THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 8, October 1983
Printed in U.S.A.

Forecasting With Adaptive Gradient Exponential Smoothing

By A. FEUER*

(Manuscript received December 20, 1982)

Exponential Smoothing (ES) as a forecasting technique has been extensively used since its introduction in the 1960s. It is simple, hence easy to implement, and in many cases performs surprisingly well. However, many phenomena require a more sophisticated forecasting technique. In this paper we introduce a new forecasting technique, Adaptive Gradient Exponential Smoothing (AGES). This technique extends the classical ES as used on simple data or on data with linear trend. For data with both linear trend and seasonal effects this extension results in a new and more general form of ES, which is presented in this paper. The new forecasting technique is tested on simulated data and some real data of the types mentioned above, and its performance in all these tests is clearly superior to ES. It is shown by analysis and by the experimentations that for certain types of data it does in fact converge to the optimal (in the mean square error sense) forecasts.

I. INTRODUCTION

The need for quick and reliable forecasts of various time series is often encountered in economic and business situations. In the Bell System, forecasting is used to help plan trunk and facilities for the telephone network,[1-3] as well as to project computer workload, to determine staffing levels for operators or service observers, and more.

Many forecasting techniques exist and different time series may require different techniques. In general, there is a clear trade-off between simplicity (resulting in cheaper implementation) and per-

* Bell Laboratories.

formance of the forecasting technique. One of the simplest forecasting techniques, Exponential Smoothing (ES), has surprisingly good performance. This technique was presented originally by Winters[4] and Brown[5] and is described briefly in Section II. In Ref. 6 the optimality properties of ES are studied and we expand on these studies and use the conclusions as the basis for a new technique we introduce here.

In fact, these studies revealed a relationship between the ES and the Autoregressive-Integrated Moving Average (ARIMA) model-fitting-based forecasting suggested by Box and Jenkins.[7] This is further discussed in Section III.

The extensive use of ES clearly indicated that for time series with nonstationary discontinuities or changes in the generating parameters, ES performance is not satisfactory. This prompted a number of researchers to develop the Adaptive Exponential Smoothing (AES) idea. In these techniques the algorithm is supposedly evaluating its own performance and correcting its parameters to obtain improved performance. Recently, the existing AESs (see, for example, Refs. 8 through 11) were reviewed critically by Ekern.[12] One of the points raised in Ref. 12 was that none of the existing AESs is supported by analysis or general performance claims (e.g., optimality). In addition, it should be pointed out that only Roberts' and Reed's AES[11] can be used on data with both linear trend and seasonal effects, while the other AESs are limited to simpler data and have no natural generalization.

In this paper we present a new AES algorithm, which we call Adaptive Gradient Exponential Smoothing (AGES). This technique naturally generalizes to data with both linear trend and seasonal effect. In addition, analysis of AGES for simple data and extensive simulations, using simple as well as more general data, strongly suggests that this technique converges to optimal performance in the mean square error (MSE) sense.

Section II presents ES as commonly used. A new, more general form is developed with a discussion of its optimal properties. The new technique, AGES, is derived and presented in Section III, while the results of experiments with this technique on both real and simulated data are presented in Section IV.

II. EXPONENTIAL SMOOTHING AND ITS OPTIMAL PROPERTIES

First we consider ES as Winters[4] did for three types of data: simple* (S), with linear trend (LT), and with both linear trend and multiplicative seasonal effects (LSM). Common to all the configurations is

*Simple data are of the form $a + n(t)$, where a is a fixed value and $n(t)$ is noise with zero mean.

the following: a time series $\{x(t)\}$ is measured every time interval T (e.g., hour, day, or week), and t is an integer representing the time tT. Then, one is interested in forecasting the value $x(t + 1)$* based on the data available up to and including t, namely $x(0), x(1), \cdots, x(t)$.

