The ARIMA model in state space form

Piet de Jong and Jeremy Penzer

Department of Statistics, London School of Economics
Houghton Street, London, WC2A 2AE, UK.

August 11, 2000

Abstract

This article explores an alternative state space representation for ARIMA models to that usually advocated. The alternative representation has minimal state order. More importantly, it has more convenient Kalman filter convergence properties. This convergence reveals the concrete connection between classical infinite sample representations based on lag polynomials and the recursive Kalman filter construction.

KEYWORDS: Filter convergence; Kalman filter smoother; State space model; Time series.

1 Introduction

State space forms are used for prediction, smoothing and likelihood evaluation (Schweppe 1965; Anderson and Moore 1979; Harvey and Phillips 1979). There are several practical state space forms for an ARMA or ARIMA process. The corresponding Kalman filter recursion converges provided the model is invertible. We advocate the use of a representation with simple and transparent convergence properties. The converged quantities makes obvious the relationship between filtering and the classical ARMA approach of conditioning on presample values. The nature of smoothing algorithm quantities also becomes apparent and the approach establishes the finite sample generalisation of methods for dealing with explanatory variables.

The results mentioned above can be derived using any practical state space representation of an ARMA process. However, using the approach advocated in this paper is particularly straightforward, making for a worthwhile tool and illustrating the usefulness of the correlated form of the state space model and the judicious labelling of state space disturbances.

2 The ARIMA model and the state space form

The ARMA(p, q) model is

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q},$$

written in the usual lag operator notation as $\phi(B)y_t = \theta(B)\varepsilon_t$ where $\varepsilon_t \sim (0, \sigma^2)$. The possibility that $\phi(B)$ has roots inside or on the unit circle, and hence that the model is nonstationary (ARIMA) is not excluded.

The general state space form is

$$y_t = Z_t\alpha_t + G_t\epsilon_t, \tag{1}$$

$$\alpha_{t+1} = T_t\alpha_t + H_t\epsilon_t, \quad t = 1, \ldots, n, \tag{2}$$

where $\epsilon_t \sim (0, \sigma^2 I)$, $\alpha_1 \sim (a_1, \sigma^2 P_1)$ and the ϵ_t and α_1 are mutually uncorrelated. The system matrices Z_t, T_t, G_t and H_t are nonrandom, typically depend on hyperparameters and, as the notation indicates, may vary over time. For a univariate model with an $s \times 1$ state vector α_t and $m \times 1$ vector of errors ϵ_t, the matrices Z_t, T_t, G_t and H_t are $1 \times s$, $s \times s$, $1 \times m$ and $s \times m$ respectively.

Pearlman (1980) puts forward the following ARMA state space representation as well known; for $t = 1, \cdots, n$, $Z_t = Z = (1, 0, \ldots, 0)$, $G_t = G = 1$,

$$T_t = T = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & 0 & \cdots & 0 & 1 \\ \phi_m & 0 & \cdots & \cdots & 0 \end{pmatrix}, \qquad H_t = H = \begin{pmatrix} \theta_1 + \phi_1 \\ \theta_2 + \phi_2 \\ \vdots \\ \vdots \\ \theta_m + \phi_m \end{pmatrix}.$$

where $m = \max(p, q)$. However, Pearlman gives no references for this form and it does not appear to be dealt with in the literature. We shall refer to it as the $\max(p, q)$ representation. In this representation ϵ_t in (1) and (2) is the same as ε_t in the original ARMA model. Note $G_t H_t' \neq 0$ implying correlated measurement and state noise.

In the literature (Brockwell and Davis 1987; Harvey 1989; Box, Jenkins, and Reinsel 1994; Hamilton 1994) the $\max(p, q)$ representation has been overlooked in favour of one in which the state vector is of length $m = \max(p, q + 1)$. In this version, Z and T are as above but $G = 0$ and $H = (1, \theta_1, \ldots, \theta_{m-1})'$. The prevalence of this form may be explained by the fact that the measurement and state noise are uncorrelated - there is no measurement noise. Uncorrelated measurement and state noise fits in with the more usual state space form where $G_t H_t' = 0$ (Anderson and Moore 1979, p. 14).

