Estimation, Prediction, and Interpolation for Nonstationary Series With the Kalman Filter

Víctor GÓMEZ and Agustín MARAVALL*

We show how our definition of the likelihood of an autoregressive integrated moving average (ARIMA) model with missing observations, alternative to that of Kohn and Ansley and based on the usual assumptions made in estimation of and forecasting with ARIMA models, permits a direct and standard state–space representation of the nonstationary (original) data, so that the ordinary Kalman filter and fixed point smoother can be efficiently used for estimation, forecasting, and interpolation. In this way, the problem of estimating missing values in nonstationary series is considerably simplified. The results are extended to regression models with ARIMA errors, and a computer program is available from the authors.

KEY WORDS: ARIMA models; Likelihood function; Missing observations; Nonstationarity; Time series

1. INTRODUCTION

The Kalman filter provides a well-established procedure to compute the likelihood of a time series which is the outcome of a stationary autoregressive moving average (ARMA) process; see, for example, Harvey and Phillips (1979), or Pearlman (1980). Further, Jones (1980) extended the procedure for the case of missing observations in the series. However, when extending the method to a nonstationary autoregressive integrated moving average (ARIMA) process, the likelihood cannot be defined in the usual sense. The main difficulty lies in the specification of the starting conditions to initialize the filter. One cannot use, as in the stationary case, the distribution of the initial state vector, because in the nonstationary case this distribution is not properly defined. Besides, when there are observations missing, it is clearly not possible to use the likelihood of the differenced series (i.e., of its stationary transformation).

There have been several attempts to overcome the difficulty (see, for example, Bell and Hillmer 1991; de Jong 1988, 1991; Harvey and Peters 1990; Harvey and Pierse 1984; and Kohn and Ansley 1986). The path-breaking contribution of Harvey and Pierse, extending the state–space methodology to ARIMA models with missing observations, presented two limitations. First, it could not be applied to series with missing values near the start or the end of the series; second, the chosen state–space representation was not minimal. Possibly, the present state of the art is the powerful methodology developed by Ansley and Kohn in a sequence of papers. To define the likelihood, the data are transformed to eliminate dependence on the starting values. Then, to obtain an efficient procedure, a modified Kalman filter is used to compute the likelihood, and a modified fixed-point smoothing algorithm interpolates the missing observations. Both are generalizations of the ordinary Kalman filter and the ordinary fixed-point smoothing algorithms for handling a partially diffuse initial state vector (see Kohn and Ansley 1986).

In this article we show how an alternative definition of the likelihood, based on the usual hypothesis made in estimation (Box and Jenkins 1976) and prediction (Brockwell and Davis 1987) of ARIMA models, permits a standard state–space representation of the nonstationary series, easy to program, that does not require any transformation of the data and provides a convenient structural interpretation of the state variable. As a consequence, the ordinary Kalman filter and the ordinary fixed-point smoothing algorithms can be efficiently used without modification for estimation, forecasting, and interpolation. In this way, the problem of missing observations in nonstationary series can be simplified considerably. The results are extended to regression models with ARIMA errors.

It is seen how our likelihood coincides with that of Harvey and Pierse (1984) when the latter is applicable and is also equal to that of Kohn and Ansley (1986). When no observation is missing, our likelihood is of course the same as that of Box and Jenkins (1976).

Because most of the results for the stationary case will be valid for the nonstationary case, we begin by briefly reviewing the use of the Kalman filter for stationary series in Section 2. We then proceed to analyze a nonstationary series that follows a general ARIMA model. First, the definition of the likelihood function is considered in Section 3.1, and the state–space representation of the series, the Kalman filter, and the appropriate starting conditions are developed in Section 3.2. Estimation of the model is explained in Section 3.3, and interpolation and prediction of the series is covered in Section 3.4. Special attention is paid to the difficulties that may arise when estimating missing observations at the beginning of the series, and the effect thereof on interpolation and prediction. The relationship between our approach and that of Kohn and Ansley is analyzed in Section 3.5, the methodology is extended to regression models with ARIMA errors and missing observations in Section 3.6, the advantages of our approach are summarized in Section 3.7. Finally, a numerical application consisting of the four data sets considered by Kohn and Ansley (1986) and an additional one that illustrates a type of difficulty not discussed by them are presented in Section 4. We have placed all proofs in Appendix A and an example to illustrate the several steps in our approach in Appendix B; the example is the same as that used by Kohn and Ansley (1986). As mentioned in Section 4, a computer program and its documentation are available from us on request.

* Víctor Gómez is Statistician, Instituto Nacional de Estadística, 28046 Madrid, Spain. Agustín Maravall is Professor, European University Institute, Badia Fiesolana, I-50016 S. Domenico di Fiesole (FI), Italy. The authors are grateful to the reviewers for their helpful comments.

Journal of the American Statistical Association
June 1994, Vol. 89, No. 426, Theory and Methods

2. STATIONARY SERIES, ARMA MODEL

2.1 Prediction Error Decomposition

Let the observed series $\mathbf{z} = (z(1), z(2), \ldots, z(N))'$ be the outcome of the ARMA model,

$$\phi(B)z(t) = \theta(B)a(t), \tag{2.1}$$

where $\phi(B) = 1 + \phi_1 B + \cdots + \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ are finite polynomials in the lag operator B, of orders p and q, and $\{a(t)\}$ is a sequence of independent $N(0, \sigma^2)$ variables. The model is assumed stationary; that is, all roots of the polynomial $\phi(B)$ lie outside the unit circle. Using the prediction error decomposition, the likelihood can be written as

$$L(z(1), z(2), \ldots, z(N)) = L(z(1)) \prod_{t=2}^{N} L(z(t)|z(t-1), \ldots, z(1)), \tag{2.2}$$

where the vertical bar denotes conditional distribution. Defining, for $t = 2, \ldots, N$,

$$\hat{z}(t|t-1) = E(z(t)|z(t-1), \ldots, z(1)) \tag{2.3a}$$

and

$$\sigma^2(t|t-1) = \operatorname{var}(z(t)|z(t-1), \ldots, z(1))/\sigma^2 = E(z(t) - \hat{z}(t|t-1))^2/\sigma^2 \tag{2.3b}$$

and using the marginal distribution of z to set the starting conditions $\hat{z}(1|0) = E(z(1)) = 0$ and $\sigma^2(1|0) = (1/\sigma^2)\operatorname{var}(z(1))$, the likelihood can be written as

$$(2\pi\sigma^2)^{-N/2} \left(\prod_{t=1}^{N} \sigma^2(t|t-1) \right)^{-1/2} \times \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{N} ((z(t) - \hat{z}(t|t-1))/\sigma(t|t-1))^2 \right).$$

Let $e(t), t = 1, \ldots, N$, denote the sequence of standardized one-period-ahead forecast errors $e(t) = (z(t) - \hat{z}(t|t-1))/\sigma(t|t-1)$, and define the vector $\mathbf{e} = (e(1), \ldots, e(N))'$. Then the log-likelihood can be expressed, apart from a constant, as

$$l = -\frac{1}{2}\left\{ N \ln(\sigma^2) + \ln\left(\prod_{t=1}^{N} \sigma^2(t|t-1) \right) + \mathbf{e}'\mathbf{e}/\sigma^2 \right\}.$$

Assuming known model parameters, l will be maximized with respect to σ^2 when

$$\hat{\sigma}^2 = (1/N)\mathbf{e}'\mathbf{e}. \tag{2.4}$$

Therefore, σ^2 can be concentrated out of the function l, yielding the concentrated log-likelihood, which, apart from a constant, is

$$l^* = -\frac{1}{2}\left\{ N \ln(\mathbf{e}'\mathbf{e}) + \ln\left(\prod_{t=1}^{N} \sigma^2(t|t-1) \right) \right\} = -\frac{N}{2} \ln S,$$

where

$$S = \left(\prod_{t=1}^{N} \sigma(t|t-1) \right)^{1/N} \mathbf{e}'\mathbf{e} \left(\prod_{t=1}^{N} \sigma(t|t-1) \right)^{1/N}. \tag{2.5}$$

Thus exact maximum likelihood (ML) estimation of the model parameters minimizes the nonlinear sum of squares S, for which Marquardt's method (see, for example, Fletcher 1987, chap. 5) provides a robust and dependable procedure. (When S is replaced by $\tilde{S} = \mathbf{e}'\mathbf{e}$, the method is often called unconditional least squares.)

The vector $\mathbf{e}$ contains the sequence of orthonormal variables obtained from the Gram–Schmidt orthonormalization process applied to the series $\mathbf{z}$ (Wecker and Ansley 1983). Hence the random variables $e(t)$, $t = 1, \ldots, N$, are independent $N(0, \sigma^2)$ variables. Furthermore, if the covariance matrix of $\mathbf{z}$ is $\sigma^2\mathbf{\Omega}$, and $\mathbf{\Omega} = \mathbf{LL}'$ is the Cholesky decomposition of $\mathbf{\Omega}$ with $\mathbf{L}$ lower triangular, then it is straightforward to verify that $\mathbf{e} = \mathbf{L}^{-1}\mathbf{z}$, $|\mathbf{\Omega}| = |\mathbf{L}|^2$, and $|\mathbf{L}| = \prod_{t=1}^{N} \sigma(t|t-1)$, so that S can be rewritten compactly as $S = |\mathbf{L}|^{1/N} \times \mathbf{e}'\mathbf{e}|\mathbf{L}|^{1/N}$.

2.2 State–Space Representation and the Kalman Filter

Among the several state–space representations of ARMA models, we select that of Jones (1980), originally proposed by Akaike (1974). It provides a minimal representation, easy to program with the Kalman filter, where the state vector has a convenient structural interpretation. (It is also found in some standard statistical packages, such as SAS.)

If $\{z(t)\}$ follows the ARMA process given by (2.1), then letting $r = \max\{p, q+1\}$ and defining $\phi_i = 0$ when $i > p$, the state–space representation is given by

$$\mathbf{x}(t) = \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}a(t) \tag{2.6a}$$

and

$$z(t) = \mathbf{H}'\mathbf{x}(t), \tag{2.6b}$$

where $\mathbf{x}(t) = (z(t), z(t+1|t), \ldots, z(t+r-1|t))'$, $\mathbf{G} = (1, \psi_1, \ldots, \psi_{r-1})'$, $\mathbf{H} = (1, 0, \ldots, 0)'$,

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -\phi_r & -\phi_{r-1} & -\phi_{r-2} & \cdots & -\phi_1 \end{bmatrix},$$

and the ψ_i weights are the coefficients obtained from $\psi(B) = \sum_{i=0}^{\infty} \psi_i B^i = \theta(B)/\phi(B)$.

The expression $z(t+j|t)$ is the orthogonal projection of $z(t+j)$ on the subspace generated by $\{z(s): s \le t\}$ and coincides with the conditional expectation $E(z(t+j)|z(s): s \le t)$. Thus the state vector $\mathbf{x}(t)$ contains the series $z(t)$ and its $(r-1)$-periods-ahead forecast function with respect to the semi-infinite sample $\{z(s): s \le t\}$. Note that when the projection is on the finite sample $\{z(s): 1 \le s \le t\}$, we represent it with a hat, as in (2.3a).

Because σ^2 can be concentrated out of the likelihood, without loss of generality we can set $\sigma^2 = 1$. (Once the parameters have been estimated, $\hat{\sigma}^2$ will be obtained with (2.4)). The Kalman filter consists of the following equations: first, the starting conditions $\mathbf{x}(0|0) = E(\mathbf{x}(0)) = 0$ and $\mathbf{\Sigma}(0|0) = \operatorname{var}(\mathbf{x}(0))$, and then, the recursions

$$\mathbf{x}(t|t-1) = \mathbf{F}\mathbf{x}(t-1|t-1),$$
$$\boldsymbol{\Sigma}(t|t-1) = \mathbf{F}\boldsymbol{\Sigma}(t-1|t-1)\mathbf{F}' + \mathbf{Q},$$
$$\mathbf{K}(t) = \boldsymbol{\Sigma}(t|t-1)\mathbf{H}(\mathbf{H}'\boldsymbol{\Sigma}(t|t-1)\mathbf{H})^{-1},$$
$$\boldsymbol{\Sigma}(t|t) = (\mathbf{I} - \mathbf{K}(t)\mathbf{H}')\boldsymbol{\Sigma}(t|t-1),$$

and

$$\mathbf{x}(t|t) = \mathbf{x}(t|t-1) + \mathbf{K}(t)(z(t) - \mathbf{H}'\mathbf{x}(t|t-1)), \quad t = 1, 2, \ldots, N, \quad (2.7)$$

where $\mathbf{Q} = \mathbf{G}\mathbf{G}'$ and

$$\mathbf{x}(t|T) = E(\mathbf{x}(t)|z(1), \ldots, z(T)) \quad (2.8a)$$

and

$$\boldsymbol{\Sigma}(t|T) = \mathrm{var}(\mathbf{x}(t)|z(1), \ldots, z(T)), \quad 1 \le t \le N, \quad 1 \le T \le N. \quad (2.8b)$$

The filter can also be initialized with $\mathbf{x}(1|0) = 0$ and $\boldsymbol{\Sigma}(1|0) = \boldsymbol{\Sigma}(0|0)$, because stationarity implies $\boldsymbol{\Sigma}(0|0) = \mathbf{F}\boldsymbol{\Sigma}(0|0)\mathbf{F}' + \mathbf{Q}$. An efficient procedure to compute the initial covariance matrix $\boldsymbol{\Sigma}(1|0)$ was provided by Jones (1980).

To maximize the likelihood, at each iteration it is necessary to obtain the values $\hat{\mathbf{z}}(t|t-1) = \mathbf{H}'\mathbf{x}(t|t-1)$ and $\sigma^2(t|t-1) = \mathbf{H}'\boldsymbol{\Sigma}(t|t-1)\mathbf{H}$ and the residual $e(t) = (z(t) - \hat{z}(t|t-1))/\sigma(t|t-1)$. Once the last iteration ($t = N$) has been completed, the objective function S is obtained with (2.5).

