Title
Adaptive filtering revisited

Journal
Journal of the Operational Research Society, 30(9)

ISSN
0160-5682

Authors
Nau, RF
Oliver, RM

Publication Date
1979

DOI
10.1057/jors.1979.193

Peer reviewed

J. Opl Res. Soc., Vol. 30, No. 9, pp. 825 to 831
Pergamon Press Ltd 1979. Printed in Great Britain

0030-3623/79/0901-0825$02.00/0

Adaptive Filtering Revisited

ROBERT F. NAU* and ROBERT M. OLIVER†

* Operations Research Center and †Department of Industrial Engineering and Operations Research, University of California, Berkeley

This paper shows that the adaptive filtering and forecasting techniques proposed by Makridakis and Wheelwright can be viewed as approximations to a more precise filtering method in which the Kalman filter is applied to a dynamic autoregressive model which is a special case of the models of Harrison and Stevens. The correct "learning" or "training factors" are shown to be data-dependent matrices rather than scalar constants.

INTRODUCTION

IN RECENT years there has been a growing interest in adaptive methods for the analysis and forecasting of time series, characterized by the use of recursive estimation techniques. These techniques have often been inspired by or directly adopted from the theory of optimal control. An advantage of adaptive methods over more traditional methods, such as those of Box–Jenkins[1], includes the capability of dealing with time series which are nonstationary in fundamental ways that cannot be overcome by simple transformations of the data, for example when the underlying parameters of the process, or its very structure, may be changing randomly with time. One may also be able to deal with the "start-up" stage in forecasting.

A number of formal Bayesian models, incorporating the structure of the Kalman filter are described in an important paper by Harrison and Stephens.[2] Separately, Makridakis and Wheelwright[3–5] have introduced an heuristic approach to adaptive forecasting which has received much attention and some criticism. The "adaptive filtering" methods of Makridakis and Wheelwright (MW) consist of algorithms for revising the coefficients of linear regression and ARMA models which add to the old parameters a correction term proportional to the product of the most recent forecast residual and the values of prior observations. Their models have essentially only two *a priori* parameters: the number of autoregressive or moving average terms to be included and an arbitrary scalar "learning constant" or "training factor." The authors claim a wide applicability for these adaptive filtering models and also that, by judicious choice of the above parameters and repeated use of the technique on a given record of historical data, one can often produce superior results to Box-Jenkins methods with considerably less effort. Although their suggested procedures are intuitively appealing and appear to yield interesting results in diverse applications, they have been questioned on both theoretical and empirical grounds. It has not been shown exactly which assumptions and models formally yield their revision rules, nor was it obvious that the parameter revisions could be derived from an equation of motion underlying the random process itself.

This paper reviews the important features of the MW model and introduces a rigorous adaptive forecasting model based on the Kalman filter approach suggested by Harrison and Stephens. The important similarities and differences in the two approaches are discussed.

THE MW ADAPTIVE FILTERING MODEL

Let the forecasts of x_t made at time $t-1$ and of x_{t+1} made at time t be

$$\hat{x}_t = \phi_{t1}x_{t-1} + \cdots + \phi_{tp}x_{t-p} \tag{1a}$$

$$\hat{x}_{t+1} = \phi'_{t1}x_t + \cdots + \phi'_{tp}x_{t-p+1}$$

$$= \phi_{t+1,1}x_t + \cdots + \phi_{t+1,p}x_{t-p+1}. \tag{1b}$$

(MW[4] incorrectly include the residual in the forecasting equation.) The suggested MW procedure for revising the parameters ϕ_{tj}, $1 \leqslant j \leqslant p$, is either[3]

$$\phi'_{tj} = \phi_{tj} + 2ke_t x_{t-j} \tag{2a}$$

or[4]

$$\phi'_{tj} = \phi_{tj} + 2\kappa e_t^* x_{t-j}^* \tag{2b}$$

where k and κ are appropriate "learning factors", and

$$e_t = x_t - \hat{x}_t \tag{2c}$$

is the error or one step ahead forecast residual and where e_t^*, x_{t-j}^* are suitably standardized values of e_t, x_{t-j}.

