrelation functions, respectively, the general appearance
of the estimators should provide clues about the appropriate choice of \( p \) and \( q \). Table 1 summarizes the likely behavior of the different functions for autoregressive, moving average, and mixed schemes. As several authors point out, a suitable choice is an art rather than a science; and linking SAF's and SPAF's to the correct specification of an ARMA process typically becomes easier through experience with actual constructions.

For a pure moving average process of order \( q \), there exists a hypothesis test due to Bartlett [2] and Anderson [1] which allows one to test whether the SAF is compatible with the hypothesis that the underlying process follows a moving average scheme of order \( q \). Bartlett derived the following approximate expression for the variance of the estimated autocorrelation coefficient for stationary normal processes:

\[
\text{Var}(r_k) = \frac{1}{T} \sum_{v=-\infty}^{\infty} \left( \rho_v^2 + \rho_{v+k}^2 - 4 \rho_v \rho_{v-k} + 2 \rho_v^2 \rho_k^2 \right)
\]

which reduces to \( \text{Var}[r_k] = \frac{1}{T} \left( 1 + 2 \sum_{v=1}^{q} \rho_v^2 \right) \) when the underlying process is moving average of order \( q \).

When \( r_k = 0 \), that is, for \( k > q \), Anderson has shown that

\[
\frac{r_k}{\sqrt{\text{Var}[r_k]}} \quad \text{is approximately, distributed as a standard normal}
\]

\[14\] The table is taken from Box and Jenkins [4, p. 79].

\[15\] Approximate in the sense that terms other than the dominant first term in the expansion of the exact result in powers of \( (1/T) \) are omitted.

\[16\] Approximate in the sense that the 5 and 1 percent significance points are essentially those of the standard normal to 2 decimal places.
<table>
<thead>
<tr>
<th>Table 1. Types of processes</th>
<th>Autoregressive Processes</th>
<th>Moving Average Processes</th>
<th>Mixed Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverted form of the model</td>
<td>$\phi(B)w_t = a_t$</td>
<td>$\theta^{-1}(B)w_t = a_t$</td>
<td>$\theta^{-1}(B)\phi(B)w_t = a_t$</td>
</tr>
<tr>
<td>Random shock form of the model</td>
<td>$w_t = \phi^{-1}(B)a_t$</td>
<td>$w_t = \theta(B)a_t$</td>
<td>$w_t = \phi^{-1}(B)\theta(B)a_t$</td>
</tr>
<tr>
<td>$\Pi$ weights</td>
<td>Finite series</td>
<td>Infinite series</td>
<td>Infinite series</td>
</tr>
<tr>
<td>$\Psi$ weights</td>
<td>Infinite series</td>
<td>Finite series</td>
<td>Infinite series</td>
</tr>
<tr>
<td>Stationarity condition</td>
<td>Roots of $\phi(B)=0$ lie outside the unit circle</td>
<td>Always stationary</td>
<td>Roots of $\phi(B)=0$ lie outside the unit circle</td>
</tr>
<tr>
<td>Invertibility condition</td>
<td>Always invertible</td>
<td>Roots of $\theta(B)=0$ lie outside the unit circle</td>
<td>Roots of $\phi(B)=0$ lie outside the unit circle</td>
</tr>
<tr>
<td>Autocorrelation function</td>
<td>Infinite (damped exponentials and/or damped sine waves) tails off</td>
<td>Finite cuts off</td>
<td>Infinite (damped exponentials and/or sine waves after first $q-p$ lags) tails off</td>
</tr>
<tr>
<td>Partial autocorrelation function</td>
<td>Finite cuts off</td>
<td>Infinite (dominated by damped exponentials and/or sine waves) tails off</td>
<td>Infinite (damped exponentials and/or sine waves after first $q-p$ lags) tails off</td>
</tr>
</tbody>
</table>
deviate.

With this result, one can test whether the SAF is compatible with the assumption that the process is moving average of order q.

If by inspection of the SAF and SPAF, an autoregressive scheme of order p is suspected, there exists a hypothesis test as to whether the theoretical partial autocorrelation coefficients beyond lag p are zero.

Quenouille [15] showed that if the underlying process is autoregressive of order p, the estimates of the partial autocorrelations are approximately independently distributed with

\[ \text{Var}(\hat{\phi}_{kk}) \approx 1/T \text{ for } k \geq p+1 \]

Box and Jenkins claim that the estimates are also approximately normal so that

\[ T^{1/2} \hat{\phi}_{kk} \sim N(0,1) \]

Thus, one can evaluate whether the SPAF agrees with the hypothesis that the process is autoregressive of order p.

B. Estimation and Verification

The usual estimates of the parameters of an ARMA process are approximations to the maximum likelihood estimates where the likelihood function is not conditional on previous or unobserved values of the w's. Actually, the estimates, called least squares estimates, do maximize a certain function but one that is an approximation to the

\[ \text{The bias in the approximation is of order } (1/T). \]
unconditional likelihood. There are several ways usually advocated of approximating the likelihood, and typically the different approximations result in different estimates. The exact unconditional likelihood function for mixed processes has already been derived in Section II. Here we will simply present the various approximations.

Recall, an ARMA (p,q) process can be written as

$$a_t = w_t - \phi_1 w_{t-1} - \ldots - \phi_p w_{t-p} - \delta + \phi_1 a_{t-1} + \ldots + \phi_q a_{t-q}$$  \hspace{1cm} (11)

Assuming the $a_t$ are $NID(0, \sigma^2)$, the density for $w_1, \ldots, w_T$ conditional on the parameters $\phi$, $\theta$, $\sigma^2$, $\delta$, $w_{1-p}, \ldots, w_0$, and $a_{1-q}, \ldots, a_0$ is given by

$$p(w_1, \ldots, w_T | \sigma^2, \delta, \phi, \theta, a_{1-q}, \ldots, a_0, w_{1-p}, \ldots, w_0) \propto \frac{1}{\sigma^T} \exp\left[- \frac{1}{2 \sigma^2} S_*(\phi, \theta, \delta)\right]$$

where

$$S_*(\phi, \theta, \delta) = \sum_{t=1}^T (w_t - \phi_1 w_{t-1} - \ldots - \phi_p w_{t-p} - \delta + \theta_1 a_{t-1} + \ldots + \theta_q a_{t-q})^2$$  \hspace{1cm} (12)

$S_*(\phi, \theta, \delta)$ is called a "conditional" sum of squares function. One approximation to the unconditional likelihood is given by (12) where $a_t (t=1-q, \ldots 0)$ is set equal to zero (its unconditional expectation) and $w_t (t=1-p, \ldots 0)$ is set equal to $\bar{w}$. The values of $\phi$, $\theta$, and $\delta$ which minimize (13) are taken as the chosen estimates. Representing estimates as $\hat{\phi}$, $\hat{\theta}$, $\hat{\delta}$, and $\hat{\sigma}^2$, $\hat{\sigma}^2$ is given by

$$S_*(\hat{\phi}, \hat{\theta}, \hat{\delta}) / T$$  \hspace{1cm} (14)

Another approximation is given by (12) with $a_{1-q}, \ldots, a_p$ set equal to zero and actually occurring values of the $w$'s used throughout. Using this
second approximation, the summation in $S_x(\phi, \theta, \delta)$ extends from $p+1$ to $T$.
Again, the values of $\phi$, $\theta$, and $\delta$ which minimize this approximation to the unconditional sum of squares function are chosen and $\sigma^2$ is again given by (14).