If $\hat{x}(t + 1)$ denotes the forecast, carried out at time t for $x(t + 1)$, from Ref. 4 we have (using our own notation for consistency with the discussions in the sequel), for S data:

$$\hat{x}(t + 1) = \alpha x(t) + (1 - \alpha)\hat{x}(t) \tag{1a}$$

$$0 \le \alpha \le 1; \tag{1b}$$

for LT data:

$$\hat{x}(t + 1) = \hat{a}(t) + \hat{b}(t) \tag{2a}$$

$$\hat{a}(t) = \alpha x(t) + (1 - \alpha)[\hat{a}(t - 1) + \hat{b}(t - 1)] \tag{2b}$$

$$\hat{b}(t) = \beta[\hat{a}(t) - \hat{a}(t - 1)] + (1 - \beta)\hat{b}(t - 1) \tag{2c}$$

$$0 \le \alpha, \beta \le 1; \tag{2d}$$

and for LSM data:

$$\hat{x}(t + 1) = (\hat{a}(t) + \hat{b}(t))\hat{c}(t - L + 1) \tag{3a}$$

$$\hat{a}(t) = \alpha \frac{x(t)}{\hat{c}(t - L)} + (1 - \alpha)[\hat{a}(t - 1) + \hat{b}(t - 1)] \tag{3b}$$

$$\hat{b}(t) = \beta[\hat{a}(t) - \hat{a}(t - 1)] + (1 - \beta)\hat{b}(t - 1) \tag{3c}$$

$$\hat{c}(t) = \gamma \frac{x(t)}{\hat{a}(t)} + (1 - \gamma)\hat{c}(t - L) \tag{3d}$$

$$0 \le \alpha, \beta, \gamma \le 1, \tag{3e}$$

where L is the known periodicity of the season.

In all the equations above, the parameters α, β, and γ are called the "smoothing coefficients".

Our first step is to rewrite eq. (1) and, more importantly, eq. (2). This provides the basis for a new form of ES for LSM data, more general than (3). The new form, which is a natural extension of (1) and (2), suggests types of data for which the ES algorithm can result in optimal (in the MSE sense) performance.

Equation (1) can be readily rewritten as

$$\hat{x}(t + 1) = \hat{\theta}_1\hat{x}(t) + (1 - \hat{\theta}_1)x(t), \tag{4a}$$

*Note that we restrict our discussions to one-interval-ahead forecasting with the understanding that it can be generalized to more time intervals ahead.

where clearly

$$\hat{\theta}_1 = 1 - \alpha. \tag{4b}$$

With some algebra one can show that eq. (2) is equivalent to

$$\begin{aligned}\hat{x}(t+1) = {} & \hat{\theta}_1\hat{x}(t) + \hat{\theta}_2\hat{x}(t-1) \\ & + (2-\hat{\theta}_1)x(t) - (1+\hat{\theta}_2)x(t-1),\end{aligned} \tag{5a}$$

where

$$\hat{\theta}_1 = 2 - \alpha(1+\beta) \tag{5b}$$

$$\hat{\theta}_2 = \alpha - 1. \tag{5c}$$

The basic difference between (2) and (5) is that (5) reflects the assumption that the noise-free part of the data $x(t)$ is generated by the difference equation

$$y(t) - 2y(t-1) + y(t-2) = 0, \tag{6}$$

while (2) reflects the assumption that the solution of (6) is

$$y(t) = a + bt. \tag{7}$$

[Note that in (2) $\hat{a}(t)$ is the current estimate of '$a + bt$' and $\hat{b}(t)$ is the current estimate of 'b.']

The ES as given in (3) for LSM data is clearly based on the assumption that the noise-free part of the data has the form

$$y(t) = (a+bt)c(t), \tag{8a}$$

where

$$c(t+L) = c(t). \tag{8b}$$

The difference equation satisfied by (8) is

$$y(t) - 2y(t-L) + y(t-2L) = 0, \tag{9}$$

and the corresponding ES

$$\begin{aligned}\hat{x}(t+1) = {} & \sum_{j=1}^{M} \hat{\theta}_J\hat{x}(t-j+1) - \sum_{j=1}^{M} \hat{\theta}_j x(t-j+1) \\ & + 2x(t-L+1) - x(t-2L+1).\end{aligned} \tag{10}$$

The parameters $\hat{\theta}_j$, $j = 1, \cdots, M$ and the constraints they have to satisfy are discussed later. Also, the claimed correspondence between (9) and (10) will become more apparent in later discussion.

At this point, however, we emphasize that while (7) is the general solution of (6), and thus (2) and (5) are equivalent, (8) is only one of many possible solutions of (9). Hence (10) represents an ES form that is more general than (3).