When $q \ge p$, an alternative representation that yields a state–space vector of minimal dimension can be obtained by eliminating the first element in both the state vector and the $\mathbf{G}$ vector and by replacing the observation equation (2.6b) with $z(t) = \mathbf{H}'\mathbf{x}(t-1) + a(t)$. The new matrix $\mathbf{F}$ in $\mathbf{x}(t) = \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}a(t)$ would be the one of (2.6a) without the first row and the first column, the new matrix $\mathbf{G}$ would be the old one without the first element, and the new matrix $\mathbf{H}$ would be the old one without the last 0. In the rest of the article we shall always refer to the representation (2.6).

2.3 ARMA Model with Missing Observations

Assume, in all generality, that only the values $\{z(t_1), z(t_2), \ldots, z(t_M)\}$, $1 \le t_1 < \cdots < t_M \le N$, are observed. To obtain the prediction error decomposition, the observation equation (2.6b) can be replaced by

$$z(t) = \mathbf{H}'(t)\mathbf{x}(t) + \alpha(t)W(t), \quad t = 1, \ldots, N, \quad (2.9)$$

where $\mathbf{H}'(t) = (1, 0, \ldots, 0)$, $\alpha(t) = 0$ if $z(t)$ is observed, and $\mathbf{H}'(t) = (0, 0, \ldots, 0)$, $\alpha(t) = 1$ if $z(t)$ is missing (Brockwell and Davis 1987, p. 494). The variable $W(t)$ represents a $N(0, 1)$ sample, independent of $\{z(t_1), \ldots, z(t_M)\}$. Thus when $z(t)$ is missing, $\mathbf{x}(t|t) = \mathbf{x}(t|t-1)$ and $\boldsymbol{\Sigma}(t|t) = \boldsymbol{\Sigma}(t|t-1)$, and both the residual and the standard error corresponding to a missing value are ignored when computing the function S of (2.5) (see Jones 1980).

Having obtained parameter estimates by minimizing the appropriate function S, estimators of the missing values can be obtained through the fixed-point smoother (FPS) (see Anderson and Moore 1979). Assume that the jth observation is missing and define the starting condition $\boldsymbol{\Sigma}^a(j|j-1) = \boldsymbol{\Sigma}(j|j-1)$. Then, letting $k = j+1, \ldots, N$, the equations of the FPS are given by

$$\boldsymbol{\Sigma}^a(k+1|k) = \boldsymbol{\Sigma}^a(k|k-1)(\mathbf{F} - \mathbf{F}\mathbf{K}(k)\mathbf{H}'(k))',$$
$$\mathbf{K}^a(k) = \boldsymbol{\Sigma}^a(k|k-1)\mathbf{H}(k)(\mathbf{H}'(k)\boldsymbol{\Sigma}(k|k-1)\mathbf{H}(k) + \alpha^2(k)\sigma_W^2)^{-1},$$
$$\mathbf{x}(j|k) = \mathbf{x}(j|k-1) + \mathbf{K}^a(k)(z(k) - \mathbf{H}'(k)\mathbf{x}(k|k-1)),$$

and

$$\boldsymbol{\Sigma}(j|k) = \boldsymbol{\Sigma}(j|k-1) - \boldsymbol{\Sigma}^a(k|k-1)\mathbf{H}(k)(\mathbf{K}^a(k))',$$

where $\mathbf{H}(k)$ and $\alpha(k)$ are as in (2.9), $\sigma_W^2 = 1$, and $\mathbf{x}(j|k)$, $\boldsymbol{\Sigma}(j|k)$, and $\mathbf{K}(k)$ are as in (2.7) and (2.8). For ARMA models, the set of equations of the FPS can be simplified by noticing that only the first element of $\mathbf{x}(j|k)$ and the (1, 1) element of $\boldsymbol{\Sigma}(j|k)$—namely, $\hat{z}(j|k)$ and $\sigma^2(j|k)$—are of interest. Defining $\mathbf{v}'(j|k) = \mathbf{H}'\boldsymbol{\Sigma}^a(k|k-1)$ and $b(k) = \mathbf{H}'\mathbf{K}^a(k)$, $k \ge j$, where $\mathbf{H}$ is as in (2.6b), the FPS simplifies to

$$\mathbf{v}'(j|j) = \mathbf{H}'\boldsymbol{\Sigma}^a(j|j-1) = \mathbf{H}'\boldsymbol{\Sigma}(j|j-1),$$
$$\mathbf{v}'(j|k+1) = \mathbf{v}'(j|k)(\mathbf{F} - \mathbf{F}\mathbf{K}(k)\mathbf{H}'(k))',$$
$$b(k) = \mathbf{v}'(j|k)\mathbf{H}(k)(\mathbf{H}(k)'\boldsymbol{\Sigma}(k|k-1)\mathbf{H}(k) + \alpha^2(k)\sigma_W^2)^{-1},$$
$$\hat{z}(j|k) = \hat{z}(j|k-1) + b(k)(z(k) - \mathbf{H}(k)'\mathbf{x}(k|k-1)),$$

and

$$\sigma^2(j|k) = \sigma^2(j|k-1) - \mathbf{v}'(j|k)\mathbf{H}(k)b(k), \quad k \ge j,$$

which only requires storage of the vector $\mathbf{v}(j|k)$ and of the scalars $\hat{z}(j|k)$ and $\sigma^2(j|k)$.

2.4. Regression Model with Missing Observations and ARMA Errors

Consider the regression model

$$z(t) = \mathbf{y}'(t)\boldsymbol{\beta} + \nu(t), \quad (2.10a)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_h)'$ is a vector of parameters, $\mathbf{y}'(t)$ is a vector of h independent variables, $z(t)$ is the dependent variable, and $\nu(t)$ is assumed to follow the ARMA model given by (2.1). If, as before, $\{z(t_1), \ldots, z(t_M)\}$ denote the observed values defining the vectors $\mathbf{z} = (z(t_1), \ldots, z(t_M))'$, $\boldsymbol{\nu} = (\nu(t_1), \ldots, \nu(t_M))'$, and the $M \times h$ matrix $\mathbf{Y}$ with the vectors $\mathbf{y}'(t)$, $t = t_1, \ldots, t_M$, as rows, then we can write

$$\mathbf{z} = \mathbf{Y}\boldsymbol{\beta} + \boldsymbol{\nu}, \quad (2.10b)$$

where the matrix $\mathbf{Y}$ is assumed of rank h. Denoting by $\sigma^2\boldsymbol{\Omega}$ the covariance matrix of $\boldsymbol{\nu}$, the likelihood corresponding to (2.10b) is given by

$$(2\pi\sigma^2)^{-M/2}|\boldsymbol{\Omega}|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{z} - \mathbf{Y}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{z} - \mathbf{Y}\boldsymbol{\beta})\right).$$

The $\boldsymbol{\beta}$ parameters and those of the ARMA model for $\nu(t)$ can be jointly estimated using the Kalman filter, as described earlier, by simply defining the state vector $\mathbf{x}(t) = (\nu(t), \nu(t+1|t), \ldots, \nu(t+r-1|t))'$ and using Equations (2.6a) and

$$z(t) = \mathbf{y}'(t)\beta + \mathbf{H}'(t)\mathbf{x}(t) + \alpha(t)W(t) \quad (2.11)$$

instead of (2.9). As before, the only residuals included in the computation of the likelihood are those for $t = t_1, \ldots, t_M$ given by $e(t) = (z(t) - \mathbf{y}'(t)\beta - \hat{\nu}(t|t-1))/\sigma(t|t-1)$, and the estimators of the missing values are obtained with the FPS. When this procedure is followed, mean squared error (MSE) estimators obtained in subsequent smoothing or forecasting operations will all be conditional on β (see Harvey and Pierse 1984).

One way to overcome this limitation, which at the same time yields a more efficient computational procedure, is to use the approach of Kohn and Ansley (1985), concentrating β and σ^2 out of the likelihood function. The best linear unbiased estimator (BLUE) of β can be obtained by GLS, and then the Kalman filter and the FPS can be used to compute minimum MSE missing observations estimators and forecasts, as well as their MSE. More specifically, let $\mathbf{\Omega} = \mathbf{LL}'$ be the Cholesky decomposition of $\mathbf{\Omega}$, with $\mathbf{L}$ lower triangular. Premultiplying (2.10b) by $\mathbf{L}^{-1}$, we obtain the OLS model

$$\mathbf{L}^{-1}\mathbf{z} = \mathbf{L}^{-1}\mathbf{Y}\beta + \mathbf{L}^{-1}\boldsymbol{\nu}. \quad (2.12)$$

Replacing $\mathbf{z}$ with $\boldsymbol{\nu}$, the Kalman filter associated with Equation (2.6a) and the observation equation (2.9) can be seen as an algorithm to obtain the residuals $\mathbf{e} = \mathbf{L}^{-1}\mathbf{z}$ and the determinant $|\mathbf{L}|$. The same algorithm applied to the columns of $\mathbf{Y}$ provides simultaneously $\mathbf{e} = \mathbf{L}^{-1}\mathbf{z}$, $\mathbf{L}^{-1}\mathbf{Y}$, and $|\mathbf{L}|$ (see Kohn and Ansley 1985 and Wecker and Ansley 1983). Therefore, applying the Kalman filter in this way, we can move from (2.10b) to (2.12).

Using the **QR** algorithm on the matrix $\mathbf{L}^{-1}\mathbf{Y}$, an orthogonal $M \times h$ matrix $\mathbf{Q}$ is obtained, such that $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{Y} = (\mathbf{R}', 0')'$, where $\mathbf{R}$ is an upper triangular $h \times h$ matrix with nonzero elements in the main diagonal. Let $\mathbf{Q}_1'$ and $\mathbf{Q}_2'$ be the submatrices formed with the rows of $\mathbf{Q}'$ such that $\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{Y} = \mathbf{R}$ and $\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{Y} = 0$. Premultiplying (2.12) by $\mathbf{Q}'$, it is found that the GLS estimator of β is given by $\hat{\beta} = \mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{z}$, from which the estimator of σ^2 is also obtained by $\hat{\sigma}^2 = (1/M)(\mathbf{z} - \mathbf{Y}\hat{\beta})'\mathbf{\Omega}^{-1}(\mathbf{z} - \mathbf{Y}\hat{\beta}) = (1/M)\mathbf{z}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{z}$. The function to be minimized becomes $S = |\mathbf{L}|^{1/M} \times \mathbf{z}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{z}|\mathbf{L}|^{1/M}$. (For more details on how to obtain the missing observation estimators and forecasts, as well as their MSE, see Kohn and Ansley 1985.)

3. NONSTATIONARY SERIES, ARIMA MODEL

3.1 The Likelihood Function

Let $\{z(t)\}$ be a nonstationary process such that the transformation $u(t) = \delta(B)z(t)$ renders it stationary, and let $\{u(t)\}$ follow the ARMA model (2.1). Then $\{z(t)\}$ follows the nonstationary model

$$\phi(B)\delta(B)z(t) = \theta(B)a(t), \quad (3.1)$$

where $\delta(B) = 1 + \delta_1 B + \cdots + \delta_d B^d$ denotes a polynomial in B with all roots on the unit circumference. Typically, $\delta(B)$ will contain regular and/or seasonal differences. Let $\mathbf{z} = (z(1), z(2), \ldots, z(N))'$ be the observed series, and let $\mathbf{u} = (u(d+1), u(d+2), \ldots, u(N))'$ be the differenced series. The nonstationarity of $\{z(t)\}$ prevents us from using the prediction error decomposition (2.2), because the distribution of $z(1)$ is not well defined. To define the likelihood properly, we use the following assumptions; the first assumption is a standard one when forecasting with ARIMA models (see Brockwell and Davis 1987, pp. 304–307):

Assumption A: The variables $\{z(1), \ldots, z(d)\}$ are independent of the variables $\{u(t)\}$.

Assumption B: The variables $\{z(1), \ldots, z(d)\}$ are jointly normally distributed.

The likelihood of ARIMA models is usually defined as the likelihood of the differenced series, $L(\mathbf{u})$ (see Box and Jenkins 1976, chap. 7). But an expression in terms of the original series itself could be very useful, because the Kalman filter could then be used directly to estimate missing observations in nonstationary series. Accordingly, we define as our likelihood the density

$$L(z(d+1), \ldots, z(N)|Z_*), \quad (3.2)$$

where $Z_* = \{z(1), z(2), \ldots, z(d)\}$. This is a well-defined likelihood because, following Bell (1984, p. 650), the variable $z(t)$ can be expressed as

$$z(t) = \mathbf{A}'(t)\mathbf{z}_* + \sum_{i=0}^{t-d-1} \xi_i u(t-i), \qquad t > d, \quad (3.3)$$

where $\mathbf{z}_* = (z(1), \ldots, z(d))'$, $\xi(B) = 1/\delta(B) = \sum_{i=0}^{\infty} \xi_i B^i$, and $\mathbf{A}(t) = (A_1(t), \ldots, A_d(t))'$ and the coefficients $A_i(t)$ are obtained recursively from

$$A_i(t) = -\delta_1 A_i(t-1) - \cdots - \delta_d A_i(t-d), \qquad t > d, \quad i = 1, \ldots, d,$$

and

$$A_i(j) = 1 \quad \text{if } i = j = 1, \ldots, d;$$
$$= 0 \quad \text{if } i, j = 1, \ldots, d; \quad i \neq j. \quad (3.4)$$

Let Ξ be the lower triangular $(N-d) \times (N-d)$ matrix with rows the vectors $(\xi_{j-1}, \xi_{j-2}, \ldots, 1, 0, \ldots, 0)$, $j = 1, \ldots, N-d$, and let $\mathbf{A}$ be the $(N-d) \times d$ matrix with rows the vectors $\mathbf{A}'(t)$, $t = d+1, \ldots, N$. Define $\mathbf{J} = (\mathbf{J}_1', \mathbf{J}_2')'$, where $\mathbf{J}_1$ and $\mathbf{J}_2$ are the $d \times N$ and $(N-d) \times N$ submatrices such that $\mathbf{J}_1 = (\mathbf{I}_d, 0)$ and $\mathbf{J}_2 = (\mathbf{A}, \Xi)$. Then the transformation

$$\mathbf{z} = \mathbf{J}[\mathbf{z}_*', \mathbf{u}']' \quad (3.5)$$

has a determinant equal to 1. Therefore, the densities will satisfy, under Assumptions A and B, $L(\mathbf{z}) = L(\mathbf{z}_*, \mathbf{u}) = L(\mathbf{z}_*)L(\mathbf{u})$, from which the following result is obtained.