We will present what is known in the literature as the "unconditional likelihood" for an ARMA $(p,q)$ process. It is unconditional on previous values of the $w$'s and $a$'s in the usual probability sense.

However, the advocated approximation to this likelihood is conditional on "back forecasted" predictions of the $w$'s and $a$'s. The back forecasts are discussed in the Appendix.

The derivation of the unconditional likelihood begins by writing an ARMA $(p,q)$ process as a finite moving average of some order $Q$. For autoregressive or mixed processes the proper choice of $Q$ is questionable. A guideline is suggested by the fact that observations which evolve from a finite moving average process of order $Q$ are uncorrelated at lags greater than $Q$. Here the unconditional likelihood is presented for finite MA($Q$) processes.\(^\text{18}\) We use $\Gamma(B)$, a $Q$th order polynomial in $B$ to represent the approximation to $\Psi(B)$ which may be infinite. Only if $\Gamma(B)=\Theta(B)$ when the underlying model is MA($q$) will the likelihood be exact rather than an approximation. We assume that the ARMA $(p,q)$ process can be approximately written as

$$\tilde{w}_t = w_t - E\frac{w_t}{a_t} + \Gamma_1 a_{t-1} + \Gamma_2 a_{t-2} + \ldots + \Gamma_Q a_{t-Q} \quad (t=1,2,\ldots,T) \quad (15)$$

\(^{18}\)For a detailed derivation, see Box and Jenkins [4].
where $\Gamma(B)$ equals the first $Q$ terms in the series $\phi^{-1}(B)\theta(B)$. If we now proceed as if the approximation is exact, assuming the process is invertible and that $p(a_{1-Q}, \ldots, a_{T} | \Gamma, \delta, \sigma) \sim N(0, \sigma^2)$ we can derive $p(\tilde{w}_1, \ldots, \tilde{w}_T | \Gamma, \delta, \sigma)$. The final result is the following:

$$p(\tilde{w} | \sigma, \delta, \Gamma) = \frac{|X'X|^{-1/2}}{\sigma^T} \exp\left[-\frac{1}{2\sigma^2}((X\hat{\alpha} + (\tilde{w})'(X\hat{\alpha} + L\tilde{w}))ight]$$

$$= \frac{|X'X|^{-1/2}}{\sigma^T} \exp\left[-\frac{1}{2\sigma^2}S(\delta, \Gamma)\right]$$

where

$$\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_T)'$$

$$X_{(T+Q)\times Q} = \begin{bmatrix} I \\ -A_{22} -1 \end{bmatrix}$$

$$L_{(T+Q)\times T} = \begin{bmatrix} 0 \\ A_{22} -1 \end{bmatrix}$$

$$\hat{\alpha} = -(X'X)^{-1}X'L\tilde{w}$$

$$A_{21} = \begin{bmatrix} \Gamma_Q & \Gamma_{Q-1} & \cdots & \Gamma_1 \\ \Gamma_Q & \cdots & \Gamma_2 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \Gamma_Q & 0 \end{bmatrix}$$
The last expression is the unconditional likelihood for the model's parameters. $S(\delta, \Gamma)$ involves only $\delta$, $\phi$, $\theta$, and observed values of the $w$'s. The appropriate estimates are taken again as the least squares estimates, that is, those values of $\delta$, $\phi$, and $\theta$ minimizing $S(\delta, \Gamma)$. In practice, however, another form of $S(\delta, \Gamma)$ which is more convenient computationally is used. The alternative form for $S(\delta, \Gamma)$ is

$$S(\delta, \Gamma) = \sum_{t=1-Q}^{T} (E[a_t|\bar{w}, \sigma, \Gamma])^2 = \sum_{t=1-Q}^{T} [a_t]^2$$  \hspace{1cm} (16)$$

Taking expectations in (15),

$$[a_t] = [w_t] - \delta - \Gamma_1 [a_{t-1}] - \Gamma_2 [a_{t-2}] - \ldots - \Gamma_Q [a_{t-Q}]$$ \hspace{1cm} (17)$$

where

$$[w_t] = w_t \text{ for } t = 1, 2, \ldots, T.$$ 

Given calculated back forecasts of $[w_{1-Q}], \ldots, [w_0], [a_{1-2Q}], \ldots, [a_0]$, and values for the parameters, the system (17) can be used recursively to compute an approximation to (16). For preliminary back-
forecasted estimates of unobserved quantities, and preliminary estimates of $\delta$, $\phi$, $\theta$, and $\sigma^2$, see the Appendix.

Analogous to approximating any ARMA $(p,q)$ model by an MA of order $Q$ and deriving the likelihood of the approximating MA($Q$) process, one can approximate an ARMA($p,q$) as an AR of order $p$ and use the density for purely autoregressive schemes given in Section I.

Since $S(\delta, \Gamma)$ is nonlinear in $\phi$, $\theta$, and $\delta$, a nonlinear estimation routine is used to obtain least squares estimates $\hat{\delta}$, $\hat{\phi}$, and $\hat{\theta}$. Most routines of this nature are based on a Taylor series approximation of $S(\delta, \Gamma)$ about an initial guess for the parameter values.

A procedure known as Marquardt's iterative procedure was used to obtain the sampling theory least squares estimates for the entertained models of the empirical portion of this paper.$^{19}$

Having estimated the tentatively chosen model, diagnostic checks can be applied to the model in order to check its adequacy. A typical diagnostic technique called "overfitting" entails fitting a more elaborate model which includes parameters covering suspected areas of inadequacy. A hypotheses test is then applied to the additional parameters to see whether they are essentially zero.$^{20}$ Other diagnostic tests involve the analysis of the estimated residuals,

---

$^{19}$ For an explanation of the algorithm, the interested reader is referred to D. W. Marquardt [11].

$^{20}$ For details, see Nelson [13].
\[ \hat{a}_t = \vartheta(B) \hat{\varphi}(B) \hat{\gamma}_t \quad (t = 1, 2, \ldots, T). \]

C. Forecasting

In the analysis that follows, only unobservable quantities beyond time \( t \) are treated as random variables. Denote the forecast of \( \omega_t \) for period \( t + \ell \) and based on all the \( \omega \)'s (or equivalently, all the \( a \)'s) up to time \( t \) as \( \hat{\omega}_t(\ell) \). \( \hat{\omega}_t(\ell) \) shall be called the origin-\( t \) forecast for lead time \( \ell \).