Similarly, data with linear trend and additive seasonal effects* (LSA) have the underlying difference equation

$$y(t) - y(t-1) - y(t-L) + y(t-L-1) = 0 \qquad (11)$$

and the corresponding ES is

$$\hat{x}(t+1) = \sum_{j=1}^{M} \hat{\theta}_j \hat{x}(t-j+1) - \sum_{j=1}^{M} \hat{\theta}_j x(t-j+1) + x(t) + x(t-L+1) - x(t-L). \qquad (12)$$

To unify and simplify the discussions ahead we introduce the following notation. Let D be a unit delay operator, namely $Dx(t) = x(t-1)$, and let $A(D)$ be a polynomial in D such that

$$A(D) = \begin{cases} 1 & \text{for S data} \\ 2 - D & \text{for LT data} \\ 2D^{L-1} - D^{2L-1} & \text{for LSM data} \\ 1 + D^{L-1} - D^{L} & \text{for LSA data.} \end{cases} \qquad (13)$$

With these definitions (4), (5), (10), and (12) can be unified as

$$\hat{x}(t+1) = \sum_{j=1}^{M} \hat{\theta}_j D^{j-1}(\hat{x}(t) - x(t)) + A(D)x(t), \qquad (14)$$

where $M = 1$ will result in (4) and $M = 2$ in (5).

It should also be pointed out that the ES as given by eq(s). (1) [(2) or (3)] has an implicit assumption in it. The assumption is that one (two or three) coefficient(s) can, in fact, smoothen the data. In other words, M in (14) is equal to one (two or three). However, its general form, (14), allows for a larger number of coefficients to get better approximations.

To observe the optimal properties of the ES forecasts we define the forecast error as

$$e(t) = x(t) - \hat{x}(t) \qquad (15)$$

and use as our criteria for the forecast quality the mean square error (MSE), i.e., $E\{e^2(t)\}$. With this in mind, it is clear that optimal performance is achieved if the $e(t)$ becomes a white noise sequence (i.e., independent and identically distributed with zero mean). Namely, the ES technique, while assuming knowledge of the generating process for the noise-free component of the data, attempts to "whiten" the

*This type of data was not addressed in Ref. 4 and, as far as we know, no form of ES applicable to it was proposed before the one here.

noise component. This attempt implies an underlying assumption that the data are generated through, or at least approximated by, the process

$$[1 - DA(D)]x(t) = \left[1 - \sum_{j=1}^{M} \theta_j D^j\right] \epsilon(t), \tag{16}$$

where $\epsilon(t)$ is a white noise with variance σ^2.

Substituting (14) and (16) into (15) results in

$$\left[1 - \sum_{j=1}^{M} \hat{\theta}_j D^j\right] e(t) = \left[1 - \sum_{j=1}^{M} \theta_j D^j\right] \epsilon(t). \tag{17}$$

This equation satisfied by $e(t)$ is the basis of our claims for correspondence between eqs. (9) and (10), and (11) and (12). Equation (17) immediately suggests the conditions for optimal forecasting. First, to get bounded MSE one must require:

Condition 1: All zeros of the polynomial $[1 - \sum_{j=1}^{M} \hat{\theta}_j \lambda^j]$ are outside the unit circle.

If, in addition, we also require:

Condition 2: $\hat{\theta}_j = \theta_j$, $j = 1, 2, \cdots, M$,

then, clearly, from eq. (17), $e(t)$ will converge to $\epsilon(t)$ and optimal forecasting (in the MSE sense) is achieved.

Remark 1: As we discussed here, the sufficiency of Conditions 1 and 2 is quite obvious; however, they are also necessary. This is argued in Appendix A.

Remark 2: In Ref. 4 α and β for LT data are restricted to interval [0, 1], which corresponds to the set S_2 in Fig. 1. The actual constraints follow from applying Condition 1 to the $M = 2$ case. This results in the set S_1 in Fig. 1, which clearly contains S_2 and is considerably larger. Allowing for a larger constraint set for $\hat{\theta}_1$ and $\hat{\theta}_2$ (or, correspondingly, α, β) will result in more cases for which ES could result in optimal performance.