MW[4] have specifically suggested the standardization

$$x_{t-j}^* = \frac{x_{t-j}}{\left(\sum_{j=1}^{p} x_{t-j}^2\right)^{1/2}} \qquad 1 \leqslant j \leqslant p \tag{3a}$$

$$e_t^* = \frac{e_t}{\left(\sum_{j=1}^{p} x_{t-j}^2\right)^{1/2}}. \tag{3b}$$

This standardization has the important effect of making the data and residuals (hence also the learning factor) dimensionless quantities, and therefore independent of the units of observation, which eases comparisons between different time series. Besides its simplicity the adaptive filtering model in (2), (3) has the advantage that (i) the amount of change in a given parameter is proportional to both the forecast error and to the relative amount that parameter contributes to the forecast, and (ii) like a good steepest-descent algorithm for minimizing the squared residual, revises the parameter vector in the direction of the gradient

$$\left(\frac{\partial e_t^2}{\partial \phi_{t,1}}, \frac{\partial e_t^2}{\partial \phi_{t,2}}, \ldots, \frac{\partial e_t^2}{\partial \phi_{t,p}}\right)$$

where

$$\frac{\partial e_t^2}{\partial \phi_{t,j}} = 2e_t \frac{\partial e_t}{\partial \phi_{t,j}} = -2e_t x_{t-j}.$$

The term on the right is, except for the "learning factor" k or κ, identical to the correction term in (2a), (2b). The model also has a superficial resemblance to the parameter updating equations of the Kalman filter, in which the magnitude of the revision in a parameter estimate is equal to the most recent forecast error multiplied by a gain factor which depends on values of recent observations; and the parameter revision rule resembles the forecast updating equations for exponential smoothing (EWMA models). In particular, the AR(1) model in adaptive filtering leads to a revised parameter which is a linear interpolation between the old parameter value and the ratio, x_t/x_{t-1}, which may be regarded as a noisy measurement of the scalar autoregressive parameter. Finally, in contrast to the exponential smoothing models where the correction or adjustment term in forecasts is proportional to e_t, in MW the adjustment of the parameters is proportional to the product $e_t x_{t-j}$ and therefore to terms of the form $x_t x_{t-j}$ since e_t, the forecast residual equals x_t minus the forecast. Thus, to the extent that parameter estimates depend linearly on estimates of autocorrelations (the Yule–Walker equations) it is probably reasonable to include terms of the form $e_t x_{t-j}$ for calculating a revised parameter estimate. However, the precise form that such revisions should take is not theoretically justified in any of MW's papers.

The MW model has important weaknesses. First, it is vague as regards a possible underlying mathematical or physical model and therefore fails to distinguish between the separate sources of uncertainty that may contribute to forecast error, e.g. uncertainty in the prior estimates of parameters, possible changes in the actual values of the parameters or noise in the observations. Instead, it is left to the user to incorporate his experience and subjective feelings about various aspects of the process into a single scalar "learning constant". The important questions of the variance and possible bias of parameter estimates and forecasts are left unresolved.

It appears that the use of a constant learning factor may lead to spurious revisions and unstable forecasts, especially in the standardized model. In a pure AR(p) process, for example, if p consecutive observations are close to zero, then the next observation can be expected to represent almost pure noise, thus giving little information regarding the autoregressive parameters. In the MW adaptive filtering model, however, if p small observations are followed by a "normal" value, the standardized residual e_t^* can become arbitrarily large whereas the standardized observations will always remain of the order of $1/\sqrt{p}$. Thus for any fixed value of the training constant, κ, large changes in parameter estimates can conceivably be generated by situations in which almost no actual learning occurs. Conversely, a run of exceptionally large values might represent a temporary improvement in the signal-to-noise ratio and hence an occasion for accelerated learning. In such cases the effect of the standardization is to decrease the learning.