Suppose \( \omega_t \) follows an ARMA process (the parameters of the model known exactly). Using the random shock form of the model, let

\[ \omega_{t+\ell} = \varphi^{-1}(B) \theta(B) a_{t+\ell} \]

\[ = \Psi(B) a_{t+\ell} \quad \text{(where } \Psi_0 = 1) \]

\[ = (a_{t+\ell} + \psi_1 a_{t+\ell-1} + \ldots + \psi_{\ell-1} a_{t+1}) + \sum_{j=0}^{\infty} \psi_{\ell+j} a_{t-j} \]

\[ = \epsilon_t(\ell) + \hat{\omega}_t(\ell) \]

Taking expectations in (19) conditional on past \( \omega \)'s, we have

\[ \hat{\omega}_t(\ell) = \epsilon_t(\ell) \]

\[ \text{for three different diagnostic tests dealing with the residuals, see Box and Jenkins [4].} \]
\[
E(w_{t+\ell} \mid w_t, w_{t-1}, \ldots) = E(e_t(\ell) \mid w_t, w_{t-1}, \ldots) + E(\hat{\omega}_t(\ell) / w_t, w_{t-1}, \ldots)
\]

\[
= 0 + E(\hat{\omega}_t(\ell) \mid w_t, w_{t-1}, \ldots)
\]
since by assumption, unobserved shocks are independent of the past

\[
\omega_t(\ell)
\]
since \(\hat{\omega}_t(\ell)\) is determined by \(w_t, w_{t-1}, \ldots\).

Hence, in this context, \(\hat{\omega}_t(\ell)\) minimizes the expected mean square forecast error given by

\[
E(e_t(\ell) \mid w_t, w_{t-1}, \ldots) = E(w_{t+\ell} - \hat{\omega}_t(\ell))^2
\]

\[
= E(\hat{\omega}_t(\ell) - w_{t+\ell})^2
\]

\[
= \text{Var}[e_t(\ell)]
\]

\[
= (1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_{\ell-1}^2) \sigma^2
\]

(20)

In practice, to compute \(\hat{\omega}_t(\ell) = \sum_{j=0}^{\infty} \psi_{\ell+j} a_{t-j}\), the \(a_t\) are replaced by the estimated errors, \(\hat{a}_t\) (which for moderate degrees of accuracy are negligible beyond some point), that is, the residuals from the estimated equation, and the \(\psi_{\ell+j}\) are replaced by their estimates \(\hat{\psi}_{\ell+j}\) after fitting the model.

A more conventional and common method for computing the forecasts is to use the difference equation form of the process. The computation of \(\hat{\omega}_t(\ell)\) begins by finding

\[
\hat{\omega}_t(1) = \hat{\phi}_1 w_t + \ldots + \hat{\phi}_p w_{t-p+1} + \hat{\delta} \hat{a}_t - \ldots - \hat{\delta} \hat{a}_{t-q+1}
\]

Since \(a_{t+1}\) is independent of \(w_t, w_{t-1}, \ldots\), it is set to its expectation; zero. Also for \(t<0\), \(w_t\) is taken to be the backforecasted value,
and similarly for $\hat{a}_t$. Next $\hat{\omega}_t(2)$ is obtained via the equation

$$\hat{\omega}_t(2) = \hat{\phi}_1 \hat{\omega}_t(1) + \ldots + \hat{\phi}_p \hat{\omega}_{t-p+2} + \hat{\delta} \hat{a}_t \ldots - \hat{\theta}_q \hat{a}_{t-q+2}$$

where again, $a_{t+1}$ and $a_{t+2}$ are set to zero. Continuing in this manner, we arrive at

$$\hat{\omega}_t(\ell) = \hat{\phi}_1 \hat{\omega}_t(\ell-1) + \ldots + \hat{\phi}_p \hat{\omega}_{t-p+\ell} + \hat{\delta} \hat{a}_t \ldots - \hat{\theta}_q \hat{a}_{t-q+\ell}$$

For $q < \ell < p$, the forecast involves only an autoregressive portion. For $\ell > \max[p,q]$ we have

$$\hat{\omega}_t(\ell) = \hat{\phi}_1 \hat{\omega}_t(\ell-1) + \ldots + \hat{\phi}_p \hat{\omega}_{t-p}$$

From (20) we see that the variance of the forecast error is given by

$$(1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_\ell^2) \sigma^2$$

when the weights are exactly known. But in practice, since the $\psi_k$ are replaced by their estimates which are themselves random variables, the variance of the forecast error will in general be larger. The variances and the bias$^{22}$ of the estimates $\hat{\psi}_k$ when the approximating nonlinear estimation routine is used are difficult to determine.

Although the Bayesian prescription for the computation of forecasts is considerably more complicated, the evaluation of the variance of forecasts is accomplished straightforwardly by numerical integration routines.

---

$^{22}$The estimates of $\psi_k$ are typically obtained via $\hat{\Psi}(B) = \hat{\phi}^{-1}(B)\hat{\Theta}(B)$. Since the $\psi_k$'s are in general nonlinear functions of the $\hat{\theta}_i$ and $\hat{\phi}_j$, they will usually be biased even if $\hat{\theta}$ and $\hat{\phi}$ are unbiased for $\theta$ and $\phi$ respectively.
III. POST-BAYESIAN APPROACH: IDENTIFYING THE TRADE-OFFS

A. Choice of Methods

As presented here, both the Bayesian and Orthodox approaches utilize assumptions simplifying analysis of the model but neither approach attempts to measure the implicit trade-offs occurring between the level of complexity and accuracy of analysis.

Two points on which the Bayesian analysis was simplified were first, proceeding throughout the analyses on specified values of p, d, and q, that is, p, d, and q were treated as known constants rather than as random variables similar to \( \phi, \theta, \delta, \) and \( \sigma \); and second, the choice of a uniform prior over the parameters of interest rather than one which reflects most accurately prior personal beliefs.

Several points on which the Orthodox analysis was simplified are found in: the use of pictorial techniques and hypothesis tests in identifying p, d, and q; the use of a finite moving average process

---

23 Although more complex computationally, the approach of treating the parameters, p, d, and q as random variables is more in the Bayesian spirit. Ideally, by taking a suitable prior over \( \phi, \theta, \delta, \sigma^2, p, d, \) and q, the predictive density \( p(z_{t+1} | z_1, \ldots, z_t) \) could be implemented for forecasting without ever specifying particular values for the parameters which include in this instance, p, d, and q.

24 Significance tests (used not only in identifying p, d, and q, but at any stage of model analysis) can be viewed as crude procedures attempting to detect whether the gain in simplicity resulting from accepting the null hypothesis outweighs the inaccuracy resulting from accepting it. For a discussion of hypothesis tests in a Post-Bayesian framework, see Faden and Rausser [7].
to approximate autoregressive and mixed schemes; the computation of backforecasted values of unobservable quantities for the evaluation of the unconditional sum of squares function; the substitution of least squares estimates for maximum likelihood estimates; and the simple recursive calculation of the period $T+\ell$ forecast by substitution of earlier forecasts, $T+\ell-1$, $T+\ell-2$, ..., $T+1$, in the difference equation form of the model.

The problem of model specification and choice of procedures for analysis is actually one of balancing the inaccuracy of competing strategies against the complexities associated with each strategy. As more resources are invested into an analysis, for example, man hours of designing and learning an increasingly complex model, available computer use funds, man hours of analyzing and checking the appropriateness of a proposed model, higher levels of model accuracy can be obtained. And as accuracy increases, the losses associated with implementing the model in the formulation of policy decisions decline. The strategy minimizing the sum of costs due to the level of complexity of an analysis and expected costs due to model inaccuracy is taken as optimal in the Post-Bayesian framework.