Lemma 1. $L(z(d+1), \ldots, z(N)|Z_*) = L(\mathbf{u})$.

Lemma 1 establishes the equality between the classical definition of the likelihood in ARIMA models and our definition (3.2). The prediction error decomposition associated with the latter is given by

$$L(z(d+1), \ldots, z(N) \mid Z_*) = \prod_{t=d+1}^{N} L(z(t) \mid z(t-1), \ldots, z(1)). \quad (3.6)$$

For $t = d+2, \ldots, N$, define $\hat{z}(t|t-1)$ and $\sigma^2(t|t-1)$ as in (2.3) and let, for $t = d+1$,

$$\hat{z}(d+1 \mid d) = E(z(d+1) \mid Z_*) = \mathbf{A}'(d+1)\mathbf{z}_*$$

and

$$\sigma^2(d+1 \mid d) = E[(z(d+1) - \hat{z}(d+1 \mid d))^2]/\sigma^2 = \operatorname{var}(u(d+1))/\sigma^2.$$

Then the residuals corresponding to the prediction error decomposition (3.6) are $e(t) = (z(t) - \hat{z}(t|t-1))/\sigma(t|t-1)$, $t = d+1, \ldots, N$. Defining for $t = d+2, \ldots, N$, $\hat{u}(t|t-1) = E(u(t)|u(t-1), \ldots, u(d+1), Z_*) = E(u(t)|u(t-1), \ldots, u(d+1))$, the following lemma shows the equality between the residuals corresponding to (3.6) and those obtained in the Box and Jenkins likelihood. The proof is given in Appendix A.

Lemma 2. Under Assumptions A and B,

$$z(t) - \hat{z}(t|t-1) = u(t) - \hat{u}(t|t-1)$$

and

$$\sigma^2(t|t-1) = \operatorname{var}(u(t)|u(t-1), \ldots, u(d+1))/\sigma^2, \quad t \geq d+1,$$

where $\hat{u}(d+1 \mid d) = E(u(d+1)) = 0$ and $\sigma^2(d+1 \mid d) = \operatorname{var}(u(d+1))/\sigma^2$.

3.2 State–Space Representation of a Nonstationary Series and the Kalman Filter

Bell (1984, pp. 649–651) proved that there is a one-to-one correspondence between $\{z(t)\}$ and $\{Z_*, \{u(t)\}\}$. Using Bell's backward representation as well as (3.3), it is not difficult to check that for any $t \geq d+1$, the subspaces generated by $\{z(s): s \leq t\}$ and $\{Z_*, \{u(s): s \leq t\}\}$ coincide. Therefore, under Assumptions A and B, if we define $z(s|t) = E(z(s)|z(\nu): \nu \leq t)$ and $u(s|t) = E(u(s)|z(\nu): \nu \leq t)$ for $s \geq t \geq d+1$, we have

$$z(s|t) = E(z(s)|z(\nu): \nu \leq t) = E(z(s)|Z_*, \{u(\nu): \nu \leq t\}) \quad (3.7a)$$

and

$$u(s|t) = E(u(s)|z(\nu): \nu \leq t) = E(u(s)|u(\nu): \nu \leq t). \quad (3.7b)$$

Consider model (3.1). Let $\phi^*(B) = \phi(B)\delta(B)$, $\psi^*(B) = \theta(B)/\phi^*(B) = \sum_{i=0}^{\infty} \psi_i^* B^i$, $\phi_i^* = 0$, when $i > p+d$, and $r = \max\{p+d, q+1\}$. Then the following lemma, proved in Appendix A, will allow us to preserve the same state–space representation as in the stationary case.

Lemma 3. $z(t+r-1|t) = -\phi_r^* z(t-1) - \phi_{r-1}^* z(t|t-1) - \cdots - \phi_1^* z(t+r-2|t-1) + \psi_{r-1}^* a(t)$.

As a consequence, for the nonstationary case the state–space representation is also given by (2.6), with the ϕ and ψ coefficients replaced by the ϕ^* and ψ^* coefficients. The Kalman filter can then be applied in an identical manner to compute the likelihood through the prediction error decomposition (3.6). The starting conditions will be different, however. Considering that $\mathbf{x}(d+1) = A_*\mathbf{z}_* + \Xi\mathbf{U}$, where $\mathbf{A}_* = [\mathbf{A}(d+1), \mathbf{A}(d+2), \ldots, \mathbf{A}(d+r)]'$, Ξ is the lower triangular $r \times r$ matrix with rows the vectors $(\xi_{j-1}, \xi_{j-2}, \ldots, 1, 0, \ldots, 0)$, $j = 1, \ldots, r$, and $\mathbf{U} = [u(d+1), u(d+2|d+1), \ldots, u(d+r|d+1)]'$, the starting conditions for the nonstationary case are given by

$$\mathbf{x}(d+1 \mid d) = \mathbf{A}_*\mathbf{z}_* \quad (3.8a)$$

and

$$\Sigma(d+1 \mid d) = \Xi E(\mathbf{U}\mathbf{U}')\Xi' = \Xi\tilde{\Sigma}(d+1 \mid d)\Xi', \quad (3.8b)$$

where $\tilde{\Sigma}(d+1 \mid d) = E(\mathbf{U}\mathbf{U}')$ can be computed from the stationary process $\{u(t)\}$, which follows model (2.1), as in Section 2.2. The dimension of the state–space representation (2.6) is minimal, and hence smaller than that of Harvey and Pierse (1984), and equal to the one of Kohn and Ansley (1986). Our representation has the appeal of its simplicity, of the direct interpretation of the state vector (i.e., the nonstationary series and its $(r-1)$-periods-ahead forecast function), and of the easiness in computing the starting conditions. A simple example to illustrate this and the following sections is discussed in Appendix B.

3.3 Missing Observations

As in the two previous sections, let $\{z(t)\}$ follow the ARIMA process (3.1) with Assumptions A and B holding. Let the available observations be $\{z(t_1), z(t_2), \ldots, z(t_M)\}$, with $1 \leq t_1 < \cdots < t_M \leq N$. If there are no missing observations among the first d values of the series, then one could proceed as in the stationary case, with the fixed-point smoother and the observation equation (2.9), using (3.8) as starting conditions. This is the case considered by Harvey and Pierse (1984). Given that their starting conditions and equations are the same as ours, the two likelihoods coincide.

For the general case, assume that $\mathbf{z}_I = (z(t_1), \ldots, z(t_k))'$ is the vector of observations in $\mathbf{z}_* = (z(1), \ldots, z(d))'$. Letting $\mathbf{z}_J$ denote the vector of missing observations in $\mathbf{z}_*$, then the observed values in $\mathbf{z}$ can be expressed as

$$z(t_i) = \mathbf{A}'(t_i)\mathbf{z}_*, \qquad i = 1, \ldots, k,$$

and

$$z(t_i) = \mathbf{A}'(t_i)\mathbf{z}_* + \tilde{u}(t_i) = \mathbf{B}'(t_i)\mathbf{z}_I + \mathbf{C}'(t_i)\mathbf{z}_J + \tilde{u}(t_i), \qquad i = k+1, \ldots, M,$$

where $\tilde{u}(t) = \sum_{j=0}^{t-d-1} \xi_j u(t-j)$, $t > d$, the vectors $\mathbf{A}'(s)$ have been defined in (3.3) and (3.4), and $\mathbf{B}'(s)$ and $\mathbf{C}'(s)$ are the subvectors of $\mathbf{A}'(s)$ such that $\mathbf{A}'(s)\mathbf{z}_* = \mathbf{B}'(s)\mathbf{z}_I + \mathbf{C}'(s)\mathbf{z}_J$, $s > 0$. Let $\mathbf{z}_{II}$ and $\tilde{\mathbf{u}}$ denote the vectors $(z(t_{k+1}), \ldots, z(t_M))'$ and $(\tilde{u}(t_{k+1}), \ldots, \tilde{u}(t_M))'$, and let $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ denote the $M \times d$, $(M-k) \times k$, and $(M-k) \times (d-k)$ matrices with rows $\mathbf{A}'(s)$, $s = t_1, \ldots, t_M$, $\mathbf{B}'(s)$, and $\mathbf{C}'(s)$, $s = t_{k+1}, \ldots, t_M$. Define $\mathbf{D} = (\mathbf{D}_1', \mathbf{D}_2')'$, where $\mathbf{D}_1$ and $\mathbf{D}_2$ are the $k \times d$

and $(\mathbf{M}-\mathbf{k}) \times d$ submatrices, such that $\mathbf{D}_1 = (\mathbf{I}_k, 0)$ and $\mathbf{D}_2 = (\mathbf{B}, \mathbf{C})$. Then the preceding equations can be rewritten as

$$\mathbf{z} = [\mathbf{z}_I', \mathbf{z}_{II}']' = \mathbf{A}\mathbf{z}_* + [0', \tilde{\mathbf{u}}']' = \mathbf{D}[\mathbf{z}_I', \mathbf{z}_J']' + [0', \tilde{\mathbf{u}}']'. \tag{3.9}$$

Defining $\mathbf{y} = \mathbf{z}_{II} - \mathbf{B}\mathbf{z}_I$, (3.9) implies that

$$\mathbf{y} = \mathbf{C}\mathbf{z}_J + \tilde{\mathbf{u}}. \tag{3.10}$$

A natural way of extending our likelihood (3.2) to the case of missing observations is to consider the likelihood of the observations $\mathbf{z}_{II}$ conditional on $\mathbf{z}_*$ and to treat $\mathbf{z}_J$ as additional parameters. This definition of the likelihood is equivalent to considering (3.10) as a regression model whose errors $\tilde{\mathbf{u}}$ have a known covariance matrix $\sigma^2\Delta$. We define as our likelihood that associated with (3.10); that is,

$$(2\pi\sigma^2)^{-(M-k)/2} |\Delta|^{-1/2} \times \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{C}\mathbf{z}_J)'\Delta^{-1}(\mathbf{y}-\mathbf{C}\mathbf{z}_J)\right), \tag{3.11}$$

where the unknown parameters are σ^2, $\mathbf{z}_J$, and the coefficients $(\phi, \theta) = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ of model (2.1). Using the prediction error decomposition and concentrating σ^2 out of the likelihood, it is seen that maximizing (3.11) over (ϕ, θ) and $\mathbf{z}_J$ is equivalent to minimizing

$$S = |\mathbf{L}|^{1/(M-k)}\mathbf{e}'\mathbf{e}|\mathbf{L}|^{1/(M-k)}$$

over (ϕ, θ) and $\mathbf{z}_J$, where $\mathbf{L}$ is the lower triangular matrix such that $\Delta = \mathbf{L}\mathbf{L}'$ is the Cholesky decomposition of Δ, and $\mathbf{e} = \mathbf{L}^{-1}(\mathbf{y} - \mathbf{C}\mathbf{z}_J) = (e(t_{k+1}), \ldots, e(t_M))'$ is the vector of standardized residuals. For given values of (ϕ, θ) and $\mathbf{z}_J$, the function S can be computed with the Kalman filter. The equations to use are (2.6a) and (2.9), with starting conditions (3.8). In this way, we can jointly estimate the unknown parameters (ϕ, θ) and $\mathbf{z}_J$. As starting values for $\mathbf{z}_J$, linear combinations of adjacent observations can be used.

However, as stated in Section 2.4, a more efficient computational procedure is obtained by using the approach of Kohn and Ansley (1985), whereby σ^2 and $\mathbf{z}_J$ can be concentrated out of the likelihood. This approach not only will allow us to compute MSE not conditional on $\mathbf{z}_J$, but also will place on a sound theoretical basis the discussion of the difficulties that may arise when estimating $\mathbf{z}_J$ and their implications for prediction and interpolation. To see how this approach can be used in the present context, let the Kalman filter be applied to the model $\mathbf{y} = \mathbf{z}_{II} - \mathbf{B}\mathbf{z}_I = \tilde{\mathbf{u}}$. This allows us to compute $\mathbf{L}^{-1}\mathbf{y}$ and $|\mathbf{L}|$. The equations would again be (2.6a) and (2.9); the starting conditions, $\mathbf{x}(d+1|d) = [\mathbf{B}(d+1), \ldots, \mathbf{B}(d+r)]'\mathbf{z}_I$ and (3.8b). Note that it is not necessary to compute the entire matrix $\mathbf{B}$ and that the Kalman filter is applied to the vector of observations $\mathbf{z}_{II}$. The same algorithm applied to the columns of the matrix $\mathbf{C}$, with starting conditions $\mathbf{x}(d+1|d) = 0$ and (3.8b), also permits us to compute $\mathbf{L}^{-1}\mathbf{C}$. We thus consider the model

$$\mathbf{L}^{-1}\mathbf{y} = \mathbf{L}^{-1}\mathbf{C}\mathbf{z}_J + \mathbf{L}^{-1}\tilde{\mathbf{u}}. \tag{3.12}$$

Let the rank of $\mathbf{C}$ be $r_C \le d-k$. Then the matrix $\mathbf{L}^{-1}\mathbf{C}$ has also rank r_C, and two cases need to be distinguished:

1. $\operatorname{rank}(C) = r_C = d-k$. Then the **QR** algorithm applied to $\mathbf{L}^{-1}\mathbf{C}$ yields an orthogonal $(M-k)\times(M-k)$ matrix $\mathbf{Q}$ such that $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C} = (\mathbf{R}', 0')'$, where $\mathbf{R}$ is an upper triangular $(d-k)\times(d-k)$ matrix with nonzero elements in the main diagonal.
2. $\operatorname{rank}(C) = r_C < d-k$. A slight modification of the **QR** algorithm yields an orthogonal $(M-k)\times(M-k)$ matrix $\mathbf{Q}$ such that $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C} = (\mathbf{E}', 0')'$, where $\mathbf{E}$ is an $r_C \times (d-k)$ matrix in row-echelon form. Permuting the variables in $\mathbf{z}_J$ if necessary, which implies permuting the columns of $\mathbf{E}$, we can always assume without loss of generality that the matrix $\mathbf{E}$ has the form $\mathbf{E} = (\mathbf{R}, \mathbf{S})$, where $\mathbf{R}$ is an $r_C \times r_C$ upper triangular matrix with nonzero elements in the main diagonal.