III. ADAPTIVE GRADIENT EXPONENTIAL SMOOTHING

In the previous section we argued that for data that can be approximated by (16), forecasting with ES of the form (14) can result in optimal performance in the MSE sense. To achieve this, Conditions 1 and 2 must be satisfied. However, while Condition 1 can be satisfied by proper choice of $\hat{\theta}_j$, Condition 2 is, in general, hard to satisfy since the values of θ_j in eq. (16) are not known. Basically, the ARIMA

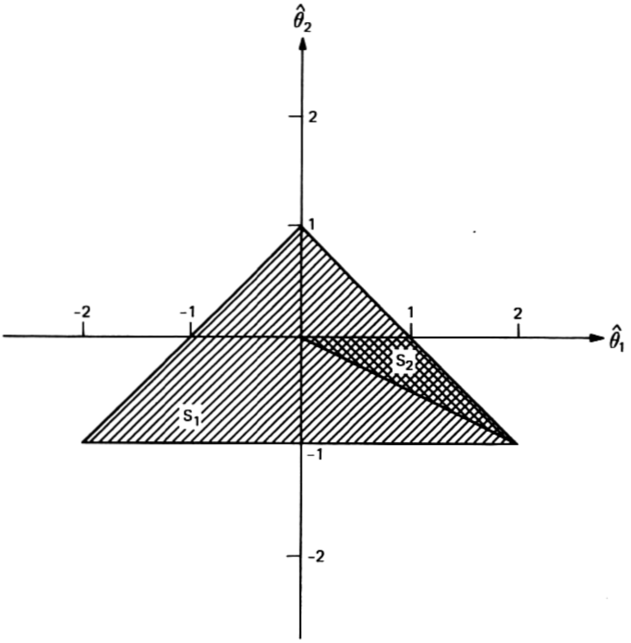


Fig. 1—The constraint sets:

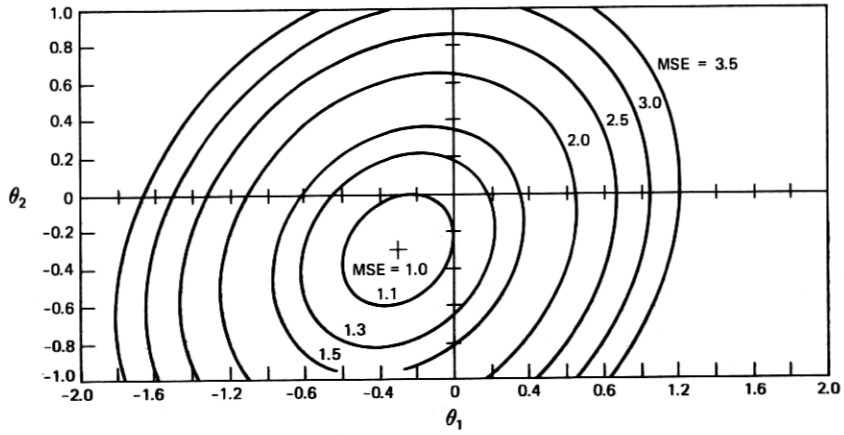
$$S_1 = \{(\hat{\theta}_1, \hat{\theta}_2): |\hat{\theta}_2| < 1, \hat{\theta}_2 + \hat{\theta}_1 < 1, \hat{\theta}_2 - \hat{\theta}_1 < 1\}$$

$$S_2 = \{(\hat{\theta}_1, \hat{\theta}_2): \hat{\theta}_1 = 2 - \alpha(1 + \beta), \hat{\theta}_2 = \alpha - 1, 0 < \alpha, \beta < 1\}.$$

model-fitting-based forecasting[7] deals with exactly this type of problem. The θ_j's of eq. (16) are estimated and these estimates are then used as the $\hat{\theta}_j$'s in eq. (14) in an attempt to satisfy Condition 2. In the ES algorithm no such attempt is made. In practice, the forecasters using ES choose some fixed values for the θ_j, which satisfy Condition 1 [or even more restrictively, e.g., eq. (2d)]. These values are based on intuition, experience, and familiarity with the data they forecast.

However, considerable differences between the underlying θ_j's and the chosen $\hat{\theta}_j$'s can result in significant performance degradation. This is demonstrated in Fig. 2 for the case $M = 2$. The MSE for this case was computed in a closed form as a function of θ_1 and θ_2 for some fixed $\hat{\theta}_1$ and $\hat{\theta}_2$ and graphed in the figure. Together with phenomena like nonstationary discontinuity* and changes in the data-generating process (i.e., the θ_j change values), this resulted in unsatisfactory performance of the ES. The realization of what may cause this poor performance brought about the idea of using adaptive schemes where

* Step-like changes in the data.