The set of possible competing strategies for choosing an appropriate specification and implementing an ARIMA $(p,d,q)$ model is extremely large. An attempt to measure the complexity-inaccuracy costs associated with each would be extremely time demanding and costly. Even if accomplished, the results could not prescribe an absolute rule for all investigators because of the subjective nature of portions of both Bayesian and Orthodox
procedures.25

As mentioned earlier, the decision to choose a model from the class of ARIMA models is frequently made for the sake of model simplicity. That is, even though one has the choice of investigating more elaborate models such as large simultaneous equations models, the search for an appropriate model is limited to the class of ARIMA models because the gain in model simplicity (and the associated costs of analysis) outweighs an expected possible loss in model inaccuracy.

As the complexity of model analysis increases one would expect a gain in model accuracy. If, however, in striving to achieve higher levels of accuracy, additional resources allocated to analyzing a model are misallocated, inaccuracy may remain unchanged or possibly rise.26

Hypothetically, for each strategy of analysis designed to describe a variable, z, there is an associated level of complexity (which often can be directly translated into a pecuniary cost) and also, an expected level of inaccuracy. For simplicity, assume complexity, \( C \), is measured in dollars and expected inaccuracy, \( I_a \), is measured by a conventional statistic such as

\[ I_a = \frac{C}{\text{expected inaccuracy}} \]

25 For example, investigators will typically differ on the interpretation and prescriptions of sample correlograms. Also, it is unlikely that different investigators will reflect identical personal probabilities in their priors used in a rigorous Bayesian analysis.

26 For an example in reference to time series analysis, see Box and Jenkins [4, pp. 248-250].
\[ \sum_{t=1}^{T} \frac{1}{T} (z_t - \hat{z}_t)^2 \text{ or } \left( \sum_{t=1}^{T} \frac{(z_t - \hat{z}_t)^2}{\sum z^2} \right)^{1/2}. \]

where \( z_t \) is the actual value at time \( t \) and \( \hat{z}_t \) is the predicted value.

Let \( S \) represent the set of all possible strategies and consider representing each strategy, say \( S_i \), (\( i = 1, 2, \ldots \)) by its associated components, \( (C_i, I_i) \) in the complexity-inaccuracy plane (see Figure 1).

The "efficiency frontier" is the set of points reflecting the lowest attainable level for \( C \) given a level of \( I \). It is the inner envelope curve of the set of all possible strategies.

Minimizing the sum of expected complexity and inaccuracy costs over the set \( S \) implies that an investigator choose a strategy on the efficiency frontier. Figure 2 graphically depicts the appropriate solutions.

As an example, when expected total costs,

\[ T = C + a.I_a \] (Case 1)

at the point of optimality the isocline is given by linear segment, \( T_1D_1 \). It has slope, \(-a\), and its intercept, \( T_1 \), is equal to the minimal expected total costs of the analysis. As the cost of anticipated inaccuracy decreases, that is, as the slope of the isocline increases, the optimal strategy and the associated expected complexity-inaccuracy mix will change; an optimal strategy will be less complex as well as less accurate, but with a lower total cost.

Suppose more generally that expected total costs are given by

\[ T = C + a.I_a^2 + b.I_a \] (Case 2)
Figure 1. Alternative strategies

Figure 2. Optimal choices
the isocline is given by the curved segment, $T_2D_2$.

The optimal strategy is located at the point of tangency \(^{27}\) between the isocline and the efficiency frontier, at the point corresponding to $S_2$. The point $T_2$ measures total minimum costs when $S_2$ is implemented.

Similar to the result earlier, if either $a$ or $b$ decrease, reflecting a decrease in the losses occurring from expected model inaccuracy, the slope of the isocline will increase, causing tangency to occur at a point of lower expected complexity and higher expected inaccuracy but with a smaller total cost.

B. Inaccuracy and Complexity Costs \(^{28}\)

The degree of inaccuracy in a model or analysis of a model is inversely related to the benefits arising from employing the model in the formulation of policy decisions. This is because such decisions are based at least in part on the premise that model forecasts are accurate. Forecasts, or more generally, model inaccuracy, may result in costly ineffective policies.

For this reason, the more important the problem which the model is designed to investigate, the greater the level of complexity which is

---

\(^{27}\) When the efficiency frontier is not smooth, the tangency condition does not apply.

\(^{28}\) The treatment of complexity here should be distinguished from another different but related approach to complexity (familiar to computer scientists) which attempts to measure the least number of steps required (or computer resources required) to solve various mathematical, combinatorial, and sorting problems, e.g. invert a matrix, evaluate a function of a given complexity, determine the limit of a sequence, order a set of family names lexicographically, etc. . . . For an extensive bibliography on this area of study see Savage [17].
Hypothetically, for policy on which a set of forecasts is based, the loss due to inaccuracy can be translated into a pecuniary cost. Although such an inaccuracy cost function might be difficult to measure, it is probably the most useful way of examining inaccuracy since it facilitates comparison with the second aspect of model analysis, complexity, which has corresponding costs: the initial man hour costs of formulating and learning the model, the computer use and man hour costs of fitting the model, the computer use and man hour costs of maintaining, interpreting, and communicating results of the model.

It will usually be true that as the inaccuracy of a model decreases, complexity costs, that is, the sum costs of various components of model complexity increase; and vice versa, that is, the accuracy of a model increases as complexity increases, but that beyond some very high level of complexity, additional returns in terms of accuracy decrease.

1. **Measuring inaccuracy and complexity**

Several familiar measures of model inaccuracy are given below where

\[
\begin{align*}
    w_t &= \text{actual value at time } t \\
    \hat{w}_t &= \text{value predicted from the model at time } t \\
    T &= \text{last period in which observations are used to fit the model}
\end{align*}
\]

---

29 As the model becomes extremely complex, human errors in terms of theoretically understanding, logically analyzing, and communicating the results of a model will cause accuracy to decline. Also, computer capabilities will constrain the level of model complexity.
Theil's U-Statistic given by

$$U = \left( \frac{1}{T} \sum_{t=1}^{T} (w_t - \hat{w}_t)^2 / \sum_{t=1}^{T} w_t^2 \right)^{1/2}$$

is one measure of model inaccuracy. $U = 0$ if and only if all forecasts are perfect. If the $w_t$ are the differences of a series, $z_t$, then $U = 1$ has the interpretation that the model does as well as the naive no-change extrapolation of $z_t$ (that is, forecasting according to the rule: $\hat{z}_t = z_{t-1}$). $U$ greater than (less than) unity means the model does worse than (better than) the naive model in terms of residual mean square error given by

$$\text{RMS} = \frac{1}{T} \sum_{t=1}^{T} (w_t - \hat{w}_t)^2$$

RMS is another conventional measure of model inaccuracy.

For forecasts 1, 2, ..., and $l$ periods beyond the sample period, mean absolute error defined by

$$\text{MAE} = \frac{1}{l} \sum_{i=1}^{l} |w_{T+1} - \hat{w}_{T+1}|$$

and the mean absolute percent error defined by

$$\text{MAPE} = \frac{1}{l} \sum_{i=1}^{l} \left| \frac{w_{T+1} - \hat{w}_{T+1}}{w_{T+1}} \right|$$

are two measures of model inaccuracy. RMS and Theil's U-Statistic can also be evaluated over the set of forecasts beyond time $T$.