Thus, in general, let the **QR** algorithm when $r_C = d-k$, or its slight modification when $r_C < d-k$, be applied to the matrix $\mathbf{L}^{-1}\mathbf{C}$. Then an orthogonal $(M-k)\times(M-k)$ matrix $\mathbf{Q}$ is obtained such that $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C} = (\mathbf{E}', 0')'$, where $\mathbf{E} = \mathbf{R}$ if $r_C = d-k$ or $\mathbf{E} = (\mathbf{R}, \mathbf{S})$ if $r_C < d-k$. Let $\mathbf{Q}_1'$ and $\mathbf{Q}_2'$ be the submatrices of $\mathbf{Q}'$ formed with the rows of $\mathbf{Q}'$, so that $\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{C}$ consists of the first r_C nonzero rows of $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C}$, and $\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{C} = 0$. If (3.12) is premultiplied by $\mathbf{Q}'$, then it is obtained that

$$\mathbf{Q}_1'\mathbf{L}^{-1}y = \mathbf{E}\mathbf{z}_J + \mathbf{Q}_1'L^{-1}\tilde{\mathbf{u}} \tag{3.13a}$$

and

$$\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y} = \mathbf{Q}_2'\mathbf{L}^{-1}\tilde{\mathbf{u}}. \tag{3.13b}$$

The normal system of equations corresponding to (3.13) is $\mathbf{E}'\mathbf{E}\mathbf{z}_J = \mathbf{E}'\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y}$. This system of linear equations is always consistent, and its general solution is (see Rao 1973, pp. 222–223)

$$\hat{\mathbf{z}}_J = (\mathbf{E}'\mathbf{E})^{-}\mathbf{E}'\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} + (\mathbf{I}_{d-k} - (\mathbf{E}'\mathbf{E})^{-}\mathbf{E}'\mathbf{E})\mathbf{x}, \tag{3.14a}$$

where $\mathbf{x}$ is an arbitrary vector of dimension $d-k$ and $(\mathbf{E}'\mathbf{E})^{-}$ is a generalized inverse of $\mathbf{E}'\mathbf{E}$. If $r_C = d-k$, then $\mathbf{E} = \mathbf{R}$, an upper triangular matrix, and (3.14a) becomes

$$\hat{\mathbf{z}}_J = \mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y}. \tag{3.14b}$$

Therefore, instead of maximizing the likelihood (3.11) (or, equivalently, minimizing the function S) over (ϕ, θ) and $\mathbf{z}_J$, we can concentrate σ^2 and $\mathbf{z}_J$ out of the likelihood by using the regression (3.13) to obtain $\hat{\mathbf{z}}_J$ and $\hat{\sigma}^2 = (1/M-k)(\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{C}\hat{\mathbf{z}}_J)'(\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{C}\hat{\mathbf{z}}_J)$. In this case, we seek to minimize the function

$$S^* = |\mathbf{L}|^{1/(M-k)}(\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{C}\hat{\mathbf{z}}_J)' \times (\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{C}\hat{\mathbf{z}}_J)|\mathbf{L}|^{1/(M-k)}$$

over (ϕ, θ) to obtain $(\hat{\phi}, \hat{\theta})$; it is easy to verify that

$$S^* = |\mathbf{L}|^{1/(M-k)}[(\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} - \mathbf{E}\hat{\mathbf{z}}_J)'(\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} - \mathbf{E}\hat{\mathbf{z}}_J) + \mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}]|\mathbf{L}|^{1/(M-k)}.$$

Because the matrix $\mathbf{E}(\mathbf{E}'\mathbf{E})^{-}\mathbf{E}'$ is symmetric and idempotent and has rank r_C (see Rao 1973, p. 25), there exists an orthogonal $r_C \times r_C$ matrix $\mathbf{Q}_\mathbf{E}$ such that $\mathbf{Q}_\mathbf{E}\mathbf{E}(\mathbf{E}'\mathbf{E})^{-}\mathbf{E}'\mathbf{Q}_\mathbf{E}' = \mathbf{I}_{r_C}$ and, therefore, $\mathbf{E}(\mathbf{E}'\mathbf{E})^{-}\mathbf{E}' = \mathbf{I}_{r_C}$. This implies

$$\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} - \mathbf{E}\hat{\mathbf{z}}_J = \mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} - \mathbf{E}(\mathbf{E}'\mathbf{E})^{-}\mathbf{E}'\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} + (\mathbf{E} - \mathbf{E}(\mathbf{E}'\mathbf{E})^{-}\mathbf{E}'\mathbf{E})\mathbf{x} = 0,$$

and S^* can be rewritten as

$$S^* = |\mathbf{L}|^{1/(M-k)}\mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}|\mathbf{L}|^{1/(M-k)}. \quad (3.15)$$

The estimator $\hat{\sigma}^2$ becomes

$$\hat{\sigma}^2 = (1/(M-k))\mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}. \quad (3.16)$$

In summary, to estimate model (3.1) concentrating $\mathbf{z}_J$ and σ^2 out of the likelihood, we apply the Kalman filter with Equations (2.6a) and (2.9) and starting conditions $\mathbf{x}(d+1 \mid d) = [\mathbf{B}(d+1), \ldots, \mathbf{B}(d+r)]'\mathbf{z}_I$ and (3.8b) to the vector of observations $\mathbf{z}_{II}$ to obtain $\mathbf{L}^{-1}\mathbf{y}$ and $|\mathbf{L}|$. We apply the same algorithm with starting conditions $\mathbf{x}(d+1 \mid d) = 0$ and (3.8b) to the columns of the matrix $\mathbf{C}$ to obtain $\mathbf{L}^{-1}\mathbf{C}$. Then the **QR** algorithm, or a slight modification of it, is used to obtain $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C} = (\mathbf{E}', 0')'$. Finally, we compute $\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}$ and S^* in (3.15). Minimizing S^* yields the estimators $(\hat{\phi}, \hat{\theta})$; then we estimate $\mathbf{z}_J$ and σ^2 by (3.14) and (3.16).

3.4 Interpolation and Prediction

Suppose that first we concentrate σ^2 and $\mathbf{z}_J$ out of the likelihood. For interpolation and prediction, we need the estimator (3.14) in (3.8a) to start the Kalman filter and the FPS. When $r_C = d - k$, we saw that $\hat{\mathbf{z}}_J$ could be obtained with (3.14b).

If $r_C < d - k$, then, as stated in the previous section, by permuting the variables in $\mathbf{z}_J$, if necessary, we can assume without loss of generality that $\mathbf{E} = (\mathbf{R}, \mathbf{S})$, where $\mathbf{R}$ is an upper triangular matrix. A generalized inverse of $\mathbf{E}'\mathbf{E}$ is the $(d-k) \times (d-k)$ matrix $(\mathbf{E}'\mathbf{E})^{-} = (\mathbf{T}_1', \mathbf{T}_2')'$, where $\mathbf{T}_1$ and $\mathbf{T}_2$ are the $r_C \times (d-k)$ and $(d-k-r_C) \times (d-k)$ submatrices such that $\mathbf{T}_1 = ((\mathbf{R}'\mathbf{R})^{-1}, 0)$ and $\mathbf{T}_2 = (0, 0)$. If $\mathbf{z}_J = (\mathbf{z}_J^{*\prime}, \mathbf{z}_J^{**\prime})'$, where $\mathbf{z}_J^*$ and $\mathbf{z}_J^{**}$ are the subvectors of $\mathbf{z}_J$ of dimension r_C and $d - k - r_C$, then (3.14a) implies $\hat{\mathbf{z}}_J^* = \mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} - \mathbf{R}^{-1}\mathbf{S}\tilde{\mathbf{x}}$ and $\hat{\mathbf{z}}_J^{**} = \tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}}$ is an arbitrary vector of dimension $d - k - r_C$. It is interesting to note that the set $\mathbf{R}\mathbf{z}_J^* + \mathbf{S}\mathbf{z}_J^{**}$ is a maximal set of independent linear combinations that can be estimated from the data y without dependence on $\tilde{\mathbf{x}}$.

The covariance matrix $\mathrm{var}(\hat{\mathbf{z}}_J)$ of $\hat{\mathbf{z}}_J$ is obtained by GLS. We defined in Section 3.3 the vectors $\mathbf{B}'(s)$ and $\mathbf{C}'(s)$ as the subvectors of $\mathbf{A}'(s)$ such that $\mathbf{A}'(s)\mathbf{z}_* = \mathbf{B}'(s)\mathbf{z}_I + \mathbf{C}'(s)\mathbf{z}_J$, $s > 0$. Let $1 \le t < t_k$, and let $z(t)$ be an unobserved value that we want to estimate. Because $z(t) = \mathbf{A}'(t)\mathbf{z}_* = \mathbf{C}'(t)\mathbf{z}_J$, we have

$$\hat{z}(t) = \mathbf{C}'(t)\hat{z}_J. \quad (3.17a)$$

The MSE of $\hat{z}(t)$ is $\mathbf{C}'(t)\mathrm{var}(\hat{\mathbf{z}}_J)\mathbf{C}(t)$. Let now $z(t) = \mathbf{B}'(t)\mathbf{z}_I + \mathbf{C}'(t)\mathbf{z}_J + \tilde{u}(t)$ be an unobserved value that we want to estimate with $t > t_k$. If $t < t_M$, we are interpolating; if $t > t_M$, we are predicting. Following Kohn and Ansley (1985), we have

$$\hat{z}(t) = \mathbf{B}'(t)\mathbf{z}_I + \mathbf{C}'(t)\hat{\mathbf{z}}_J + \mathbf{P}(t)(\mathbf{y} - \mathbf{C}\hat{\mathbf{z}}_J), \quad (3.17b)$$

where $E(\tilde{u}(t)|\tilde{\mathbf{u}}) = \mathbf{P}(t)\tilde{\mathbf{u}} = \mathrm{cov}(\tilde{u}(t), \tilde{\mathbf{u}})\mathrm{var}^{-1}(\tilde{\mathbf{u}})\tilde{\mathbf{u}}$. The MSE of $\hat{z}(t)$ is given by

$$E(z(t) - \hat{z}(t))^2 = (\mathbf{P}(t)\mathbf{C} - \mathbf{C}'(t))\mathrm{var}(\hat{\mathbf{z}}_J) \times (\mathbf{P}(t)\mathbf{C} - \mathbf{C}'(t))' + S(t), \quad (3.18)$$

where $S(t) = E(\tilde{u}(t) - \mathbf{P}(t)\tilde{\mathbf{u}})^2$. The Kalman filter with equations (2.6a) and (2.9) and starting conditions (3.8) if $t > t_M$, or the FPS if $t < t_M$, can be used to compute both $\mathbf{P}(t)(\mathbf{y} - \mathbf{C}\hat{\mathbf{z}}_J)$ and $S(t)$. The same algorithm applied to the columns of $\mathbf{C}$ yields $\mathbf{P}(t)\mathbf{C}$.

If $r_C < d - k$, then it may happen that $\mathbf{C}'(t)\hat{\mathbf{z}}_J = \mathbf{C}'(t)(\hat{\mathbf{z}}_J^{*\prime}, \hat{\mathbf{z}}_J^{**\prime})' = \mathbf{C}'(t)(\hat{\mathbf{z}}_J^{*\prime}, \tilde{\mathbf{x}}')'$ in (3.17a) or that (3.17b) will depend on the arbitrary vector $\tilde{\mathbf{x}}$. The vector $\mathbf{C}'(t)$ lies in the space generated by the rows of the matrix $(\mathbf{R}, \mathbf{S})$, which is the same as the space generated by the rows of the matrix $\mathbf{C}$, if and only if there exists a vector $\mathbf{c}'(t)$ such that $\mathbf{C}'(t) = \mathbf{c}'(t)(\mathbf{R}, \mathbf{S})$. Therefore, the vector $\mathbf{C}'(t)\hat{\mathbf{z}}_J = \mathbf{C}'(t) \times [\hat{\mathbf{z}}_J^{*\prime}, \tilde{\mathbf{x}}']' = \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y})', 0']' + \mathbf{C}'(t)[(-\mathbf{R}^{-1}\mathbf{S})', \mathbf{I}_{d-k-r_C}']'\tilde{\mathbf{x}}$ will not depend on $\tilde{\mathbf{x}}$ if and only if $\mathbf{C}'(t)$ lies in the space generated by the rows of the matrix $(\mathbf{R}, \mathbf{S})$. Note that if $r_C < d - k$, then the matrix $\mathrm{var}(\hat{\mathbf{z}}_J)$ should be replaced by $\hat{\sigma}^2(\mathbf{E}'\mathbf{E})^{-}$, and that this makes sense only if $\mathbf{C}'(t)\hat{\mathbf{z}}_J$ does not depend on $\tilde{\mathbf{x}}$. Thus a necessary and sufficient condition for an unobserved value $z(t)$ to be interpolated or predicted from the data without further information (or, equivalently, without dependence on $\tilde{\mathbf{x}}$) is that its $\mathbf{C}'(t)$ vector lie in the space generated by the rows of $\mathbf{C}$.

Suppose now that we concentrate only σ^2 out of the likelihood and that we jointly estimate $\mathbf{z}_J$ and (ϕ, θ). We would use the estimator $\hat{\mathbf{z}}(t)$ thus obtained in (3.17). But in (3.18) we would only obtain $S(t)$, and thus the MSE of $\hat{z}(t)$ would be conditional on $\mathbf{z}_J$.

3.5 The Relationship Between Our Approach and that of Kohn and Ansley

The following theorem specifies the precise relationship between our likelihood and that of Kohn and Ansley (1986). The proof is given in Appendix A.

Theorem 1. Concentrating σ^2 and $\mathbf{z}_J$ out of our likelihood (3.11), the same function to minimize is obtained as when σ^2 is concentrated out of the likelihood defined by Kohn and Ansley.

It is shown in the proof of Theorem 1 that the only difference between our estimator of σ^2 and Kohn and Ansley's estimator of σ^2 lies in the denominator. By Lemma 4 in Appendix A, $M - k - r_C$ is the denominator corresponding to the approach of Kohn and Ansley, whereas our denominator is $M - k$. The next two theorems show that the estimators of unobserved values, as well as their MSE's, obtained by the approach of Kohn and Ansley (1986) and those obtained by our approach coincide. The proofs are given in Appendix A.