THE DYNAMIC AUTOREGRESSIVE MODEL AND KALMAN FILTER

Using a vector notation to formulate a dynamic autoregressive model with ϕ_t denoting a p element vector of parameters and H_{t-1} the vector of p prior observations $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$:

$$\phi_t = \begin{bmatrix} \phi_{t1} \\ \phi_{t2} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{tp} \end{bmatrix}, \quad H_{t-1} = \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \cdot \\ \cdot \\ \cdot \\ x_{t-p} \end{bmatrix}. \tag{4a,b}$$

The dynamic model is defined by the two concurrent random equations of motion:

$$x_t = H_{t-1}^T \phi_t + a_t \qquad a_t \sim N(0, \sigma_a^2) \tag{5a}$$

$$\phi_t = \phi_{t-1} + b_t \qquad b_t \sim N(0, Q). \tag{5b}$$

(5a) is the AR(p) equation of motion, (5b) is the postulated motion of the parameters. Both a_t and b_t are independently identically distributed and a_t is a stationary (scalar) Gaussian noise sequence and b_t is a stationary (vector) Gaussian noise sequence with covariance matrix Q. Also a_t is assumed independent of all ϕ_t's and all b_t's and b_t of all prior x_t's.

Equation (5b) generates a value of ϕ_t starting from ϕ_{t-1} using the random walk equation. These values only weight p past observations which, when added to a new noise term, yield a new observation x_t. The scalar x_t is measured but the vector ϕ_t is not. The problem is to use x_{t-1} to filter and make good estimates of what values of ϕ_{t-1} occurred, to estimate future values of ϕ_t from (5b) and then use (5a) to forecast future x_t's, repeating, as needed, the cycle of filtering and forecasting.

The inner product of prior observations is the scalar quantity

$$H_{t-1}^T H_{t-1} = x_{t-1}^2 + \ldots + x_{t-p}^2$$

and the symmetric matrix outer product of prior observations is

$$H_{t-1}H_{t-1}^T = \begin{bmatrix} x_{t-1}^2 & x_{t-1}x_{t-2} & \cdots & x_{t-1}x_{t-p} \\ x_{t-2}x_{t-1} & x_{t-2}^2 & \cdots & x_{t-2}x_{t-p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{t-p}x_{t-1} & \cdots & & x_{t-p}^2 \end{bmatrix}.$$

Define the conditional expectations of ϕ_t prior and posterior, respectively, to the observation x_t as

$$\bar{\phi}_t = E[\phi_t \mid \mathscr{H}_{t-1}^{(x)}] = E[\phi_t \mid x_{t-1}, x_{t-2}, \ldots] \tag{6a}$$

$$\hat{\phi}_t = E[\phi_t \mid \mathscr{H}_t^{(x)}] = E[\phi_t \mid x_t, x_{t-1}, \ldots]. \tag{6b}$$

In terms of (6a) the conditional one-step-ahead forecast of x_t is given by:

$$\bar{x}_t = E[x_t \mid \mathscr{H}_{t-1}^{(x)}] = H_{t-1}^T[\phi_t \mid \mathscr{H}_{t-1}^{(x)}] = H_{t-1}^T\bar{\phi}_t \tag{7}$$

where $\mathscr{H}_t^{(x)}$ is the entire history of real observations of the x-process up to and including time t. Note that the time at which the forecast is made is t^{-1}. $\bar{x}_t$ is a forecast made at time $t-1$ for period t. The covariance matrices for the parameter distributions are defined by

$$\bar{P}_t \equiv \mathrm{Cov}\,[\phi_t, \phi_t^T \mid \mathscr{H}_{t-1}^{(x)}] = E[(\phi_t - \bar{\phi}_t)(\phi_t - \bar{\phi}_t)^T \mid \mathscr{H}_{t-1}^{(x)}] \tag{8a}$$

$$\hat{P}_t \equiv \mathrm{Cov}\,[\phi_t, \phi_t^T \mid \mathscr{H}_t^{(x)}] = E[(\phi_t - \hat{\phi}_t)(\phi_t - \hat{\phi}_t)^T \mid \mathscr{H}_t^{(x)}]. \tag{8b}$$