Using the Kalman filter the observations y_t are transformed to innovations v_t. In general, for $t = 1, \ldots, n$,

$$\begin{aligned} v_t = y_t - Z_t a_t, \qquad & F_t = Z_t P_t Z_t' + G_t G_t', \\ & K_t = (T_t P_t Z_t' + H_t G_t') F_t^{-1}, \\ a_{t+1} = T_t a_t + K_t v_t, \qquad & P_{t+1} = T_t P_t L_t' + H_t J_t', \end{aligned} \tag{3}$$

where $L_t = T_t - K_t Z_t$ and $J_t = H_t - K_t G_t$. The slight simplification of (3) made possible by the $\max(p, q + 1)$ representation must be balanced against desirable features of the

$\max(p,q)$ form. It is our contention that the arguments in favour of the $\max(p,q)$ version are compelling. First, when $q \geq p$ the state vector is shorter providing a slight computational advantage. Second, the converged quantities in the $\max(p,q)$ representation take on convenient and readily interpretable forms.

3 Convergence properties for filtering

Convergence of the Kalman filtering is established using the properties of the underlying ARMA model. Put $Y_{t,-\infty} = [y_t, y_{t-1}, \ldots, y_1, y_0, \ldots]$, the linear space spanned by the entire past of the series. If the series is a pure AR(p) then, for minimum mean square error linear prediction, $Y_{t,-\infty}$ can be replaced by $Y_t = [y_t, \ldots, y_1]$ whenever $t > p$. By appropriate choice of c, an invertible ARMA(p,q) model can be approximated, to any degree of accuracy, by an AR(c) process. Thus with an invertible ARMA(p,q), we can find c so that, for purposes of prediction, $Y_{t,-\infty}$ can be replaced by Y_t whenever $t > c$. The size of c depends on how close the roots of the MA polynomial $\theta(B)$ are to the unit disc, that is the closeness of the model to noninvertibility.

Consider the $\max(p,q)$ representation. Let $\alpha_{j,t}$ denote component j of α_t for $j = 1, \ldots, m$. From (1), $\alpha_{1,t} = y_t - \varepsilon_t$. Using the state equation (2), for $j = 1, \ldots, m$,

$$\begin{aligned} \alpha_{j,t+1} &= \phi_j(y_t - \varepsilon_t) + \alpha_{j+1,t} + (\theta_j + \phi_j)\varepsilon_t \\ &= \phi_j y_t + \alpha_{j+1,t} + \theta_j \varepsilon_t \\ &= \phi_j y_t + \cdots + \phi_m y_{t-m+j} + \theta_j \varepsilon_t + \cdots + \theta_m \varepsilon_{t-m+j}, \end{aligned} \tag{4}$$

provided the linear combination in (4) does not extend back to the presample period, that is, provided $t \geq m$. Thus, if the model is invertible, $\alpha_{t+1} \in Y_{t,-\infty}$ for $t > m$ and so, for $t \geq c$, we can assume $\alpha_{t+1} \in Y_t$. By definition of the filter $a_{t+1} = E(\alpha_{t+1}|Y_t)$ so, for $t > c$ we have $a_t = \alpha_t$. This implies that, once the filter has converged, $P_t = \sigma^{-2}\text{var}(\alpha_t - a_t) = 0$ and hence $F_t = 1$, $K_t = H$, $J_t = 0$ and L_t has the same form as T but with θ_j's replacing the ϕ_j's. Thus, for $t > c$, the Kalman filter collapses to the prediction error computation

$$v_t = y_t - \phi_1 y_{t-1} - \cdots - \phi_m y_{t-m} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_m \varepsilon_{t-m} = \frac{\phi(B)}{\theta(B)} y_t = \varepsilon_t,$$

which is conceptually and computationally convenient. Kalman filtering implicitly inverts the moving average polynomial and this inversion is achieved recursively without assumptions about presample values.