Theorem 2. If $z(t)$ is an unobserved value, then $z(t)$ is estimable in the sense of Kohn and Ansley if and only if our estimator $\hat{z}(t)$ does not depend on the arbitrary vector $\tilde{\mathbf{x}}$.

Theorem 3. If $z(t)$ is an unobserved value estimable in the sense of Kohn and Ansley, then the estimator $\bar{z}(t)$ obtained by the method of Kohn and Ansley and by our estimator coincide. If the same estimator of σ^2 is used, then the MSE's are also equal.

3.6 Regression Model With ARIMA Errors and Missing Values

Consider the regression model (2.10a), where the vectors β and $\mathbf{y}(t)$ are as in Section 2.4 and the residuals $\{\nu(t)\}$ follow the ARIMA model (3.1), with $z(t)$ replaced by $\nu(t)$. Defining the vectors $\mathbf{z}$ and ν and the matrix $\mathbf{Y}$ as in Section 2.4, Equation (2.10b) still holds and, similarly to the stationary case, we can proceed in two ways. First, the β parameters, the initial missing values $\mathbf{z}_J$, and the parameters of the ARIMA model can be jointly estimated. The state-space representation would be given by Equation (2.6a), with $\mathbf{z}$ replaced by ν, and Equation (2.11). The starting conditions would be (3.8b) and $\mathbf{x}(d+1\mid d) = [\mathbf{A}(d+1), \mathbf{A}(d+2), \ldots, \mathbf{A}(d+r)]'(\mathbf{z}_{**} - \mathbf{Y}_*\beta)$, where $\mathbf{z}_{**}$ is the vector formed by the first k observed values $(\mathbf{z}_I)$ and some starting values for the missing observations in the first d values of the series $(\mathbf{z}_J)$, and $\mathbf{Y}_*$ denotes the $d \times h$ matrix formed by the rows $\mathbf{y}'(t)$, $t = 1, \ldots, d$. The second procedure concentrates σ^2, β, and $\mathbf{z}_J$ out of the likelihood function and is analogous to that described in Section 2.4. Because $\{\nu(t)\}$ follows the ARIMA model (3.1), with the notation of Section 3.3, we can write $\nu_{II} = \mathbf{B}\nu_I + \mathbf{C}\nu_J + \tilde{\mathbf{u}}$, where ν_{II}, ν_I, and ν_J are the vectors of errors corresponding to the vectors of observations $\mathbf{z}_{II}$, $\mathbf{z}_I$, and $\mathbf{z}_J$ defined at the beginning of Section 3.3. Let $\mathbf{Y}_I$, $\mathbf{Y}_{II}$, and $\mathbf{Y}_J$ be the matrices with rows the vectors $\mathbf{y}'(t)$ corresponding to the vectors ν_I, ν_{II}, and ν_J. Replacing ν_I by $\mathbf{z}_I - \mathbf{Y}_I\beta$ and ν_J by $\mathbf{z}_J - \mathbf{Y}_J\beta$ in the preceding expression, the following regression model is obtained:

$$\mathbf{z}_{II} = \mathbf{B}\mathbf{z}_I + \mathbf{C}\mathbf{z}_J + \mathbf{Y}_{II}\beta - \mathbf{B}\mathbf{Y}_I\beta - \mathbf{C}\mathbf{Y}_J\beta + \tilde{\mathbf{u}},$$

where the regression parameters are z_J and β. Letting $\mathbf{y} = \mathbf{z}_{II} - \mathbf{B}\mathbf{z}_I$, this model can be rewritten as

$$\mathbf{y} = [\mathbf{C}, \mathbf{Y}_{II} - \mathbf{B}\mathbf{Y}_I - \mathbf{C}\mathbf{Y}_J][\mathbf{z}_J', \beta']' + \tilde{\mathbf{u}}$$
$$= [\mathbf{C}, \mathbf{Y}_{II} - \mathbf{A}_{II}\mathbf{Y}_*][\mathbf{z}_J', \beta']' + \tilde{\mathbf{u}},$$

where $\mathbf{A}_{II}$ is the matrix defined by $\mathbf{B}\mathbf{Y}_I + \mathbf{C}\mathbf{Y}_J = \mathbf{A}_{II}\mathbf{Y}_*$. The Kalman filter applied to the model $\mathbf{y} = \mathbf{z}_{II} - \mathbf{B}\mathbf{z}_I = \tilde{\mathbf{u}}$ yields $\mathbf{L}^{-1}\mathbf{y}$ and $|\mathbf{L}|$, where $\mathbf{L}$ is as in Section 3.3. The starting conditions would be (3.8b) and $\mathbf{x}(d+1\mid d) = [\mathbf{B}(d+1), \ldots, \mathbf{B}(d+r)]'\mathbf{z}_I$. Note that, as in Section 3.3, it is not necessary to compute the vector $\mathbf{y}$ and that the Kalman filter is applied to the vector of observations $\mathbf{z}_{II}$. The same algorithm applied to the columns of the matrix $[\mathbf{C}, \mathbf{Y}_{II} - \mathbf{A}_{II}\mathbf{Y}_*]$ yields the product of $\mathbf{L}^{-1}$ by this matrix. The starting conditions would be (3.8b) and $\mathbf{x}(d+1\mid d) = 0$. Then the **QR** algorithm can be applied to the transformed model

$$\mathbf{L}^{-1}\mathbf{y} = \mathbf{L}^{-1}[\mathbf{C}, \mathbf{Y}_{II} - \mathbf{A}_{II}\mathbf{Y}_*][\mathbf{z}_J', \beta']' + \mathbf{L}^{-1}\tilde{\mathbf{u}},$$

and we can proceed as in Section 2.4.

3.7 Final Remarks

The approach that we have presented in this article leads to simple and efficient algorithms for likelihood evaluation, interpolation, and prediction for ARIMA models with missing data. We summarize its merits:

1. The definition of the likelihood is intuitive and does not require any theoretical development. The state-space representation is also intuitive, and initial conditions for the Kalman filter are directly obtained. No computations are necessary for $t = 1, 2, \ldots, d$. Proofs are short and direct.

2. The case of a nonstationary series with missing data is reduced to the case of a regression model with errors having a known covariance structure. Therefore, the efficient method of Kohn and Ansley (1985) for GLS models can be applied. The method can be extended without modification to handle regression parameters.

3. The algorithm for likelihood evaluation consists of a computationally trivial extension of the Kalman filter plus the **QR** algorithm. This extension consists of replacing the vector recursion for the state estimate with a matrix recursion (i.e., one state for the data and one state for each of the columns of the regression matrix). Also, we need to obtain at each iteration, in addition to the residual for the data, one residual for each of the columns of the regression matrix. This amounts to replacing one scalar recursion with a vector recursion. The algorithm is easy to program, can incorporate regression parameters without modification, and is conceptually simple and computationally efficient compared to the modified Kalman filter of Ansley and Kohn (1985) (see Kohn and Ansley 1986, pp. 754–755).

4. For prediction, the same trivial extension of the Kalman filter is used; for interpolation, the simple version of the FPS given in Section 2.3 is used. This represents an advantage compared to using the modified FPS of Ansley and Kohn (1985), as becomes apparent from an inspection of the equations in the two algorithms (see Kohn and Ansley 1986, pp. 759–760). Also, the storage requirements are minimal.

4. APPLICATION

We have written a program in Fortran, available for mainframes and PC's under MS-DOS, that can be obtained directly from us. This program performs estimation, forecasting, and interpolation of regression models with missing observations and ARIMA errors (see Gomez and Maravall 1992). Recently, three more capabilities have been added to the program. Automatic ARIMA modeling, automatic specification and correction of four kinds of outlier effects, and handling of missing data by means of intervention analysis.

The regression variables can be inputed by the user or generated by the program. The variables that can be gener-

Table 1. Parameter Estimates and Standard Errors

Data set	θ	Θ	σ
1. (N = 144)	−.402 (.080)	−.557 (.084)	.037
2. (N = 78)	−.457 (.096)	−.758 (.227)	.042
3. (N = 139)	−.405 (.081)	−.566 (.083)	.037
4. (N = 130)	−.430 (.081)	−.573 (.085)	.037
5. (N = 130)	−.401 (.082)	−.563 (.089)	.037

Table 2. 1961 Forecasts and RMSE's (in Parentheses) for Logged Data

Data set	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1. (N = 144)	6.110 (.037)	6.054 (.043)	6.172 (.049)	6.199 (.053)	6.233 (.058)	6.369 (.062)	6.507 (.066)	6.503 (.069)	6.325 (.073)	6.209 (.076)	6.064 (.079)	6.168 (.082)
2. (N = 78)	6.084 (.053)	6.091 (.059)	6.247 (.064)	6.205 (.069)	6.199 (.073)	6.308 (.077)	6.409 (.080)	6.414 (.083)	6.299 (.086)	6.174 (.089)	6.043 (.091)	6.174 (.087)
3. (N = 139)	6.110 (.038)	6.054 (.044)	6.173 (.049)	6.199 (.054)	6.232 (.058)	6.367 (.062)	6.497 (.068)	6.503 (.070)	6.325 (.073)	6.209 (.077)	6.064 (.080)	6.168 (.083)
4. (N = 130)	6.111 (.037)	6.055 (.043)	6.174 (.048)	6.200 (.053)	6.233 (.057)	6.368 (.061)	—* (.080)	6.503 (.068)	6.326 (.071)	6.209 (.074)	6.064 (.077)	6.169 (.080)
5. (N = 130)	—* (.058)	6.055 (.043)	6.172 (.048)	6.199 (.053)	6.232 (.057)	6.369 (.061)	6.507 (.065)	6.503 (.069)	6.325 (.072)	6.209 (.075)	6.064 (.078)	6.168 (.081)

* Value depends on a free parameter.

ated are trading day and easter effect (see Hillmer, Bell, and Tiao 1983) and certain types of intervention variables (see Box and Tiao 1975). Estimation of the regression parameters (including the missing observations among the first d values of the series) plus the ARIMA model parameters can be made by concentrating them out of the likelihood or by joint estimation, as previously described. Several algorithms can be used for computing the likelihood or, more precisely, the nonlinear sum of squares to be minimized. When the differenced series can be used, we use the algorithm of Morf, Sidhu, and Kailath (1974), as discussed by Pearlman (1980) and improved by Mélard (1984).

For the nondifferenced series, it is possible to use the ordinary Kalman filter, as described in this article (default option), or its square root version (see Anderson and Moore 1979). The latter is adequate when numerical difficulties arise; however, it is markedly slower and does not permit (at present) concentrating the regression parameters out of the likelihood. By default, the exact maximum likelihood method is used; the unconditional least squares method is available as an option. Nonlinear maximization of the likelihood function and computation of the parameter estimates standard errors is made using Marquardt's method and first numerical derivatives.

For forecasting and interpolation, the ordinary Kalman filter or the square root filter options are available. Interpolation of missing values is made with the simplified FPS, as described herein. When concentrating z_J and β out of the likelihood, MSE's of the forecasts and interpolations are obtained following the approach of Kohn and Ansley (1985), also as described herein. If the rank of the **C** matrix is smaller than $d - k$, the program indicates which initial missing values are free parameters (flagging the elements of the vector $\hat{\mathbf{z}}_J^{**} = \tilde{\mathbf{x}}$ in Sec. 3.4) and also which forecasts or interpolations will depend on the arbitrary vector $\tilde{\mathbf{x}}$. The user can then assign arbitrary values (typically, very large or very small) to the free parameters and then rerun the program. Proceeding in this way, all parameters of the ARIMA model can be estimated because, as seen in Sections 3.3 and 3.4, the function to minimize does not depend on the free parameters. Moreover, it will be evident which forecasts and interpolations are affected by these arbitrary values, because these will strongly deviate from the rest of the estimates. But if all unknown parameters are jointly estimated, the program may not flag all free parameters. It may happen (as in data set 5 in the following paragraph) that there is convergence to a valid arbitrary set of solutions for $\mathbf{z}_J$ (i.e., that some linear combinations of the initial missing observations, including the free parameters, are estimable.)

Following the tradition set up by Harvey and Pierse (1984) and Kohn and Ansley (1986), we apply our procedure to the series of 144 monthly observations on international airline passengers (January 1949–December 1960), given and analyzed by Box and Jenkins (1976, chap. 9). The model identified in all three cases is the multiplicative ARIMA (0, 1, 1) × (0, 1, 1) model

$$\nabla_{12}\nabla z_t = (1 + \theta B)(1 + \Theta B^{12})a_t,$$

applied to the logs of the data. We consider the same four data sets of Kohn and Ansley (1986) and add a fifth data set:

- data set 1: all 144 observations
- data set 2: the 78 observations that result from treating January through November in the last 6 years as missing
- data set 3: the 139 observations that result from treating July 1949, June, July, and August 1957, and July 1960 as missing
- data set 4: the 130 observations that result from treating all July values and June and August 1957 as missing
- data set 5: the 130 observations that result from treating all January values and February 1951 and February 1954 as missing.

To facilitate the comparison of our results with those of Kohn and Ansley (1986), we use the estimator of σ^2 of Ansley

Table 3. Data Set 2 (N = 78) Interpolation for January-November 1957: Estimates and RMSE's (in Parentheses)

	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sep.	Oct.	Nov.
Estimate	5.733 (.046)	5.738 (.050)	5.893 (.053)	5.850 (.055)	5.843 (.056)	5.951 (.056)	6.051 (.056)	6.055 (.055)	5.938 (.053)	5.812 (.050)	5.680 (.046)
Actual	5.753	5.707	5.875	5.852	5.872	6.045	6.146	6.146	6.001	5.849	5.720

Table 4. Data Set 3 ($N = 139$): Estimates and RMSE's (in Parentheses) for Logarithms of Data

	July 1949	June 1957	July 1957	Aug. 1957	July 1960
Estimate	5.013	6.024	6.147	6.148	6.409
	(.031)	(.030)	(.031)	(.030)	(.032)
Actual	4.997	6.045	6.146	6.146	6.433

Table 5. Data Set 4 ($N = 130$): Estimates and RMSE's (in Parentheses) for Logarithms of Data

	June 1957	Aug. 1957
Estimate	6.023	6.147
	(.030)	(.030)
Actual	6.045	6.146

and Newbold (1981), where the denominator is $M - k - r_C - 2$, because there are two model parameters to estimate. Note that data set 2 is the example considered by Harvey and Pierse (1984) and that, because there are not 13 consecutive observations available at either the start or the end of series in the last three data sets, their methodology cannot be applied in these three cases. In data set 3 only one initial observation (number 7) is missing ($k = 1$), and the **C** matrix has rank 1. Thus all missing observations are estimable. In data set 4 there is also one initial observation (number 7) missing ($k = 1$), and the **C** matrix has rank 0. Therefore, the initial missing observation will be a free parameter. Finally, in data set 5 there are two initial missing observations (numbers 1 and 13, $k = 2$) and the **C** matrix has rank 1. This happens because the $\mathbf{C}'(t)$ vectors corresponding to the observed values are all multiples of $(-1, 1)$ for $t > 13$. Thus, although $z(1)$ and $z(13)$ are not estimable, the linear combination $z(13) - z(1)$ can be estimated from the available observations. Therefore, if we assign (for example) the value 0 to $z(1)$ and rerun the program, then the interpolation of $z(13)$ will be the minimum MSE estimator of the annual difference $(z(13) - z(1))$. Note that the approach of Kohn and Ansley (1986) would simply tell us that observations 1 and 13 are not estimable.