For comparison with complexity costs, the level of model inaccuracy as measured by one of these statistics, must be transformed into the corresponding costs of specified levels of inaccuracy. Assuming the
model has been identified, easily measured components of model complexity are:

\[ M_s = \text{man hours of setting up the data and computer program for parameter estimation} \]

\[ C = \text{computer use costs of analysis} \]

\[ M_T = \text{man hours of interpreting and tabulating computer results} \]

\[ T = \text{number of observations in the data set} \]

\[(p,d,q) = \text{vector with components: } p, \text{ order of the AR portion of the model, } d, \text{ degree of differencing to achieve stationarity; and } q, \text{ order of the MA portion of the model.} \]

\[ M_s, C, \text{ and } M_T \text{ are each positively related to } T, p, d, \text{ and } q, \text{ but their functional forms should be expected to vary according to whether an Orthodox or Bayesian analysis is used. For simplicity, complexity costs will be measured as} \]

\[ C_i = h \cdot M_{s_i} + C_i + h \cdot M_{T_i} \]

\[ = h \cdot M_{s_i}(T,p,d,q) + C_i(T,p,d,q) + h \cdot M_{T_i}(T,p,d,q) \]

when

\[ i = 1 \text{ corresponds to using the Bayesian form of analysis} \]

and

\[ i = 2 \text{ corresponds to using the Orthodox form of analysis} \]

and

\[ h \text{ is the wage rate of the investigator.} \]

For a Bayesian form of analysis, the following functional form for \( C_i \),
although crude, is suggested.

\[ C_1 = hM_{s1}(T,p,d,q) + C_1(T,p,d,q) + hM_{T1}(T,p,d,q) \]

where

\[ M_{s1} = \beta_1 + \beta_2 T + \beta_3 (p+2q+1) \]

\[ C_1 = \beta_4 T + \beta_5 (p+2q) + \beta_6 T(p+2q+1) + \beta_7 T(p+2q+1)^2 \]

\[ M_{T1} = \beta_8 (p+2q)+1 \]

Wherever \( q \) appears, it is multiplied by the factor 2. This is done as an attempt to incorporate in a simple fashion the fact that moving average and mixed processes are more complicated than autoregressive processes.

Of \( M_{s1} \), the constant term \( \beta_1 \) reflects the time needed to write standard subroutines for the numerical computation of parameters of the marginal distributions, the time needed to write a program section transforming the original series to the \( d \)th differences of the series, and program debugging time. The term, \( \beta_2 T \) reflects the time needed to tabulate and check the data; and \( \beta_3 (p+2q+1) \) measures the time needed to write the \( p+q+1 \) integration routines and to write portions of the program which compute parameters of the \( p+q+1 \) marginals.

Of the four components of \( C_1 \), \( \beta_4 T \) measures the cost of compiling the data, printing predictions, and computing certain statistics \( \beta_5 (p+2q) \) reflects the cost of compiling the remainder of the program deck. \( \beta_6 T(p+2q+1) \) measures the cost of computing predictions and obtaining plots of the \( p+q+1 \) marginals. \( \beta_7 T(p+2q+1) \) measures the cost of performing the \( p+q+1 \) integrations.
Lastly, $M_{T_1} = \beta_8 (p+2q+1)$ reflects the time needed to read and interpret the results of the analysis.

If an Orthodox analysis is pursued, the following functional form for $C_2$ is suggested.

$$C_2 = hM_{n_2}(T,p,d,q) + C_2(T,p,d,q) + hM_{T_2}(T,p,d,q)$$

where

$$M_{s2} = \alpha_1 + \alpha_2 T$$

$$C_2 = \alpha_3 T(p+2q) + \alpha_4 T + \alpha_5 T^2(p+2q)^2$$

$$M_{T_2} = \alpha_6 (p+2q+1)$$

As before, $q$ is multiplied by a factor of 2 to reflect the fact that moving average and mixed processes are more difficult to work with than autoregressive processes.

As components of $M_{s2}$, $\alpha_1$ measures the time required to put standard packaged programs such as PDQ, ESTIMATE, and FORECAST (used in this study) on disk, and $\alpha_2 T$ reflects the time needed to tabulate and check the data.

As parts of $C_2$, $\alpha_3 T(p+2q)$ measures the computer costs of getting initial estimates of the parameters via the program PDQ; $\alpha_4 T$ reflects the cost of compiling the data; and $\alpha_5 T^2(p+2q)^2$ measures the cost of estimating the $p+q+1$ parameters.

$M_{T_2} = \alpha_6 (p+2q+1)$ measures the time spent reading and interpreting results.

Figure 3 illustrates graphically a possible relationship between forecasting costs, expected inaccuracy and complexity costs for either $C_2$.
Costs Associated with Forecasting

Figure 3. Relationship between complexity and inaccuracy
or $C_2$. (The indicated optimal inaccuracy level should not be taken literally since the cost of complexity is a multidimensional function.)
IV. EMPIRICAL ANALYSES

A. The Experiment and Results

The empirical portion of this paper is directed toward some of the various complexity and inaccuracy costs associated with the choice of Bayesian procedures or Orthodox procedures (both as presented earlier) as they apply to purely autoregressive processes. Purely moving average and mixed schemes were not examined since both the time allowed for the study and available computer funds were scarce. Also, the project of measuring complexity costs and inaccuracy costs for a wide range of alternative specifications of p and d when using either Bayesian or Orthodox estimation procedures was not attempted for similar reasons. The values of p and d for the time series examined were suggested by Nelson [13] who fitted ARIMA models to the series in an Orthodox framework.

The series and model specifications examined are presented in Table 2.\(^{30}\)

There were 3 computer programs used to perform an Orthodox analysis of each time series. The programs, called PDQ, ESTIMATE, and FORECAST were made available by Charles R. Nelson of the University of Chicago. They are written in FORTRAN IV for the IBM 360 and how they were utilized is briefly explained below.

Program PDQ provided estimates and plots of the sample autocorrelation and partial autocorrelation functions for user-specified degrees of

\(^{30}\)Unfortunately, during the time of this study, data on additional time series to be analyzed was accidentally thrown out by a night janitor.
Table 2. ARIMA models for the analyzed series

<table>
<thead>
<tr>
<th>Series Number</th>
<th>Seriesa</th>
<th>Hypothesized Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gross National Product (nominal dollars) 1948-01 - 1966-04</td>
<td>$(z_t - z_{t-1}) = \phi_1 (z_{t-1} - z_{t-2}) + \phi_2 (z_{t-2} - z_{t-3}) + \eta$</td>
</tr>
<tr>
<td>2</td>
<td>Yield on United States Treasury Bills 1948-01 - 1966-04</td>
<td>$(z_t - z_{t-1}) = \phi_1 (z_{t-1} - z_{t-2}) + \phi_2 (z_{t-2} - z_{t-3}) + \eta$</td>
</tr>
<tr>
<td>3</td>
<td>Yield on Corporate Bonds 1948-01 - 1966-04</td>
<td>$(z_t - z_{t-1}) = \phi_1 (z_{t-1} - z_{t-2}) + \phi_2 (z_{t-2} - z_{t-3}) + \eta$</td>
</tr>
<tr>
<td>4</td>
<td>Yield on Commercial Paper 1948-01 - 1966-04</td>
<td>$(z_t - z_{t-1}) = \phi_1 (z_{t-1} - z_{t-2}) + \phi_2 (z_{t-2} - z_{t-3}) + \eta$</td>
</tr>
</tbody>
</table>

aThe four series were taken from the data set used in the 1969 version of the Federal Reserve Board - MIT, quarterly econometric model of the U.S. economy.
differencing of each series.