With the $\max(p, q+1)$ representation $y_t = \alpha_{1,t}$ and $\phi(B)y_t = \theta(B)\varepsilon_{t-1}$. Hence $\varepsilon_t \in Y_{t+1,-\infty} - Y_{t,-\infty}$ and, for $t > c$, $a_{j,t+1} = \alpha_{j,t+1} - \theta_{j-1}\varepsilon_t$. Thus for $t > c$, $P_t = HH'$, $F_t = 1$, $K_t = (\phi_1, \ldots, \phi_m)'$, L_t is the same as T except for the top left entry where it is 0, $J_t = H$ while $v_t = \varepsilon_{t-1}$ and $a_t = \alpha_t - H\varepsilon_{t-1}$. Thus, with the $\max(p, q+1)$ representation, the state estimate does not converge to the state and the interpretation of filtered quantities is awkward.

4 Convergence properties for smoothing

Smoothing quantities under the $\max(p, q)$ representation also converge to convenient and readily interpretable constructs. The smoothing filter (De Jong 1988; Kohn and Ansley 1989) corresponding to the general state space representation takes the following form. Put $r_n = 0$, $N_n = 0$ and for $t = n, \ldots, 1$,

$$\begin{aligned} u_t &= F_t^{-1}v_t - K_t'r_t, \qquad & M_t &= F_t^{-1} + K_t'N_tK_t, \\ r_{t-1} &= Z_t'u_t + T_t'r_t, \qquad & N_{t-1} &= Z_t'F_t^{-1}Z_t + L_t'N_tL_t, \end{aligned}$$

The smoothations u_t contain information about departures from the model (De Jong and Penzer 1998) and have the interpolation characterization

$$u_t = \{\mathrm{var}(y_t|Y_n^t)\}^{-1}\{y_t - E(y_t|Y_n^t)\},$$

where $Y_n^t = [y_n, \ldots, y_{t+1}, y_{t-1}, \ldots, y_1]$, the punctured space.

An appealing expression for the smoothations is readily derived using the $\max(p, q)$ representation. First note that $\alpha_t \in Y_{t-1,-\infty} \subset Y_{n,-\infty}$ and hence for $t = c+1, \ldots, n$, $\alpha_t = E(\alpha_t|Y_n)$. Similarly, $\varepsilon_t \in Y_{t,-\infty} \subset Y_{n,-\infty}$ implying for $t = c+1, \ldots, n$,

$$\varepsilon_t = E(\varepsilon_t|Y_n) = G_t'u_t + H_t'r_t = u_t + H'r_t.$$

where the second equality follows from De Jong (1988) and Koopman (1993). Thus, $u_t = \varepsilon_t - H'r_t$ and iterating this identity using $r_{t-1} = Z'u_t + T'r_t$ yields

$$u_t = \varepsilon_t - \phi_1\varepsilon_{t+1} - \cdots - \phi_m\varepsilon_{t+m} - \theta_1u_{t+1} - \cdots - \theta_mu_{t+m}$$

or, in the lag polynomial notation, for $t = c+1, \ldots, n$

$$u_t = \frac{\phi(B^{-1})}{\theta(B^{-1})}\varepsilon_t = \frac{\phi(B^{-1})\phi(B)}{\theta(B^{-1})\theta(B)}y_t \tag{5}$$

where ε_t and y_t are interpreted as zero if $t > n$. This expression is exact provided that the Kalman filter has converged so the result only requires invertibility. The expression (5) mirrors the infinite sample Wiener-Kolmogorov interpolation formula (Whittle 1984). The KFS computes exact finite sample interpolation errors. Provided Kalman filter convergence, these interpolation errors coincide with the infinite sample interpolation errors for $t \leq n - m$. There is no requirement for smoothing filter convergence or stationarity.

5 Other advantages of the $\max(p, q)$ representation

The $\max(p, q)$ representation can be used to clarify a number of related areas.

5.1 Maximum likelihood estimation

ARMA model parameters are often estimated using normal based maximum likelihood. With the $\max(p, q)$ representation the log-likelihood takes the form, ignoring constants,

$$-\frac{1}{2}\left[n\log\sigma^2 + \sum_{t=1}^{c}\left\{\log F_t + \frac{v_t^2}{\sigma^2 F_t}\right\} + \sum_{t=c+1}^{n}\frac{v_t^2}{\sigma^2}\right],$$

where $v_t = \varepsilon_t = \{\phi(B)/\theta(B)\}y_t$ for $t > c$. Initial conditions are handled exactly and there is no need for such tools as backcasting (Box, Jenkins, and Reinsel 1994, p. 289).