Tables 1–5 are equivalent to Tables 1–5 of Kohn and Ansley (1986), and they present the results obtained with our program. Table 6 displays the results for data set 5. The differenced series was used for data set 1, and the ordinary Kalman filter and simplified FPS were applied to the nonstationary levels (as described herein) for data sets 2, 3, 4, and 5. The estimation method was always exact maximum likelihood, and in data sets 3 and 5 the initial missing values that are not free parameters (i.e., observation 7 in data set 3 and observation 1 in data set 5) were concentrated out of the likelihood function. As for the free parameters, we assigned value 10^{-9} to observation 7 in data set 4 and, as stated before, assigned value 1 to observation 13 in data set 5. Note that no adjustment in the denominator of σ^2 is needed, because we increase the number of observations by 1 and at the same time also increase k (the number of nonmissing initial values) by 1.

The results obtained with joint estimation of all parameters were practically identical (although CPU time increased by close to 15%), except for data set 5. In this case the program did not flag any free parameter, and there was convergence to a valid set of solutions for the initial missing values. Note that two valid sets of solutions are equal up to an additive constant (the free parameter in this case). Table 1 presents the estimates of the ARIMA model parameters and their standard errors. The estimate of θ is the same as the one in Kohn and Ansley (1986), except for a small difference in data set 3 ($-.405$ vs. $-.408$). The estimate of Θ is identical in all cases. Table 2 displays the 1–12-months-ahead forecasts (and their root MSE's [RMSE's]) for the five data sets. The results are practically identical to those obtained by Kohn and Ansley (1986); the only difference is the forecast for September in data set 4 (6.326 vs. 6.333). Given that the forecasts for September are 6.325 in data sets 1, 3, and 5, it is possible that the second forecast contains a typo. There is a discrepancy of .001 in the RMSE's of three forecasts (March for data set 1, October for data set 2, and October for data set 3) that could be attributed to Kohn and Ansley's use of N-14 instead of N-15 as denominator to compute their residual variance in data sets 1, 2, and 3 (although the rank of their A matrix is 13). Table 3 contains the 11 values interpolated for the months January–November 1957 for data set 2. Table 4 presents the estimates for the five missing values in data set 3. Table 5 shows the estimates of the missing values in data set 4 that do not depend on the free parameter. Finally, Table 6 shows the estimates of the missing values in data set 5 that do not depend on the free parameter and also the estimate $z(13) - z(1)$. The results in Tables 3, 4, and 5 are identical to those obtained by Kohn and Ansley (1982). For data sets 4 and 5, as mentioned before, to determine whether or not the estimator of an unobserved value depends on a free parameter, it may help to notice that the estimate value is very different from the other estimated values. But the precise condition to check is whether its $\mathbf{C}'(t)$ vector lies in the space generated by the **C** matrix. For example, we cannot predict January 1961 in data set 5 because its $\mathbf{C}'(t)$ vector is $(-11, 12)$, which does not lie in the space generated by the $(-1, 1)$ vector.

APPENDIX A: PROOFS

Proof of Lemma 2

From (3.5) it is immediately seen that the subspaces generated by $\{Z_*, z(d+1), \ldots, z(t)\}$ and by $\{Z_*, u(d+1), \ldots, u(t)\}$ are the same for $t \geq d + 1$. Thus $\hat{z}(t|t-1) = E(z(t)|z(t-1), \ldots, z(d+1), Z_*) = E(z(t)|u(t-1), \ldots, u(d+1), Z_*)$ and, considering (3.3),

Table 6. Data Set 5 ($N = 130$): Estimates and RMSE's (in Parentheses) for Logarithms of Data

	Jan. 1950– Jan. 1949	Feb. 1951	Feb. 1954
Estimate	.068	5.020	5.327
	(.040)	(.029)	(.028)
Actual	.026	5.011	5.236

$$\hat{z}(t|t-1) = \mathbf{A}'(t)\mathbf{z}_* + \sum_{i=0}^{t-d-1} \xi_i E(u(t-i)|u(t-1), \ldots, u(d+1), Z_*)$$
$$= \mathbf{A}'(t)\mathbf{z}_* + E(u(t)|u(t-1), \ldots, u(d+1)) + \sum_{i=1}^{t-d-1} \xi_i u(t-i)$$
$$= \mathbf{A}'(t)\mathbf{z}_* + \hat{u}(t|t-1) + \sum_{i=1}^{t-d-1} \xi_i u(t-i).$$

Hence $z(t) - \hat{z}(t|t-1) = u(t) - \hat{u}(t|t-1)$.

Proof of Lemma 3

Using (3.7) in (3.3), it is obtained that

$$z(t+r-1) - z(t+r-1|t)$$
$$= \xi_0(u(t+r-1) - u(t+r-1|t)) + \cdots + \xi_{r-2}(u(t+1) - u(t+1|t)).$$

Because $\{u(t)\}$ follows the model (2.1), we have $u(t+j) - u(t+j|t) = \sum_{i=0}^{j-1} \psi_i a(t+j-i)$, $j = 1, \ldots, r-1$, and hence, given that $\psi^*(B) = \xi(B)\psi(B)$,

$$z(t+r-1) - z(t+r-1|t)$$
$$= \xi_0\psi_0 a(t+r-1) + \cdots + (\xi_0\psi_{r-2} + \cdots + \xi_{r-2}\psi_0)a(t+1)$$
$$= \psi_0^* a(t+r-1) + \cdots + \psi_{r-2}^* a(t+1).$$

Therefore, $z(t+r-1|t) = z(t+r-1|t-1) + \psi_{r-1}^* a(t)$, and considering that $z(t+r-1|t-1) = -\sum_{i=1}^{r-1} \phi_{r-i}^* z(t+i-1|t-1) - \phi_r^* z(t-1)$, the lemma follows directly.

Proof of Theorem 1.

As we showed in Section 3.3, maximizing (3.11) concentrated with respect to $\mathbf{z}_J$ and σ^2 is the same as minimizing (3.15). Kohn and Ansley (1986) defined a likelihood coinciding with that of Box and Jenkins when there are no missing observations and with that of Harvey and Pierse (1984) when there are no missing observations among the first d values of the series. In these cases, therefore, their definition coincides with ours. For the general case, Kohn and Ansley defined $\eta = (z(1-d), \ldots, z(0))'$ and considered $z(t_j) = \mathbf{F}'(t_j)\eta + w(t_j), j = 1, \ldots, M$, where $w(s) = \sum_{j=0}^{t-1} \xi_j u(t-j)$ and the $F'(s)$, $s \geq 1$, are generated similarly to the $\mathbf{A}'(s)$ of (3.3) and (3.4). If $\mathbf{z} = (z(t_1), \ldots, z(t_M))'$, $\mathbf{w} = (w(t_1), \ldots, w(t_M))'$, and $\mathbf{F}$ is the matrix having as rows the vectors $\mathbf{F}'(s)$, $s = t_1, \ldots, t_M$, then the previous equations can be written as

$$\mathbf{z} = [\mathbf{z}_I', \mathbf{z}_{II}']' = \mathbf{F}\eta + \mathbf{w}. \quad (A.1)$$

To see the relationship between expressions (A.1) and (3.9), let $\mathbf{F}_*$ be the $d \times d$ matrix having as rows the vectors $\mathbf{F}'(s)$, $s = 1, \ldots, d$, and let $\mathbf{w}_* = (w(1), \ldots, w(d))'$. By definition, equality $\mathbf{z}_* = \mathbf{F}_*\eta + \mathbf{w}_*$ holds. Substituting in (3.9), it is obtained that

$$\mathbf{z} = [\mathbf{z}_I', \mathbf{z}_{II}']' = \mathbf{A}\mathbf{F}_*\eta + \mathbf{A}\mathbf{w}_* + [0', \tilde{\mathbf{u}}']'. \quad (A.2a)$$

Therefore, considering (A.1), the following equalities are found:

$$\mathbf{F} = \mathbf{A}\mathbf{F}_*, \qquad \mathbf{A}\mathbf{w}_* + [0', \tilde{\mathbf{u}}']' = \mathbf{w}. \quad (A.2b)$$

Further, in (3.9) matrix $\mathbf{D}$ can be obtained from matrix $\mathbf{A}$ by permuting some of its columns. Thus the following equality also holds:

$$\mathbf{D} = \mathbf{A}\mathbf{T}_A, \quad (A.3a)$$

where $\mathbf{T}_A$ is the nonsingular matrix obtained from the unit matrix by permuting the columns in the same way. Then we can rewrite Equation (A.2a) as

$$\mathbf{z} = [\mathbf{z}_I', \mathbf{z}_{II}']' = \mathbf{D}\mathbf{T}_A^{-1}\mathbf{F}_*\eta + \mathbf{D}\mathbf{T}_A^{-1}\mathbf{w}_* + [0', \tilde{\mathbf{u}}']'. \quad (A.3b)$$

The transformation that permits us to move from (3.10) to (3.13) will not have a unit determinant, because $|\mathbf{L}^{-1}|$ generally is different than 1. Multiplying $\mathbf{L}^{-1}$ by a positive number α, such that $|\alpha\mathbf{L}^{-1}| = 1$, it is immediately seen that when $\alpha = |\mathbf{L}|^{1/(M-k)}$, the matrix $\alpha\mathbf{Q}'\mathbf{L}^{-1}$ has a unit determinant. Premultiplying both sides of (A.3b) by

$$\mathbf{J} = \begin{bmatrix} \mathbf{I}_k & 0 \\ -\alpha\mathbf{Q}'\mathbf{L}^{-1}\mathbf{B} & \alpha\mathbf{Q}'\mathbf{L}^{-1} \end{bmatrix},$$

which also has a unit determinant, it is obtained that

$$\mathbf{J}\begin{bmatrix} \mathbf{z}_I \\ \mathbf{z}_{II} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_k & 0 \\ 0 & \alpha\mathbf{E} \\ 0 & 0 \end{bmatrix} \mathbf{T}_A^{-1}\mathbf{F}_*\eta + \begin{bmatrix} \mathbf{I}_k & 0 \\ 0 & \alpha\mathbf{E} \\ 0 & 0 \end{bmatrix} \mathbf{T}_A^{-1}\mathbf{w}_* + \begin{bmatrix} 0 \\ \alpha\mathbf{Q}_1'\mathbf{L}^{-1}\tilde{\mathbf{u}} \\ \alpha\mathbf{Q}_2'\mathbf{L}^{-1}\tilde{\mathbf{u}} \end{bmatrix}. \quad (A.4)$$

Premultiplying (A.1) by $\mathbf{J}$ yields

$$\mathbf{J}\begin{bmatrix} \mathbf{z}_I \\ \mathbf{z}_{II} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_k & 0 \\ 0 & \alpha\mathbf{E} \\ 0 & 0 \end{bmatrix} \mathbf{T}_A^{-1}\mathbf{F}_*\eta + \mathbf{J}\mathbf{w}. \quad (A.5)$$

Let $\mathbf{J}_1 = (-\alpha\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{B}, \alpha\mathbf{Q}_1'\mathbf{L}^{-1})$ and $J_2 = (-\alpha\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{B}, \alpha\mathbf{Q}_2'\mathbf{L}^{-1})$. Equating (A.4) and (A.5), it follows that

$$\mathbf{J}_2[\mathbf{z}_I', \mathbf{z}_{II}']' = \alpha\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y} = \alpha\mathbf{Q}_2'\mathbf{L}^{-1}\tilde{\mathbf{u}} = \mathbf{J}_2\mathbf{w}. \quad (A.6)$$

Because $\mathbf{J}$ is a transformation matrix of the type defined by Kohn and Ansley (1986), their likelihood is the density of $\mathbf{J}_2\mathbf{w}$, which, from (A.6), is equal to that of $\alpha\mathbf{Q}_2'\mathbf{L}^{-1}\tilde{\mathbf{u}}$. This latter likelihood is given by

$$(2\pi\sigma^2)^{-(M-k-r_C)/2}|\mathbf{L}|^{-(M-k-r_C)/(M-k)} \times \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}\right). \quad (A.7)$$

Concentrating σ^2 out of this likelihood using

$$\hat{\sigma}^2 = (1/(M-k-r_C))\mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}, \quad (A.8)$$

it is seen that maximizing (A.7) is equivalent to minimizing $|\mathbf{L}|^{1/(M-k)}\mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2\mathbf{L}^{-1}\mathbf{y}|\mathbf{L}|^{1/(M-k)}$, which is identical to (3.15).

Lemma 4. Using the same notation as in the proof of Theorem 1, the matrix $\mathbf{F}_*$ is nonsingular and $\text{rk}(\mathbf{D}) = \text{rk}(\mathbf{F}) = \text{rk}(\mathbf{A}) = k + r_C$.