The model in (5a), (5b) yields an explicit formula for $\hat{P}_t$ which, surprisingly, is independent of the most recent observation, x_t. Thus it follows that the one-step-ahead variance of x_t at time origin $t-1$ is obtained from (5a) and (8a) as

$$\begin{aligned} \bar{v}_t &\equiv \mathrm{Var}\,[x_t \mid \mathscr{H}_{t-1}^{(x)}] \\ &= H_{t-1}^T\,\mathrm{Cov}\,[\phi_t, \phi_t^T \mid \mathscr{H}_{t-1}^{(x)1}]H_{t-1} + \sigma_a^2 \\ &= H_{t-1}^T\bar{P}_tH_{t-1} + \sigma_a^2. \end{aligned} \tag{9}$$

The quadratic form in (9) is a scalar which measures the variance contribution in the uncertainty of the parameters. Covariance terms are missing because noise in the AR equation is assumed independent of noise in the random walk of the parameters. If x_t in (5a) is the measurement equation for a Kalman filter with (5b) as the underlying equation of motion, then the actual observation, x_t is used to update the prior expectations and covariances $\bar{\phi}_t$, $\bar{P}_t$ to obtain the posteriors $\hat{\phi}_t$, $\hat{P}_t$. The latter are then used to calculate the prior forecasts for the next time period. i.e. $\bar{\phi}_{t+1}$, $\bar{P}_{t+1}$ and $\bar{x}_{t+1}$.

From (5b) the expectations and covariances of next period's parameters are

$$\bar{\phi}_{t+1} = E[\phi_{t+1} \mid \mathscr{H}_t^{(x)}] = E[\phi_t \mid \mathscr{H}_t^{(x)}] = \hat{\phi}_t \tag{10a}$$

$$\begin{aligned} \bar{P}_{t+1} &= \mathrm{Cov}\,[\phi_{t+1}, \phi_{t+1}^T \mid \mathscr{H}_t^{(x)}] \\ &= \mathrm{Cov}\,[\phi_t, \phi_t^T \mid \mathscr{H}_t^{(x)}] + \mathrm{Cov}\,[b_{t+1}, b_{t+1}^T \mid \mathscr{H}_t^{(x)}] \\ &= \hat{P}_t + Q. \end{aligned} \tag{10b}$$

If the *a priori* distribution of ϕ_t is Gaussian, e.g.

$$\phi_t \mid \mathscr{H}_{t-1}^{(x)} \sim N(\bar{\phi}_t, \bar{P}_t),$$

then it follows from (5a) and Bayes' Theorem that the *a posteriori* distribution is

$$\phi_t \mid \mathscr{H}_t^{(x)} \sim N(\hat{\phi}_t, \hat{P}_t)$$

with

$$\bar{\phi}_{t+1} = \hat{\phi}_t = \bar{\phi}_t + \frac{1}{\sigma_a^2} \hat{P}_t H_{t-1} e_t, \tag{11a}$$

where

$$\hat{P}_t = \left(\bar{P}_t^{-1} + \frac{1}{\sigma_a^2} H_{t-1} H_{t-1}^T \right)^{-1} \tag{11b}$$

and

$$e_t = x_t - \bar{x}_t = x_t - H_{t-1}^T \bar{\phi}_t. \tag{11c}$$

Equation (11a) shows that the revised parameter contains a correction term proportional to the product of the forecast residual and recent observations.

Notice that if $H_{t-1} \equiv 0$ then the right-hand product terms in (11a) are zero and no corrections are made to $\bar{\phi}_t$ to obtain $\hat{\phi}_t$. In general if $H_t \neq 0$ then H_t^T has a pseudo-inverse

$$(H_t^T)^{-1} = \frac{H_t}{H_t^T H_t}.$$

COMPARISON OF MODELS

To help compare the Kalman filter and the MW model, define the matrix

$$L_t = \frac{1}{\sigma_a^2} \hat{P}_t H_{t-1} H_{t-1}^T; \tag{12a}$$

it follows from (11b) that an equivalent expression is

$$L_t = \hat{P}_t (\hat{P}_t^{-1} - \bar{P}_t^{-1}) = I - \hat{P}_t \bar{P}_t^{-1}. \tag{12b}$$