Maximizing the log-likelihood with respect to σ^2 yields

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{t=1}^{c}\frac{v_t^2}{F_t} + \sum_{t=c+1}^{n}\varepsilon_t^2\right).$$

Substituting back gives the concentrated log-likelihood which is maximized by minimizing

$$n\log\hat{\sigma}^2 + \sum_{t=1}^{c}\log F_t,$$

with respect to the remaining parameters. This is least squares provided that we can ignore the determinental term $\sum_{t=1}^{c}\log F_t$. Kalman filtering is required only for the initial c observations. These expressions are the similar to those for the $\max(p, q+1)$ representation, except that in the $\max(p, q+1)$ parametrization the computation of the converged quantities is less convenient.

5.2 Vector ARMA models

No new issues arise for this case. The ϕ_j, θ_j and ψ_j are now square matrices and 1's in the state space representation are replaced by identity matrices. Also ε_t in the measurement equation is replaced by $\theta_0\varepsilon_t$ where the square matrix θ_0 models the contemporaneous correlation amongst the components of y_t. A corresponding adjustment is made to the definition of $\theta(B)$. In the vector case, upon convergence, $v_t = \theta_0\varepsilon_t$ and $F_t = \theta_0\theta_0'$.

5.3 Regression with ARMA(p, q) errors

This model can be written as $y_t = x_t'\beta + z'\alpha_t + \varepsilon_t$ where x_t is a vector of regression variables and $z'\alpha_t + \varepsilon_t$ is the ARMA(p, q) error as given in the max(p, q) state space representation. Initially suppose disturbances $z'\alpha_t + \varepsilon_t$ are AR(p). Then the Cochran-Orcutt procedure (Johnston 1984) can be used to perform generalized least squares supposing the AR coefficients are known. In particular y_t is transformed to $y_t^* = \phi(B)y_t$ and x_t to $x_t^* = \phi(B)x_t$ and y_t^* is regressed on x_t^*. This reduces the problem of efficiently estimating β to one of weighted ordinary least squares provided the initial data (y_t, x_t), $t = 1, \ldots, p$ are handled appropriately. In particular, the generalized least squares estimate is

$$\hat{\beta} = \left(\sum_{t=1}^{p} \frac{x_t^* x_t^{*\prime}}{F_t} + \sum_{t=p+1}^{n} x_t^* x_t^{*\prime} \right)^{-1} \left(\sum_{t=1}^{p} \frac{x_t^* y_t^*}{F_t} + \sum_{t=p+1}^{n} x_t^* y_t^* \right)$$

provided the y_t^*, x_t^* and F_t are appropriately defined for $t = 1, \ldots, p$. With an AR(1), $y_1^* = y_1$, $x_1^* = x_1$ and $F_1 = 1 - \phi_1^2$. In the general ARMA(p, q) case the expression for $\hat{\beta}$ is as above but with p replaced by c.

With the max(p, q) representation, the computation of the x_t^* and y_t^* can be stated in terms of augmented or diffuse Kalman filtering (Rosenberg 1973; Wecker and Ansley 1983; De Jong 1991). The basic Kalman filter equations are augmented with extra recursions relating to the presence of the regression variables x_t. Define, for $t = 1, \ldots, n$,

$$x_t^{*\prime} = x_t' - z'A_t \quad , \quad A_{t+1} = TA_t + K_t x_t^{*\prime} \tag{6}$$

where $A_1 = 0$. For $t = c + 1, \ldots, n$, using the same argument as in §2, this transformation can be written as

$$x_t^* = x_t - \phi_1 x_{t-1} - \cdots - \phi_m x_{t-m} - \theta_1 x_{t-1}^* - \cdots - \theta_m x_{t-m}^* = \frac{\phi(B)}{\theta(B)} x_t$$

which is the generalized Cochran-Orcutt transformation. The remaining issue relates to the initial portion of the data $t = 1, \ldots, c$, before convergence. It can be shown that (6) performs the appropriate transformation.