Proof of Lemma 4

It is easy to check that $\text{rank}(\mathbf{D}) = k + r_C$. As we saw in the proof of Theorem 1, $\mathbf{F} = \mathbf{A}\mathbf{F}_*$ and $\mathbf{D} = \mathbf{A}\mathbf{T}_A$. Hence we need only show that $\mathbf{F}_*$ is nonsingular. The rows $\mathbf{F}'(t)$, $t = 1, \ldots, d$, of $\mathbf{F}_*$ satisfy $\mathbf{F}'(t) = -\delta_1\mathbf{F}'(t-1) - \cdots - \delta_d\mathbf{F}'(t-d)$. Letting $t = d$, we have $\mathbf{F}'(d) = -\delta_1\mathbf{F}'(d-1) - \cdots - \delta_{d-1}\mathbf{F}'(1) - \delta_d\mathbf{F}'(0)$ and $-\delta_d\det(\mathbf{F}_*) = -\delta_d\det[\mathbf{F}(1), \ldots, \mathbf{F}(d-1), \mathbf{F}(0)]$. Letting $t = d-1$, we have $\mathbf{F}'(d-1) = -\delta_1\mathbf{F}'(d-2) - \cdots - \delta_{d-1}\mathbf{F}'(0) - \delta_d\mathbf{F}'(-1)$ and $\delta_d^2\det(\mathbf{F}_*) = \delta_d^2\det[\mathbf{F}(1), \ldots, \mathbf{F}(d-2), \mathbf{F}(-1), \mathbf{F}(0)]$. Proceeding in a similar manner, it is obtained that $(-\delta_d)^d\det(\mathbf{F}_*) = (-\delta_d)^d\det[\mathbf{F}(1-d), \ldots, \mathbf{F}(-1), \mathbf{F}(0)] = (-\delta_d)^d \neq 0$, because $|\delta_d| = 1$.

Proof of Theorem 2

Using the same notation as in the proof of Theorem 1, let

$$z(t) = \mathbf{F}'(t)\eta + w(t) = \mathbf{A}'(t)\mathbf{z}_* + \tilde{u}(t)$$
$$= \mathbf{B}'(t)\mathbf{z}_I + \mathbf{C}'(t)\mathbf{z}_J + \tilde{u}(t), \qquad t \geq 1, \quad (A.9)$$

where $\tilde{\mathbf{u}}(t) = 0$ if $t < t_k$ is an unobserved value that we want to estimate. Kohn and Ansley defined $x(t)$ as estimable if the vector

$\mathbf{F}'(t)$ lies in the space generated by the rows of the matrix $\mathbf{F}$. This happens if and only if there exists a vector $\mathbf{d}'(t)$ such that $\mathbf{F}'(t) = \mathbf{d}'(t)\mathbf{F}$. We now show that $\mathbf{F}'(t)$ lies in the space generated by the rows of the matrix $\mathbf{F}$ if and only if $\mathbf{C}'(t)$ lies in the space generated by the rows of the matrix $\mathbf{C}$. By definition, $\mathbf{z}_* = \mathbf{F}_*\eta + \mathbf{w}_*$; by Lemma 4, this implies $\eta = \mathbf{F}_*^{-1}\mathbf{z}_* - \mathbf{F}_*^{-1}\mathbf{w}_*$. Substituting in (A.9), we have $\mathbf{F}'(t)\eta + w(t) = \mathbf{F}'(t)\mathbf{F}_*^{-1}\mathbf{z}_* + w(t) - \mathbf{F}_*^{-1}\mathbf{w}_* = \mathbf{A}'(t)\mathbf{z}_* + \tilde{u}(t)$ and, therefore, $\mathbf{F}'(t)\mathbf{F}_*^{-1} = \mathbf{A}'(t)$. Given that the vector $(\mathbf{B}'(t), \mathbf{C}'(t))$ is obtained from $\mathbf{A}'(t)$ by a permutation, the following equality holds:

$$\mathbf{F}'(t)\mathbf{F}_*^{-1}\mathbf{T}_A = (\mathbf{B}'(t), \mathbf{C}'(t)). \quad (A.10)$$

Suppose that there exists a vector $\mathbf{d}'(t)$ such that $\mathbf{F}'(t) = \mathbf{d}'(t)\mathbf{F}$. By (A.2b), (A.3a), and (A.10), this implies $(\mathbf{B}'(t), \mathbf{C}'(t)) = \mathbf{d}'(t)\mathbf{D}$. If we define $\mathbf{d}'(t) = (\mathbf{b}'(t), \mathbf{c}'(t))$, where $\mathbf{b}'(t)$ and $\mathbf{c}'(t)$ are the subvectors of $\mathbf{d}'(t)$ of dimension $1 \times k$ and $1 \times (M-k)$, we finally obtain $\mathbf{C}'(t) = \mathbf{c}'(t)\mathbf{C}$. To prove the "if" part, assume that there exists a vector $\mathbf{c}'(t)$ such that $\mathbf{C}'(t) = \mathbf{c}'(t)\mathbf{C}$, and define $\mathbf{b}'(t) = \mathbf{B}'(t) - \mathbf{c}'(t)\mathbf{B}$ and $\mathbf{d}'(t) = (\mathbf{b}'(t), \mathbf{c}'(t))$. It is easy to check that $\mathbf{d}'(t)\mathbf{D} = (\mathbf{B}'(t), \mathbf{C}'(t))$; therefore, by (A.10), (A.2b), and (A.3a), we have $\mathbf{F}'(t) = \mathbf{d}'(t)\mathbf{F}$. Because we showed in Section 3.4 that the vector $\mathbf{C}'(t)\hat{\mathbf{z}}_J$ does not depend on the arbitrary vector $\tilde{\mathbf{x}}$ if and only if $\mathbf{C}'(t)$ lies in the space generated by the rows of $\mathbf{C}$, the theorem is proved.

Proof of Theorem 3

The estimator of Kohn and Ansley (1986) can be expressed as $\bar{z}(t) = \mathbf{d}'(t)\mathbf{z} + E[w(t) - \mathbf{d}'(t)\mathbf{w} \mid \mathbf{J}_2\mathbf{w}]$, where $\mathbf{w}$ and $\mathbf{J}_2\mathbf{w}$ are those of (A.1) and (A.6) and $\mathbf{d}'(t)$ is a vector such that $\mathbf{F}'(t) = \mathbf{d}'(t)\mathbf{F}$. This vector exists because $z(t)$ is estimable. As we showed in the proof of Theorem 2, $\mathbf{d}'(t)$ also satisfies $(\mathbf{B}'(t), \mathbf{C}'(t)) = \mathbf{d}'(t)\mathbf{D}$. This, by (A.1) and (3.9), implies

$$\begin{aligned} z(t) &= \mathbf{d}'(t)\mathbf{z} + w(t) - \mathbf{d}'(t)\mathbf{w} \\ &= \mathbf{d}'(t)\mathbf{z} + \tilde{u}(t) - \mathbf{d}'(t)[0', \tilde{\mathbf{u}}']' \end{aligned} \quad (A.11)$$

and, therefore,

$$w(t) - \mathbf{d}'(t)\mathbf{w} = \tilde{u}(t) - \mathbf{d}'(t)[0', \tilde{\mathbf{u}}']'. \quad (A.12)$$

As stated previously, by permuting the variables in $\mathbf{z}_J$, if necessary, we can assume without loss of generality that $\mathbf{E} = (\mathbf{R}, \mathbf{S})$ in Equation (3.13a), where $\mathbf{R}$ is an upper triangular matrix with nonzero elements in the main diagonal. If $\mathbf{z}_J = (\mathbf{z}_J^{*\prime}, \mathbf{z}_J^{**\prime})'$, as in Section 3.4, then from Equation (3.13a) we have $\mathbf{z}_J^* = \mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y} - \mathbf{R}^{-1}\mathbf{S}\mathbf{z}_J^{**} - \mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\tilde{\mathbf{u}}$. Substituting in Equation (A.9), it is obtained that

$$\begin{aligned} z(t) &= \mathbf{B}'(t)\mathbf{z}_I + \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y})', 0']' \\ &\quad + \mathbf{C}'(t)[-(\mathbf{R}^{-1}\mathbf{S})', \mathbf{I}]'\mathbf{z}_J^{**} + \tilde{u}(t) \\ &\quad - \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\tilde{\mathbf{u}})', 0']'. \end{aligned} \quad (A.13)$$

The third term on the right is 0 because $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C} = ((\mathbf{R}, \mathbf{S})', 0')'$ and, as we showed in the proof of Theorem 2, if $\mathbf{d}'(t) = (\mathbf{b}'(t), \mathbf{c}'(t))$, then $\mathbf{C}'(t) = \mathbf{c}'(t)\mathbf{C}$. Therefore, from (A.11) and (A.13),

$$\mathbf{d}'(t)\mathbf{z} = \mathbf{B}'(t)\mathbf{z}_I + \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y})', 0']' \quad (A.14a)$$

and

$$\tilde{u}(t) - \mathbf{d}'(t)[0', \tilde{\mathbf{u}}']' = \tilde{u}(t) - \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\tilde{\mathbf{u}})', 0']'. \quad (A.14b)$$

As we showed in Section 3.4, our estimator of $z(t)$ is $\hat{z}(t)$ in (3.17), where $\mathbf{C}'(t)\hat{\mathbf{z}}_J = \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y})', 0']'$ (because $\mathbf{C}'(t) = \mathbf{c}'(t)\mathbf{C}$) and $E(\tilde{u}(t)\mid\tilde{\mathbf{u}}) = \mathbf{P}(t)\tilde{\mathbf{u}} = \text{cov}(\tilde{u}(t), \tilde{\mathbf{u}})\text{var}^{-1}(\tilde{\mathbf{u}})\tilde{\mathbf{u}}$. Note that if $1 \le t < t_k$, then $\tilde{u}(t) = 0$ and $\mathbf{P}(t) = 0$. Now, (3.17), (A.12), and (A.14) imply $\hat{z}(t) = \mathbf{d}'(t)\mathbf{z} + \mathbf{P}(t)(\mathbf{y} - \mathbf{C}\hat{\mathbf{z}}_J)$ and $\bar{z}(t) = \mathbf{d}'(t)\mathbf{z} + E[\tilde{u}(t) - \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\tilde{\mathbf{u}})', 0']' \mid \mathbf{J}_2\mathbf{w}]$. It remains to show the equality of the second terms in the last two equations. Given that $\mathbf{Q}_1'\mathbf{L}^{-1}\tilde{\mathbf{u}}$ and $\mathbf{Q}_2'\mathbf{L}^{-1}\tilde{\mathbf{u}}$ are two orthogonal sets of variables and that, by (A.6), $\mathbf{J}_2\mathbf{w} = \sigma\mathbf{Q}_2'\mathbf{L}^{-1}\tilde{\mathbf{u}}$, we have

$$E[\tilde{u}(t) - \mathbf{C}'(t)[(\mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\tilde{\mathbf{u}})', 0']' \mid \mathbf{J}_2\mathbf{w}] = E[\tilde{u}(t)\mid\mathbf{J}_2\mathbf{w}].$$

By definition of $\mathbf{P}(t)$, the following equalities hold:

$$\begin{aligned} \mathbf{P}(t)(\mathbf{y} - \mathbf{C}\hat{z}_J) &= \text{cov}(\tilde{u}(t), \tilde{\mathbf{u}})\text{var}^{-1}(\tilde{\mathbf{u}})(\mathbf{y} - \mathbf{C}\hat{\mathbf{z}}_J) \\ &= E(\tilde{u}(t)\tilde{\mathbf{u}}')(\mathbf{L}^{-1})'\mathbf{Q}\mathbf{Q}'\mathbf{L}^{-1}(\mathbf{y} - \mathbf{C}\hat{\mathbf{z}}_J) \\ &= E(\tilde{u}(t)\tilde{\mathbf{u}}')((\mathbf{L}^{-1})'\mathbf{Q}_1, (\mathbf{L}^{-1})'\mathbf{Q}_2)(0', (\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y})')' \\ &= E(\tilde{u}(t)(\mathbf{Q}_2'\mathbf{L}^{-1}\tilde{\mathbf{u}})')\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y} \\ &= E(\tilde{u}(t)(\mathbf{J}_2\mathbf{w})')\text{var}^{-1}(\mathbf{J}_2\mathbf{w})\mathbf{J}_2\mathbf{w} \\ &= E(\tilde{u}(t)\mid\mathbf{J}_2\mathbf{w}), \end{aligned}$$

where we have used the fact that $\mathbf{Q}_1'\mathbf{L}^{-1}(\mathbf{y} - \mathbf{C}\hat{z}_J) = 0$ and $\mathbf{Q}_2'\mathbf{L}^{-1}(\mathbf{y} - \mathbf{C}\hat{\mathbf{z}}_J) = \mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}$. Because $z(t) - \hat{z}(t) = z(t) - \bar{z}(t)$, if the same estimator of σ^2 is used, it follows that the MSE are equal.