Program ESTIMATE used Marquardt's iterative procedure to find estimates of each model's parameters which minimized the unconditional sum of squares function via backforecasting of presample observations. The user told the program exactly what the model looked like, what parameters were to be estimated, and provided initial estimates of the parameters.

Program FORECAST provided predictions for the series over the sample period and predictions of the series for ten periods beyond the horizon using the recursive calculation procedure discussed in Section II.

Three programs written by the author in FORTRAN IV for the IBM 360 were utilized for the Bayesian analysis of the examined series. The posterior marginal density function for each autoregressive parameter was evaluated and a plot of each density provided. The mean and variance of each autoregressive parameter was also evaluated numerically. Predictions of the series over the sample period as well as for ten periods beyond the sample horizon were evaluated. 31

For the first order autoregressive process,

\[ w_t = \delta + \phi_1 w_{t-1} + a_t \]

the forecast of \( w_{T+\ell} \) conditional only on \( w_1, w_2, \ldots, w_T \) was evaluated by first solving the difference equation above recursively. Doing this yields

31 Copies of PDQ, ESTIMATE, and FORECAST in machine language (deck form), and copies of all other programs in FORTRAN IV can be obtained from the author on request.
\[ w_{t+\ell} = \sum_{i=1}^{\ell} \delta \phi_1^{\ell-1} + \phi_1^\ell w_t + \sum_{i=1}^{\ell} \phi_1^{\ell-1} a_{t+i} \]

Taking expectations conditional on \( w_1, w_2, \ldots, w_T \) (with \( t=T \)) we have

\[
\hat{w}_{T+\ell} = E(w_{T+\ell} | w_1, w_2, \ldots, w_T) = \sum_{i=1}^{\ell} E(\delta \phi_1^{\ell-1} | w_1, w_2, \ldots, w_T) + E(\phi_1^\ell | w_1, w_2, \ldots, w_T) w_T
\]

For the second order process,

\[ w_t = \phi_1^0 w_{t-1} + \phi_2^0 w_{t-2} + a_t \]

the forecasts of \( w_{T+\ell} (\ell=1, 2, \ldots, 10) \) were obtained by first solving the second order difference equation with characteristic equation

\[ x^2 - \phi_1 x - \phi_2 = 0 \]

for the latent roots,

\[ x_1 = \frac{1}{2}(\phi_1 + (\phi_1^2 + 4\phi_2)^{1/2}), \quad x_2 = \frac{1}{2}(\phi_1 - (\phi_1^2 + 4\phi_2)^{1/2}) \]

This yields the result,

\[ w_{t+\ell} = \frac{1}{2} c_1 (\phi_1 + (\phi_1^2 + 4\phi_2)^{1/2}) \hat{w}_t + \frac{1}{2} c_2 (\phi_1 - (\phi_1^2 + 4\phi_2)^{1/2}) \hat{w}_{t-1} + \text{(terms involving } \phi_1, \phi_2, a_t, a_{t+1}, \ldots, a_{t+\ell}) \]

Taking expectations conditional on \( w_1, w_2, \ldots, w_T \) (with \( t=T \)) we have
\[
\hat{\theta}_{T+\ell} = E(w_{T+\ell} | w_1, w_2, \ldots, w_T)
\]
\[
= \frac{1}{2} c_1 E((\phi_1 + (p_1^2 + 4\phi_2)^{1/2}) | w_1, w_2, \ldots, w_T) w_T
\]
\[
+ \frac{1}{2} c_2 E((\phi_1 - (p_1^2 + 4\phi_2)^{1/2}) | w_1, w_2, \ldots, w_T) w_T
\]

with \(c_1\) and \(c_2\) determined by \(w_T\) and \(w_{T-1}\).

The ARIMA models specified in Table 2 were fitted to the series first for the period 1948-01 through 1966-04 (that is, \(T=76\)) and second, for the period 1948-01 through 1956-02 (\(T=38\)). The length of the series was varied in order to compare the changes in complexity and inaccuracy costs associated with each approach.

In Table 3, components of complexity costs were tabulated for each analysis. For series 1 and 2, the hours of setting up the programs, \(M_s\), and the hours of reading and interpreting the results, \(M_a\), are large relative to those of series 3 and 4. This is so because the time spent on series 1 and 2 reflects writing and preparing the necessary programs. The large discrepancy between the time spent preparing the Bayesian analysis and time spent preparing the Orthodox analysis occurs because packaged programs performing an Orthodox analysis were available, as mentioned earlier.

It should be noted that if a large enough number of series (which included moving average and mixed processes in addition to autoregressive schemes) had been analyzed, the parameters of the complexity cost functions could actually be estimated by regression analysis.

In Table 4 selected measures of model inaccuracy over the sample
Table 3. Approximate complexity costs\(^{a}\)

<table>
<thead>
<tr>
<th>Measures of Complexity</th>
<th>Series 1 Bayes</th>
<th>Orthodox</th>
<th>Series 2 Bayes</th>
<th>Orthodox</th>
<th>Series 3 Bayes</th>
<th>Orthodox</th>
<th>Series 4 Bayes</th>
<th>Orthodox</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M_s)</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>(C)</td>
<td>3.35</td>
<td>2.19</td>
<td>4.22</td>
<td>2.22</td>
<td>4.21</td>
<td>2.20</td>
<td>4.20</td>
<td>2.22</td>
</tr>
<tr>
<td>(M_A)</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>(= h \cdot M_s + C + h \cdot M_A)</td>
<td>8.35</td>
<td>7.19</td>
<td>9.22</td>
<td>7.22</td>
<td>9.21</td>
<td>7.20</td>
<td>9.20</td>
<td>7.22</td>
</tr>
<tr>
<td>(T = 38)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M_s)</td>
<td>12.5</td>
<td>3.5</td>
<td>15.5</td>
<td>.5</td>
<td>.5</td>
<td>1.0</td>
<td>.5</td>
<td>1.0</td>
</tr>
<tr>
<td>(C)</td>
<td>4.68</td>
<td>2.32</td>
<td>5.71</td>
<td>2.39</td>
<td>5.55</td>
<td>2.37</td>
<td>5.96</td>
<td>2.41</td>
</tr>
<tr>
<td>(M_A)</td>
<td>5.0</td>
<td>3.0</td>
<td>6.0</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>(= h \cdot M_s + C + h \cdot M_A)</td>
<td>92.18</td>
<td>34.82</td>
<td>113.21</td>
<td>7.89</td>
<td>10.55</td>
<td>9.87</td>
<td>10.96</td>
<td>9.91</td>
</tr>
<tr>
<td>(T = 76)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\)Computed with \(h = \$5.00\) per hour.
Table 4. Inaccuracy over the sample period