MSE – MEAN SQUARED ERROR

Fig. 2—The mean squared error as a function of the data-generating parameters for $M = 2$. (The smoothing coefficients are fixed at $\hat{\theta}_1 = -0.3$, $\hat{\theta}_2 = -0.3$.)

the $\hat{\theta}_j$ are not fixed but are adjusted in an attempt to improve performance.

Compared to the existing Adaptive Exponential Smoothing (AES) techniques (see, e.g., Refs. 8 through 11), the new technique we introduce here is analytically more sound and there are strong indications that it converges to opptimal performance in the MSE sense for the data approximated by (16).

This new technique is based on the gradient search for the minimum of the MSE. If the MSE would have been available as a function of the θ_j, then one could compute the gradient

$$\nabla = \frac{\partial E\{e^2(t)\}}{\partial \hat{\theta}}, \tag{18}$$

where $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_M]^T$, and recursively update the $\hat{\theta}_j$ through

$$\hat{\theta}(t+1) = \hat{\theta}(t) - \mu\nabla, \tag{19}$$

where $\mu > 0$ is the adaptation constant. This is the gradient search technique, sometimes referred to as the steepest descent technique. In general, however, the MSE is not available as a function of the $\hat{\theta}_j$; hence, neither is the gradient. Instead, we use an instantaneous estimate of this gradient. To get this estimate we replace $E\{e^2(t)\}$ by $e^2(t)$ and the gradient by

$$\hat{\nabla} = \frac{\partial e^2(t)}{\partial \hat{\theta}} = 2e(t)\,\frac{\partial e(t)}{\partial \hat{\theta}}. \tag{20}$$

Let us denote

$$s(t) = \frac{\partial e(t)}{\partial \hat{\theta}}$$

as the "sensitivity vector," since it gives an indication of how sensitive the error $e(t)$ is to the values of $\hat{\theta}_j$.

While $s(t)$ is not available we can use eq. (17) to develop a means for generating it. Let us take partial derivatives of both sides of this equation with respect to $\hat{\theta}$. Since the right-hand side does not depend explicitly on $\hat{\theta}$ we get:

$$\left(1 - \sum_{i=1}^{M} \hat{\theta}_i D^{i-1}\right) s_j(t) = D^{j-1}e(t) \qquad j = 1, 2, \cdots, M. \tag{21}$$

At this point we are ready to introduce the Adaptive Gradient Exponential Smoothing (AGES) technique. Combining eqs. (14), (19), (20), and (21) we get:

the *forecast*:

$$\hat{x}(t+1) = A(D)x(t) - \sum_{i=1}^{M} \hat{\theta}_i(t)e(t-i+1) \tag{22}$$

[see definition of $A(D)$ in eq. (13)],

the *sensitivity functions*:

$$s_j(t+1) = \sum_{i=1}^{M} \hat{\theta}_i(t)s_j(t-i+1) + e(t-j+1)$$

$$j = 1, 2, \cdots, M, \tag{23}$$

and the *coefficient adjustments*:

$$\hat{\theta}_j(t+1) = \hat{\theta}_j(t) - 2\mu e(t)s_j(t) \qquad j = 1, 2, \cdots, M. \tag{24}$$

Recall that the error $e(t) = x(t) - \hat{x}(t)$.

Both our simulations and our experiments (as described in the next section) strongly indicate that AGES converges to optimal performance through convergence of $\hat{\theta}_j(t)$ to θ_j. Namely, the error $e(t)$ is adaptively whitened. Despite these indications, since the resulting equations are quite complex, a global proof of convergence of the AGES technique is beyond the scope of this paper. However, we conclude this section by treating the special case $M = 1$ and show local convergence properties for it.