APPENDIX B: ILLUSTRATIVE EXAMPLE

To illustrate our approach, we use the example of Kohn and Ansley (1986)—namely, the model

$$z(t) = z(t-4) + a(t) + \theta a(t-1). \quad (B.1)$$

With the notation of Section 2.1, we have $d = 4$, $p = 0$, $q = 1$, $u(t) = \delta(B)z(t)$, $\delta(B = 1 - B^4$, $\theta(B) = 1 + \theta B$ and $\xi(B) = 1/\delta(B) = 1$

Table 7. Details of Kalman Filter for Example B.1

t	$x(t\mid t-1)$	$c(t\mid t-1)$	$\sigma^2(t\mid t-1)$	$K(t)$	$L^{-1}y$	$L^{-1}C_1$	t	$x(t\mid t-1)$	$c(t\mid t-1)$	$\sigma^2(t\mid t-1)$	$K(t)$	$L^{-1}y$	$L^{-1}C_1$
5	1.2	0	1.25	1	.81	0	9	1.38	0	1.05	1	−.57	0
	0	0		−.4				3.2	1		−.48		
	0	0		0				−2.42	−.48		−.19		
	−1.3	0		0				.5	0		0		
6	−.36	0	1.05	1	3.47	.98	10	3.48	1	1.01	1	−3.85	0
	0	0		−.48				−2.31	−.48		−.59		
	−1.3	0		0				.5	0		0		
	2.1	0		0				.8	0		0		
7	−1.70	−.48	1.01	0	—	—	11	−.03	−.48	1.67	0	—	—
	−1.3	0		0				.5	0		0		
	2.1	0		0				.8	0		0		
	3.2	1		0				−.4	1		0		
8	−1.3	0	1.25	1	1.61	0	12	.5	0	1.25	1	.63	0
	2.1	0		−.4				.8	0		−.4		
	3.2	1		0				−.4	1		0		
	−1.70	−.48		−.4				−.03	−.48		−.4		

Table 8. Forecasting Details for Example B.1

t	$\mathbf{x}(t\|t-1)$	$\mathbf{L}^{-1}(\mathbf{z}_{II}-\mathbf{z}_*)$	$\sigma^2(t\|t-1)$	t	$\mathbf{x}(t\|t-1)$	$\mathbf{L}^{-1}(\mathbf{z}_{II}-\mathbf{z}_*)$	$\sigma^2(t\|t-1)$	t	$\mathbf{x}(t\|t-1)$	$\sigma^2(t\|t-1)$
5	1.2	.81	1.25	9	1.38	−.57	1.05	13	.52	1.05
	3.56				3.2				−.4	
	9,999				9,998.28				10,000.39	
	−1.3				.5				1.2	
6	3.2	0	1.05	10	3.48	−3.853	1.01	14	−.4	1.25
	9,999				9,998.39				10,000.39	
	−1.3				.5				1.2	
	2.1				.8				.52	
7	9,999	—	1.01	11	10,000.67	—	1.67	15	10,000.39	2.72
	−1.3				.5					
	2.1				.8					
	3.2				−.4					
8	−1.3	1.61	1.25	12	.5	.63	1.25			
	2.1				.8					
	3.2				−.4					
	9,999				10,000.67					

$+ B^4 + B^8 + \ldots$. The $\mathbf{A}'(t) = (A_1(t), A_2(t), A_3(t), A_4(t))$ vectors of (3.3) and (3.4) are generated from the recursions

$$A_i(t) = A_i(t-4), \qquad t > 4, \qquad i = 1, 2, 3, 4,$$

and

$$\begin{aligned} A_i(j) &= 1 \quad \text{if } i = j = 1, \ldots, 4; \\ &= 0 \quad \text{if } i, j = 1, \ldots, 4; \qquad i \neq j. \end{aligned}$$

To obtain the state–space representation, we need $\psi^*(B) = \theta(B)/\phi^*(B) = \theta(B)/\delta(B) = 1 + \theta B + B^4 + \ldots$. Equations (2.6a) and (2.9) in this case become

$$\mathbf{x}(t) = \begin{bmatrix} z(t) \\ z(t+1\,|\,t) \\ z(t+2\,|\,t) \\ z(t+3\,|\,t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z(t-1) \\ z(t\,|\,t-1) \\ z(t+1\,|\,t-1) \\ z(t+2\,|\,t-1) \end{bmatrix} + \begin{bmatrix} 1 \\ \theta \\ 0 \\ 0 \end{bmatrix} a(t), \qquad \text{(B.2)}$$

and

$$z(t) = (1, 0, 0, 0)\mathbf{x}(t) + \alpha(t)W(t).$$

Because there are missing values, with the notation of Section 3.3 we have $\mathbf{z}_* = (z(1), z(2), z(3), z(4))'$, $\mathbf{z}_J = (z(2), z(3))'$, $\mathbf{z}_I = (z(1), z(4))'$, $M = 8$, $k = 2$, $\tilde{u}(t) = \sum_{j=0}^{t-5} \xi_j u(t-j)$, $t > 4$, $\mathbf{z}_{II} = (z(5), z(6), z(8), z(9), z(10), z(12))'$, and $\tilde{\mathbf{u}} = (\tilde{u}(5), \tilde{u}(6), \tilde{u}(8), \tilde{u}(9), \tilde{u}(10), \tilde{u}(12))'$. The matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ corresponding to Equation (3.9) are in this case

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} \mathbf{I}_2 & 0 \\ \mathbf{B} & \mathbf{C} \end{bmatrix}.$$

It is easy to check that matrix $\mathbf{D}$ is the result of first interchanging columns 2 and 4 and then interchanging columns 3 and 4 in matrix $\mathbf{A}$. From the form of matrix $\mathbf{C}$, it is immediate that $\text{rank}(\mathbf{C}) = 1$ and that $z(3)$ is a free parameter. Let the covariance matrix of $\tilde{\mathbf{u}}$ be $\text{var}(\tilde{\mathbf{u}}) = \sigma^2\Delta$, and let $\Delta = \mathbf{LL}'$ be the Cholesky decomposition of Δ. To compute the function S^* in (3.15), we consider the regression model $\mathbf{y} = \mathbf{Cz}_J + \tilde{\mathbf{u}}$, where $\mathbf{y} = \mathbf{z}_{II} - \mathbf{Bz}_I$, and proceed as in Section 3.3. First, we apply the Kalman filter to the model $\mathbf{y} = \mathbf{z}_{II} - \mathbf{Bz}_I = \tilde{\mathbf{u}}$ to obtain $\mathbf{L}^{-1}\mathbf{y}$ and $|\mathbf{L}|$. The equations to use are (B.2). The starting conditions for the Kalman filter are $\mathbf{x}(5\,|\,4) = (z(1), 0, 0, z(4))'$ and

$$\Sigma(5\,|\,4) = \Xi\tilde{\Sigma}\Xi' = \mathbf{I}_4\tilde{\Sigma}\mathbf{I}_4 = \tilde{\Sigma} = \begin{bmatrix} 1+\theta^2 & \theta & 0 & 0 \\ \theta & \theta^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \qquad \text{(B.3)}$$

Note that the Kalman filter is applied to the vector of observations $\mathbf{z}_{II}$ and that it is not necessary to compute the entire matrix $\mathbf{B}$. Then the same algorithm applied to the columns of matrix $\mathbf{C}$ permits us to compute $\mathbf{L}^{-1}\mathbf{C}$. The starting conditions are $\mathbf{x}(5\,|\,4) = 0$ and (B.3). Therefore, we can consider the model $\mathbf{L}^{-1}\mathbf{y} = \mathbf{L}^{-1}\mathbf{Cz}_J + \mathbf{L}^{-1}\tilde{\mathbf{u}}$. Because the second column of matrix $\mathbf{C}$ is 0, so will be the second column of $\mathbf{L}^{-1}\mathbf{C}$. Therefore, if we denote the first column of matrix $\mathbf{C}$ by $\mathbf{C}_1$, then we may write $\mathbf{L}^{-1}\mathbf{y} = \mathbf{L}^{-1}\mathbf{C}_1 z(2) + \mathbf{L}^{-1}\tilde{\mathbf{u}}$, and $z(3)$ is a free parameter. We can apply the $\mathbf{QR}$ algorithm to the vector $\mathbf{L}^{-1}\mathbf{C}_1$ to obtain an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C}_1 = [\mathbf{R}, 0]'$, where $\mathbf{R}$ is a nonzero scalar. Let $\mathbf{Q}_1'$ denote the first row of $\mathbf{Q}'$ and let $\mathbf{Q}_2'$ denote the submatrix of $\mathbf{Q}'$ formed by the other three rows. Then we have Equation (3.13) with $\mathbf{E} = (\mathbf{R}, 0)$, and from this we obtain $\hat{z}(2) = \mathbf{R}^{-1}\mathbf{Q}_1'\mathbf{L}^{-1}\mathbf{y}$, $\hat{z}(3) = \tilde{x}$ (an arbitrary value), and $S^* = |\mathbf{L}|^{1/6}\mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y}|\mathbf{L}|^{1/6}$.

We now consider the numerical example of Kohn and Ansley; that is, we suppose $\theta = -.5$, $\sigma^2 = 1$, and $\mathbf{z} = (1.2, -1.3, 2.1, 3.2, .5, .8, -.4, 1.2)'$. In Table 7, we give output from the Kalman filter applied as described previously. For each time index, we present values for the MSE $\sigma^2(t\,|\,t-1)$, the vectors $\mathbf{x}(t\,|\,t-1)$, and $\mathbf{c}(t\,|\,t-1)$ (for column $\mathbf{C}_1$), the Kalman gain $\mathbf{K}(t)$, and the corresponding elements of the vectors of standardized residuals $\mathbf{L}^{-1}\mathbf{y}$ and $\mathbf{L}^{-1}\mathbf{C}_1$. It is easy to check that if $\mathbf{Q}'$ is the matrix obtained from the unit matrix by interchanging its first and second rows, then $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{C}_1 = (.976, 0')' = (\mathbf{R}, 0)'$ and $\mathbf{Q}'\mathbf{L}^{-1}\mathbf{y} = (3.474, .805, 1.610, -.566, -3.853, .626)'$. From this we obtain $z(2) = 3.56$ with MSE 2.222, $\mathbf{y}'(\mathbf{L}^{-1})'\mathbf{Q}_2\mathbf{Q}_2'\mathbf{L}^{-1}\mathbf{y} = 18.8$, and finally $S^* = 21.406$.

As for interpolation and prediction, we have just seen that $\hat{z}(2) = 3.56$ and $z(3)$ is a free parameter. To see whether the interpolators for $z(7)$ and $z(11)$ depend on the free parameter, we must examine its $\mathbf{C}'(t)$ vectors. We have $\mathbf{C}'(7) = (0, 1)$ and $\mathbf{C}'(11) = (0, 1)$. Therefore, both interpolators will depend on the free parameter. As for the forecasts of $z(13)$, $z(14)$, and $z(15)$, we have $\mathbf{C}'(13) = (0, 0)$, $\mathbf{C}'(14) = (1, 0)$, and $\mathbf{C}'(15) = (0, 1)$. Therefore, only the forecast of $z(15)$ will depend on the free parameter. To obtain the forecasts, we put $z(2) = 3.56$ and $z(3) = 9{,}999$ (an arbitrary value) and apply the Kalman filter to the model $\mathbf{z}_{II} - \mathbf{A}\mathbf{z}_* = \tilde{\mathbf{u}}$. The starting conditions are (see Secs. 3.4 and 3.2) $\mathbf{x}(6|5) = (1.2, 3.56, 9999, -1.3)'$ and (B.3). In Table 8 we present output from the Kalman filter applied to $\mathbf{z}_{II} - \mathbf{A}\mathbf{z}_* = \tilde{\mathbf{u}}$. This time we only give the values for $x(t|t-1)$, $\mathbf{L}^{-1}(\mathbf{z}_{II} - A z_*)$, and the MSE $\sigma^2(t|t-1)$, with $\mathbf{K}(t)$ being the same as in Table 7 for $t = 1, \ldots, 12$.

[Received March 1992. Revised December 1992.]

REFERENCES

Akaike, H. (1974), "Markovian Representation of Stochastic Processes and Its Application to the Analysis of Autoregressive Moving Average Processes," *Annals of the Institute of Statistical Mathematics,* 26, 363–387.

Anderson, B., and Moore, J. (1979), *Optimal Filtering,* Englewood Cliffs, NJ: Prentice Hall.

Ansley, C. F., and Kohn, R. (1990), "Filtering and Smoothing in State–Space Models with Partially Diffuse Initial Conditions," *Journal of Time Series Analysis,* 11, 275–293.

——— (1985), "Estimation, Filtering, and Smoothing in State–Space Models with Incompletely Specified Initial Conditions," *The Annals of Statistics,* 13, 1286–1316.

Ansley, C. F., and Newbold, P. (1981), "On the Bias in Estimates of Forecast Mean Squared Error," *Journal of the American Statistical Association,* 76, 569–578.

Bell, W., (1984), "Signal Extraction for Nonstationary Series," *The Annals of Statistics,* 12, 646–664.

Bell, W., and Hillmer, S. C. (1991), "Initializing the Kalman Filter for Nonstationary Time Series Models," *Journal of Time Series Analysis,* 12, 283–300.

Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis, Forecasting and Control,* San Francisco: Holden-Day.

Box, G. E. P., and Tiao, G. C. (1975), "Intervention Analysis With Applications to Economic and Environmental Problems," *Journal of the American Statistical Association,* 70, 70–79.

Brockwell, P., and Davis, R. (1987), *Time Series: Theory and Methods,* Berlin: Springer-Verlag.

de Jong, P. (1988), "The Likelihood for the State-Space Model," *Biometrika,* 75, 165–169.

——— (1989), "Smoothing and Interpolation With the State Space Model," *Journal of the American Statistical Association,* 84, 408–409.

——— (1991), "The Diffuse Kalman Filter," *The Annals of Statistics,* 19, 1073–1083.

Fletcher, R. (1987), *Practical Methods of Optimization,* New York: John Wiley.

Gómez, V., and Maravall, A. (1992), "Time Series Regression with ARIMA Noise and Missing Observations: Program TRAM," Working Paper ECO 92/81, European University Institute.

Harvey, A. C., and Phillips, G. D. A. (1979), "Maximum Likelihood Estimation of Regression Models With Autoregressive Moving Average Disturbances," *Biometrika,* 66, 49–58.

Harvey, A. C., and Peters, S. (1990), "Estimation Procedures for Structural Time Series Models," *Journal of Forecasting,* 9, 89–108.

Harvey, A. C., and Pierse, R. G. (1984), "Estimating Missing Observations in Economic Time Series," *Journal of the American Statistical Association,* 79, 125–131.

Hillmer, S. C., Bell, W. R., and Tiao, G. E. (1983), "Modeling Considerations in the Seasonal Adjustment of Economic Time Series," in *Applied Time Series Analysis of Economic Data,* ed. A. Zellner, Washington, DC: Bureau of the Census.

Jones, R. (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations," *Technometrics,* 22, 389–395.

Kohn, R., and Ansley, C. F. (1985), "Efficient Estimation and Prediction in Time Series Regression Models," *Biometrika,* 72, 694–697.

——— (1986), "Estimation, Prediction, and Interpolation for ARIMA Models With Missing Data," *Journal of the American Statistical Association,* 81, 751–761.

Mélard, G. (1984), "A Fast Algorithm for the Exact Likelihood of Autoregressive Moving Average Models," *Applied Statistics,* 35, 104–114.

Morf, M., Sidhu, G. S., and Kailath, T. (1974), "Some New Algorithms for Recursive Estimation on Constant, Linear, Discrete Time Systems," *IEEE Transactions on Automatic Control,* AC-19, 315–323.

Pearlman, J. G. (1980), "An Algorithm for the Exact Likelihood of a High-Order Autoregressive Moving Average Process," *Biometrika,* 67, 232–233.

Rao, C. (1973), *Linear Statistical Inference and its Applications,* New York: John Wiley.

Wecker, W., and Ansley, C. F. (1983), "The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing," *Journal of the American Statistical Association,* 78, 81–89.