<table>
<thead>
<tr>
<th>Measures of Inaccuracy</th>
<th>Series 1 Bayes Orthodoxy</th>
<th>Series 2 Bayes Orthodoxy</th>
<th>Series 3 Bayes Orthodox</th>
<th>Series 4 Bayes Orthodox</th>
</tr>
</thead>
<tbody>
<tr>
<td>T=38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theil's U-Statistic</td>
<td>0.5748167</td>
<td>0.5749841</td>
<td>0.8101426</td>
<td>0.8099263</td>
</tr>
<tr>
<td></td>
<td>0.8176919</td>
<td>0.9060053</td>
<td>0.7110391</td>
<td>0.7104837</td>
</tr>
<tr>
<td>Residual Mean Square</td>
<td>18.28886</td>
<td>18.29951</td>
<td>0.02569438</td>
<td>0.02568066</td>
</tr>
<tr>
<td></td>
<td>0.01296787</td>
<td>0.07684860</td>
<td>0.02245675</td>
<td>0.02242168</td>
</tr>
<tr>
<td>One step ahead absolute</td>
<td>0.13</td>
<td>1.1009576</td>
<td>1.114215</td>
<td>1.121013</td>
</tr>
<tr>
<td>percent error</td>
<td>1.870337</td>
<td>1.106644</td>
<td>0.9395327</td>
<td>0.9455124</td>
</tr>
<tr>
<td>T=76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theil's U-Statistic</td>
<td>0.520802</td>
<td>0.5189157</td>
<td>0.8225171</td>
<td>0.820387</td>
</tr>
<tr>
<td></td>
<td>0.8987366</td>
<td>0.8979356</td>
<td>0.7805080</td>
<td>0.7790498</td>
</tr>
<tr>
<td>Residual Mean Square</td>
<td>23.23578</td>
<td>23.06776</td>
<td>0.0818253</td>
<td>0.08140194</td>
</tr>
<tr>
<td></td>
<td>0.0117701</td>
<td>0.01174818</td>
<td>0.0683355</td>
<td>0.06808043</td>
</tr>
<tr>
<td>One step ahead absolute</td>
<td>0.4699805</td>
<td>0.54174</td>
<td>1.399147</td>
<td>1.409134</td>
</tr>
<tr>
<td>percent error</td>
<td>1.51088</td>
<td>1.472269</td>
<td>1.157963</td>
<td>1.118021</td>
</tr>
</tbody>
</table>
period for the alternative approaches are tabulated. The absolute percent error of the forecast one period beyond the sample is also provided. It seems surprising that the results for each mode of analysis are so similar. For \( T=76 \), the standard approach produces a better fit over the sample period for each of the four series examined. However, for \( T=38 \), the standard approach (when inaccuracy is measured by the sample residual mean square error (SRMS) and Bayesian approach are approximately equivalent since both show two cases of smaller SRMS.

At both \( T=38 \) and \( T=76 \), each approach has two cases of smaller one-step-ahead absolute percent error.

Again, if enough series had been analyzed, an estimate of expected inaccuracy for different types of series could be obtained by regression.

In Table 5 Theil's U-Statistic and the residual mean square error are measured for the ten forecasts beyond the sample horizon. For \( T=76 \), an Orthodox approach appears to provide better forecasts since it is less inaccurate for three of the four time series. But for \( T=38 \), the Bayesian approach produces both a smaller \( U \)-statistic and residual mean square error in three of the four cases examined.

It is interesting that the Bayesian forecasts are not as accurate as the Orthodox forecasts for the larger sample period. A possible reason could be the numerical rounding errors encountered when doing multivariate integrations on the computer.

The results over the extrapolated period for \( T=38 \) possibly suggest the dependence on large sample theory of Orthodox results.
<table>
<thead>
<tr>
<th>Measures of Inaccuracy</th>
<th>Series 1</th>
<th>Series 2</th>
<th>Series 3</th>
<th>Series 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bayes</td>
<td>Orthodox</td>
<td>Bayes</td>
<td>Orthodox</td>
</tr>
<tr>
<td>T=38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theil's U-Statistic</td>
<td>0.9527230</td>
<td>0.9009387</td>
<td>1.001341</td>
<td>0.9864404</td>
</tr>
<tr>
<td>Residual Mean Square</td>
<td>70.34528</td>
<td>49.71648</td>
<td>.5438256</td>
<td>.5448262</td>
</tr>
<tr>
<td>T=76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theil's U-Statistic</td>
<td>0.8943064</td>
<td>0.6025598</td>
<td>0.9702145</td>
<td>0.9394268</td>
</tr>
<tr>
<td>Residual Mean Square</td>
<td>228.4183</td>
<td>103.6952</td>
<td>.2311764</td>
<td>.2463773</td>
</tr>
</tbody>
</table>
B. Summary

On the basis of the experimental results, there is little difference in the sample period and forecast accuracy of both approaches. In fact, the closeness of accuracy is surprising. However, there is a substantial difference in the set up costs of each initial analysis (for example, Series 1 for T=76). And the computer use costs of subsequent analyses are consistently higher for a Bayesian analysis. Thus, whether or not computer programs designed to perform a Bayesian analysis are already available, the results obtained here would suggest using an Orthodox strategy when expected inaccuracy costs are not extremely large. If the loss from inaccuracy is large, it is unclear which approach is in order since results for the two methods are quite close. I feel more series would have to be examined in order to argue persuasively one way or another.

The treatment of the ARIMA model in a Post-Bayesian framework is by no means completed. Additional experimental analyses of alternative model specifications in the manner of this section would provide a sounder guideline for the actual analysis of a time series model.

"Seasonal" ARIMA models\(^ {32} \) (useful particularly in analyzing economic data) were not investigated in this thesis. Naturally, a Post-Bayesian study of ARIMA models could be made more general and applicable by this inclusion.

\(^ {32} \)See Box and Jenkins [4, Chapter 9].
Although extremely important, the problem of initially setting $p$, $d$, and $q$ to specific values has not been studied here or elsewhere in a Bayesian or Post-Bayesian setting, although one possible Bayesian solution has been suggested.\footnote{See footnote 21.}

One problem encountered, that of rounding errors associated with numerical multiple integration procedures, was witnessed also by a colleague while working on her research. This problem can most likely be corrected by the investment of much time and funds in writing improved integration routines.

If sufficient time and money were available, various Bayesian analyses could be made into packaged computer programs. I suspect this would result in an increase of their use (provided accurate results were obtained) since the initial set up costs of doing a bayesian analysis are extremely high.

Hopefully, what has been presented in this thesis has at least stressed the need and usefulness of approaching problems of inference in terms of expected complexity and inaccuracy.
V. REFERENCES


VI. ACKNOWLEDGMENTS

My deepest thanks go to Arnold M. Faden for enlightening sessions spent together. His capability to understand and communicate complex problems and their solutions, his patience, good humor, and encouragement were a tremendous help to me.

Thank you also, Steve, for your persistent encouragement, help at keeping me at my work, and your friendship and understanding.
VII. APPENDIX

A. Preliminary Estimates of the Parameters

Initial estimates of $\phi$, $\theta$, $\delta$, and $\sigma^2$ are usually required in non-linear estimation routines.

For purely autoregressive processes, estimates $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ of $\phi_1, \phi_2, \ldots, \phi_p$ respectively can be obtained via an approximation to the Yule-Walker equations (system (4a) given in Section I). Replacing $\rho_1, \rho_2, \ldots, \rho_p$ by the estimated autocorrelations $r_1, r_2, \ldots, r_p$, the $p$ by $p$ system can be solved for $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ in terms of $r_1, r_2, \ldots, r_p$.