Since the information matrix of a parameter estimate is approximately equal to the inverse of its covariance matrix, L_t represents the *relative information gain* due to the observation x_t. Note that it is obtained by multiplying $\hat{P}_t^{-1} - \bar{P}_t^{-1}$ by the inverse of $\hat{P}_t^{-1}$. The factor premultiplying the forecast residual in (11a) can now be written as

$$\frac{1}{\sigma_a^2} \hat{P}_t H_{t-1} = \frac{1}{\sigma_a^2} \hat{P}_t H_{t-1} (H_{t-1}^T H_{t-1}^{T^{-1}}) = \frac{L_t H_{t-1}}{H_{t-1}^T H_{t-1}}, \tag{12c}$$

(where the term $\frac{1}{\sigma_a^2} \hat{P}_t H_{t-1}$ is often called the Kalman gain) which, when substituted in the parameter updating equation of (11a), yields

$$\bar{\phi}_{t+1} = \hat{\phi}_t = \bar{\phi}_t + \frac{L_t H_{t-1}}{H_{t-1}^T H_{t-1}} e_t. \tag{13}$$

Once again, e_t is the forecast residual at t and H_{t-1} is the vector of p past observations ending with the observation at $t-1$. Only e_t, directly through x_t, depends on the most recent observation x_t. The vector H_{t-1}, and the normalization $H_{t-1}^T H_{t-1}$ only depends on historical observations up to and including x_{t-1}.

The Kalman filter in (13) takes advantage of the fact that relevant information about the entire historical data, given the assumed state and measurement equations of (5a), (5b), is contained in the covariance matrix $\hat{P}_t$. Thus, truly adaptive behaviour is achieved and the tedious and redundant application of linear regression to reestimate parameters with the occurrence of each new observation is avoided. Perhaps the most useful information is not only the observation-dependent parameter estimates and forecasts, but also the observation-dependent covariance matrix of parameter estimates and the forecast variances obtained at each step.

L_t, the relative information gain obtained with each new observation is in the AR(1) case a scalar between 0 and 1, since the posterior information cannot be less than the prior information and both must be positive (they are reciprocals of scalar variances). In the more general AR(p) case, L_t is a $p \times p$ matrix of rank one satisfying the inequalities,

$$0 = H_t^T 0 H_t \leqslant H_t^T L_{t+1} H_t \leqslant H_t^T I H_t = H_t^T H_t.$$

In those extreme cases where H_t is approximately equal to zero, $\hat{P}_t$ is approximately equal to $\bar{P}_t$ and hence L_t, the relative information gain, is also approximately equal to zero. As a result, the instability problem of the standardized MW model is avoided and the prior estimates of the parameters are just equal to the most recent posterior estimate, i.e. the correction term is small or negligible. At the other extreme, $H_{t-1}^T L_t \cong H_{t-1}^T$ represents maximal information gain (zero prior information) for the elements of ϕ_t corresponding to nonzero elements of H_{t-1}. The limiting case $\bar{P}_t = 0$ with H_{t-1} uniformly nonzero implies $L_t = I$, a rank p matrix. This case is unattainable, since the existence of the finite data history in H_{t-1} is inconsistent with the assumption of a complete lack of prior information.

The asymptotic properties of $\hat{P}_t$, $\bar{P}_t$ and L_t may be worthwhile discussing in order to assess the validity of scalar approximations to these quantities. In the constant parameter case ($Q = 0$), information continually enters the system but never leaves it so that $\hat{P}_t$ will decrease monotonically. L_t will, on average, decrease with each observation and both will eventually approach zero. If the parameter undergoes a random walk ($Q \neq 0$) a sort of equilibrium is reached in which the average information entering per unit time equals the average information leaving per unit time. If Q is sufficiently small then the matrix $\hat{P}_t/\sigma_a^2$ may well be approximated by a constant. However, L_t, being given by (12a) will continue to depend significantly on the recent history of the observations regardless of the constancy of $\hat{P}_t/\sigma_a^2$. It therefore appears that it might be better to approximate $\hat{P}_t/\sigma_a^2$ by a constant than to approximate L_t by a constant. If this is true it suggests that the unstandardized MW formula may be superior in certain instances to the standardized version.