5.4 Transfer function modelling

The $\max(p, q)$ representation and augmented Kalman filtering can also be used with transfer function models. A transfer function model is $y_t = \psi(B)x_t + \varepsilon_t$ where both y_t and x_t are observed and ε_t is a white noise disturbance. The aim is to estimate the transfer function $\psi(B)$. In general, x_t is assumed to be an ARMA process $\phi(B)x_t = \theta(B)\eta_t$. One proposal (Box, Jenkins, and Reinsel 1994, p. 417) is to use prewhitening, that is, apply the linear filter $\phi(B)/\theta(B)$ to both x_t and y_t yielding

$$v_t = \frac{\phi(B)}{\theta(B)} y_t = \psi(B)\eta_t + \frac{\phi(B)}{\theta(B)}\varepsilon_t$$

Preliminary estimates of the coefficients in $\psi(B)$ are then calculated using the cross correlation function between the v_t and η_t. With the $\max(p, q)$ representation, the transformation $\phi(B)/\theta(B)$ can be applied to y_t using the Kalman filter. Further applying the augmented Kalman filter under this setup yields $\eta_t = x_t^* = \{\phi(B)/\theta(B)\}x_t$ where x_t^* is defined as in §5.3. For $t > c$, the KF calculates the prewhitened series v_t and $x_t^* = \eta_t$ exactly. For $t \leq c$ the v_t and x_t^* are uncorrelated but heteroskedastic and hence for these initial observations the cross product terms of the form $v_t\eta_{t-k}$ in the cross covariance estimate have to be appropriately standardized. The appropriate standardization for η_t is based on F_t.

6 Conclusion

This paper has dealt with a particular state space form for ARIMA models that has been overlooked in the time series literature. The form has both computational and conceptual advantages. With this form, the Kalman filter collapses, after the processing of an initial stretch of the data, to computing the exact moving average errors $\varepsilon_t = \{\phi(B)/\theta(B)\}y_t$. Collapsing is analogous to augmented Kalman filtering reducing to ordinary Kalman filtering in the nonstationary case and emphasizes that uncollapsed forms deal with the tedium of exact initialization. The advocated state space form clarifies smoothing, maximum likelihood estimation, and transformation issues for time series regression models.

References

ANDERSON, B. D. O. and MOORE, J. B. (1979) *Optimal filtering*. Englewood Cliffs, New Jersey: Prentice-Hall.

BOX, G. E. P., JENKINS, G. M. and REINSEL, G. C. (1994) *Time Series Analysis: Forecasting and Control* (3rd edn). Englewood Cliffs, New Jersey: Prentice-Hall.

BROCKWELL, P. J. and DAVIS, R. A. (1987) *Time Series: Theory and Models*. New York: Springer-Verlag.

DE JONG, P. (1988) A cross-validation filter for time series models. *Biometrika 75*, 594–600.

DE JONG, P. (1991) The diffuse Kalman filter. *Annals of Statistics 19*, 1073–83.

DE JONG, P. and PENZER, J. R. (1998) Diagnosing shocks in time series. *Journal of the American Statistical Association 93*, Forthcoming.

HAMILTON, J. D. (1994) *Time Series Analysis*. Princeton: Princeton University Press.

HARVEY, A. C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

HARVEY, A. C. and PHILLIPS, G. D. A. (1979) Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika 66*, 49–58.

JOHNSTON, J. (1984) *Econometric Methods* (3rd edn). Singapore: McGraw-Hill.

KOHN, R. and ANSLEY, C. F. (1989) A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika 76*, 65–79.

KOOPMAN, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–26.

PEARLMAN, J. G. (1980) An algorithm for the exact likelihood of a high-order autoregressive-moving average process. *Biometrika 67*, 232–3.

ROSENBERG, B. (1973) Random coefficient models: the analysis of a cross section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement 2*, 339–428.

SCHWEPPE, F. (1965) Evaluation of likelihoods for Gaussian signals. *IEEE Transactions on Information Theory 11*, 61–70.

WECKER, W. E. and ANSLEY, C. F. (1983) The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association 78*, 81–9.

WHITTLE, P. (1984) *Prediction and Regulation by Linear Least Square Methods* (2nd edn). Oxford: Blackwell.