The initial estimate, $\hat{\delta}_0$ of $\delta$ is taken as $\bar{w}(1-\hat{\phi}_1-\hat{\phi}_2-\ldots-\hat{\phi}_p)$ and the estimate $\hat{\sigma}^2_0$ of $\sigma^2$ is given by

$$T \left[ \sum_{t=1}^{T} (w_t - \bar{w})^2 (T-1) \right] [1-\hat{\phi}_1, r_1, \hat{\phi}_2, r_2, \ldots, \hat{\phi}_p, r_p].$$

For purely moving average processes, by replacing $\rho_1, \rho_2, \ldots, \rho_q$ by $r_1, r_2, \ldots, r_q$ respectively in the nonlinear system

$$r_k = (-\theta_1 + \theta_1 \theta_2 + \ldots + \theta_{k-1} \theta_k)/(1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2)$$

($k = 1, 2, \ldots, q$)

one can solve for estimates $\hat{\theta}_0, \hat{\theta}_0, \ldots, \hat{\theta}_0$. Since multiple solutions for the estimates exist, the one satisfying the invertibility condition should be chosen. $^{34}$ $\hat{\delta}_0$ is taken as $\bar{w}$ and $\hat{\sigma}^2_0$ as

$^{34}$Box and Jenkins [4, pp. 195-198], show that for a given covariance structure corresponding to an ARMA $(p,q)$ processes, it is always true that one and only one of the multiple solutions can satisfy the invertibility condition.
The calculation of initial estimates for a mixed process is done as follows. First, the autoregressive parameters are estimated by solving the system

\[ r_k = \phi_1 r_{k-1} + \phi_2 r_{k-2} + \ldots + \phi_p r_{k-p} \quad (k = q+1, q+2, \ldots, q+p) \]

for \( \hat{\theta}_{01}, \hat{\theta}_{02}, \ldots, \hat{\theta}_{0p} \). These estimates are then inserted as replacements for \( \phi_1, \phi_2, \ldots, \phi_p \) into the system

\[ r_k = \phi_1 r_{k-1} + \ldots + \phi_p r_{k-p} + [r_{wa} (k) - \theta_{01} r_{wa} (k-1) - \ldots - \theta_q r_{wa} (k-q)] \quad (k = 1, 2, \ldots, q) \]

relating the \( \phi_i 's, \theta_j 's \) and \( \rho_k 's \). \( r_{wa} (k) \) can be expressed in terms of the \( \phi_i 's, \theta_j 's, \) and \( r_k 's \). Since these equations will be nonlinear in the \( \theta_j 's, \) the invertible solution should be chosen. \( \hat{\theta}_0 \) is taken as

\[ \hat{\theta}_0^2 = \frac{T}{\sum (w_t - \bar{w})^2 / (T-1)} \left[ 1 + \hat{\theta}_{01}^2 + \hat{\theta}_{02}^2 + \ldots + \hat{\theta}_{0q}^2 \right] \]

B. Back Forecasting

In order to evaluate the unconditional sum of squares function in the form

\[ S(\delta, \Gamma) = \sum_{t=1-Q}^{T} [a_t]^2 \]

We need to estimate the unobserved values \( [a_{1-q}] \ldots [a_0] \). This can be
done by using an alternative form of the ARMA(p,q) process, viz.,

\[ \phi(B) \tilde{w} = \theta(B) a_t \]

The alternative representation turns the process around in time it is called the backward process, and is given by

\[ \phi(F) \tilde{w}_t = \theta(F) e_t \quad \text{(A-1)} \]

where

\[ F \tilde{w}_t = \tilde{w}_{t+1} \quad \text{and} \quad F^m \tilde{w}_t = \tilde{w}_{t+m} \quad \text{\(F=B^{-1}\) is the forward shift operator)} \]

The \(e_t\) have exactly the same probability structure as the \(a_t\), that is, normally and independently distributed with mean zero and common variance, \(\sigma^2\). \(\phi(\cdot)\) and \(\theta(\cdot)\) are the same polynomials as before. The backward process expresses \(w_t\) entirely in terms of future \(w\)'s and \(e\)'s. The backward process is equivalent in the sense that the covariance structure of the \(w\)'s remains unchanged; \(^{35}\) \(w_t\) is still stationary and invertible (meaning \(\phi^{-1}(F)\theta(F)\) evaluated at \(F=1\) converges).

Taking expectations in (A-1) conditional on \(\phi, \theta, \delta, \sigma^2, \) and \(w_1, \ldots, w_T\) we have

\[ \phi(F)[\tilde{w}_t] = \theta(F)[e_t], \quad t = 1, 2, \ldots, T \quad \text{(A-2)} \]

where we denote the expectation of a random variable (be it \(w_t, e_t, \) or \(a_t\)) conditional on \(\phi, \theta, \delta, \sigma^2, \) and \(w_1, \ldots, w_T\) by that variable enclosed in brackets.

\(^{35}\) See Box and Jenkins [4, pp. 197-200].
In (A-2), we have

\[ w_t = w_{t-1}, \quad t = 1, 2, \ldots, T \]

\[ e_t = 0, \quad t = 0, -1, -2, \ldots \]

since these disturbances are distributed independently of future observations.

To utilize this form of the process to forecast \( w_0, w_{-1}, \ldots, w_{-Q} \), the unknown quantities, \( [e_{T+1}], [e_{T+2}], \ldots \) are set equal to zero. Then using the preliminary parameter estimates, \( \hat{\phi}_0, \hat{\theta}_0, \) and \( \hat{\delta}_0 \), the system (A-2) is used recursively (beginning at \( t=T \) and working backwards) to solve for estimates, \( \{\hat{e}_t\} \), of \( \{e_t\} \), \( t = 1, 2, \ldots, T \).

Estimates, \( \{\hat{w}_0\}, \{\hat{w}_{-1}\}, \ldots \) of \( \{w_0\}, \{w_{-1}\}, \ldots \) respectively are then solved for using the \( \{e_0\}, \{e_{-1}\}, \ldots \) known to be zero, and the obtained \( \{\hat{e}_t\}, \ t = 1, 2, \ldots, T \).

For computation of \( S(\sigma, \Gamma) \), that is,

\[
\sum_{t=1-Q}^{T} [a_t]^2, \quad \text{the quantities, } [a_t], \ t = -Q, -(Q+1), \ldots \text{ are set to zero since for any MA(Q) process,}^{36} [a_{-Q}, a_{-Q+1}, \ldots \text{ are distributed independently of } w_1, \ldots, w_T.]
\]

Beginning at \( t=1-Q \), the \( \{a_t\}'s \) \( (t = 1-Q, \ldots, T) \) now can be solved for by evaluating

\[
\hat{\phi}_0(B)[\hat{w}_t] = \hat{\delta}_0(B)[a_t]
\]

---

\(^{36}\text{Recall an MA(Q) process is approximating the autoregressive and mixed processes.}\)
recursively with

\[ [w_t] = [\hat{\phi}_t], \quad t = 0, -1, \ldots, 1 - Q \]

\[ = w_t, \quad t = 1, 2, \ldots, T \]

At each iteration of the nonlinear estimation procedure, new back-
forecasts are computed conditional on the latest set of values for \( \hat{\phi}_k \),
\( \hat{\theta}_k \), and \( \hat{\delta}_k \).