To recast the MW equations in (1a), (1b) using vector notation, define:

$$\bar{\phi}_t^{(\mathrm{MW})} = \begin{bmatrix} \phi_{t1} \\ \phi_{t2} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{tp} \end{bmatrix}; \quad \hat{\phi}_t^{(\mathrm{MW})} = \begin{bmatrix} \phi'_{t1} \\ \phi'_{t2} \\ \cdot \\ \cdot \\ \cdot \\ \phi'_{tp} \end{bmatrix}. \tag{14}$$

The superscript (MW) is a reminder that these vectors are merely prior and posterior coefficients used by MW in their forecasting equation (they do not represent conditional expectations of random parameters). In terms of these quantities and H_t (2a) becomes:

$$\hat{\phi}_t^{(\mathrm{MW})} = \bar{\phi}_t^{(\mathrm{MW})} + 2kH_{t-1}e_t. \tag{15a}$$

Since $H_{t-1}^T H_{t-1} = x_{t-1}^2 + x_{t-2}^2 + \cdots + x_{t-p}^2$ the standardized equation (2b) is:

$$\hat{\phi}_t^{(\mathrm{MW})} = \bar{\phi}_t^{(\mathrm{MW})} + 2\kappa \frac{H_{t-1}e_t}{H_{t-1}^T H_{t-1}}. \tag{15b}$$

Both equations show close resemblance to the updating equations of the Kalman filter, except that $2k$ in (15a) replaces $(\hat{P}_t/\sigma_a^2)$ in (11a), and 2κ in (15b) replaces L_t in (13). Since $\hat{P}_t/\sigma_a^2$ rather than L_t might be approximated by a scalar constant the unstandardized MW formula would be superior to the standardized one. However if this

approximation is made, much important information on the correlation of parameter estimates and the transient response of the Kalman filter is lost. If the motivation behind standardization of data and residuals is to reduce them to dimensionless quantities of the order of unity it might be preferable to divide by an estimate of σ_a^2 rather than by the squared norm of the recent data. With such a normalization the learning constant would have the useful interpretation as a scalar estimate of $\hat{P}_t$, the posterior uncertainty in parameters.

SUMMARY AND CONCLUSIONS

It appears that the general form of the MW formulas which have been proposed in the literature as "adaptive filtering" methods can be rigorously derived from an AR(p) model in which the parameters undergo a random walk with independent increments. The application of Kalman filtering to this model results in parameter estimates that are updated by correction terms which, as a rough approximation, are proportional to the product of the most recent forecast residual and prior observations of the time series itself. The Kalman filter appears superior to the MW formulas because the learning factors are observation-dependent rather than constant over all time. In contrast to the adaptive filtering models proposed by MW, it is possible in the Kalman filter to distinguish the various sources of uncertainty from the terms Q, σ_a^2 and $\bar{P}_t$. Any arbitrariness which may be involved in assigning *a priori* values to quantities such as Q is far outweighed by the benefits of taking proper account of parameter-estimate correlations and weighting recently acquired data so as to produce forecasts with calculable variances and confidence intervals.

ACKNOWLEDGEMENT

This research was supported by the Lawrence Livermore Laboratory under Purchase Order No. 1980709 with the University of California.

REFERENCES

[1] G. E. D. Box and G. M. Jenkins (1976) *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.
[2] P. J. Harrison and C. F. Stevens (1976) Bayesian forecasting. *JRSS* **38**, 205–247.
[3] S. Makridakis and S. C. Wheelwright (1973) An examination of the use of adaptive filtering in forecasting. *Opl Res. Q.* **24**, 55–64.
[4] S. Makridakis and S. C. Wheelwright (1977) Adaptive filtering: An integrated autoregressive/moving average filter for time series forecasting. *Opl Res. Q.* **28**, 425–437.
[5] S. Makridakis and S. C. Wheelwright (1978) *Interactive Forecasting*. Holden-Day, San Francisco.