Let $M = 1$; then eqs. (17), (23), and (24) become

$$e(t+1) = \hat{\theta}_1(t)e(t) + \epsilon(t+1) - \theta_1\epsilon(t)$$

$$s_1(t+1) = \hat{\theta}_1(t)s_1(t) + e(t),$$

and

$$\hat{\theta}(t+1) = \hat{\theta}_1(t) - 2\mu e(t) s_1(t).$$

Assuming $\hat{\theta}_j(t)$ is independent of $e(t)$ and $s_1(t)$ (similar assumptions are common in convergence proofs of adaptive filters) and observing that $E\{\epsilon(t)\cdot e(t)\} = \sigma^2$, $E\{\epsilon(t)\cdot\epsilon(t+1)\} = 0$, $E\{s(t)\cdot\epsilon(t)\} = 0$ we get

$$\begin{aligned} E\{e^2(t+1)\} &= E\{\hat{\theta}_1^2(t)\}\cdot E\{e^2(t)\} \\ &\quad + \sigma^2[1 + \theta_1^2 - 2\theta_1 E\{\hat{\theta}_1(t)\}] \\ E\{s_1(t+1)\cdot e(t+1)\} &= E\{\hat{\theta}_1^2(t)\}\cdot E\{s_1(t)\cdot e(t)\} \\ &\quad + E\{\hat{\theta}_1(t)\}\cdot E\{e^2(t)\} - \theta_1 c^2\sigma^2 \\ E\{\hat{\theta}_1(t+1)\} &= E\{\hat{\theta}_1(t)\} - 2\mu E\{s_1(t)e(t)\}. \end{aligned} \tag{25}$$

If we assume in addition that $\hat{\theta}_1(t)$ has a small variance, namely $E\{\hat{\theta}_1^2(t)\} \approx [E\{\hat{\theta}_1(t)\}]^2$ (the simulation results tend to support this assumption), defining

$$\begin{aligned} \gamma_1(t) &= E\{e^2(t)\} - \sigma^2 \\ \gamma_2(t) &= E\{s_1(t)\cdot e(t)\} \\ \gamma_3(t) &= E\{\hat{\theta}_1(t)\} - \theta_1 \end{aligned} \tag{26}$$

and substituting in (25) results in

$$\begin{aligned} \gamma_1(t+1) &= [\gamma_3(t) + \theta_1]^2\gamma_1(t) + \sigma^2[\gamma_3(t)]^2 \\ \gamma_2(t+1) &= [\gamma_3(t) + \theta_1]^2\gamma_2(t) + [\gamma_3(t) + \theta_1]\gamma_1(t) + \sigma^2\gamma_3(t) \\ \gamma_3(t+1) &= \gamma_3(t) - 2\mu\gamma_2(t). \end{aligned} \tag{27}$$

Clearly, if we could prove that $\gamma_1(t)$, $\gamma_2(t)$, and $\gamma_3(t)$ converge to the origin globally (i.e., independent of the initial values), it would mean that [see eq. (26)] the MSE converges to the minimum σ^2 and $E\{\theta_1(t)\}$ converges to θ_1. However, despite strong indications from our simulations that these variables do converge globally, we can prove only local convergence. In addition, the proof provides an indication as to how to choose the parameter μ.

Let us linearize eq. (27) around the origin to get

$$\begin{aligned} \gamma_1(t+1) &= \theta_1^2\gamma_1(t) \\ \gamma_2(t+1) &= \theta_1^2\gamma_2(t) + \theta_1\gamma_1(t) + \sigma^2\gamma_3(t) \\ \gamma_3(t+1) &= \gamma_3(t) - 2\mu\gamma_2(t). \end{aligned} \tag{28}$$

The coefficients matrix is

$$A = \begin{bmatrix} \theta_1^2 & 0 & 0 \\ \theta_1 & \theta_1^2 & \sigma^2 \\ 0 & -2\mu & 1 \end{bmatrix}$$

and to ensure convergence all eigenvalues of A must be within the unit circle. The eigenvalues of A are

$$\lambda_1 = \theta_1^2$$

$$\lambda_{2,3} = 1/2\{1 + \theta_1^2 \pm [(1 - \theta_1^2)^2 - 8\mu\sigma^2]^{1/2}\},$$

and it can be verified that choosing

$$\mu < \frac{1 - \theta_1^2}{2\sigma^2} \tag{29}$$

will guarantee the convergence of eq. (28).

Condition (29) implies that if $|\theta_1|$ is close to one, μ must be chosen very small and the convergence will be slow. Again, our simulation experiments verified this observation.

IV. SIMULATION RESULTS

We divide our experiments with AGES into two parts. In the first part we applied both ES and AGES on data generated by the computer

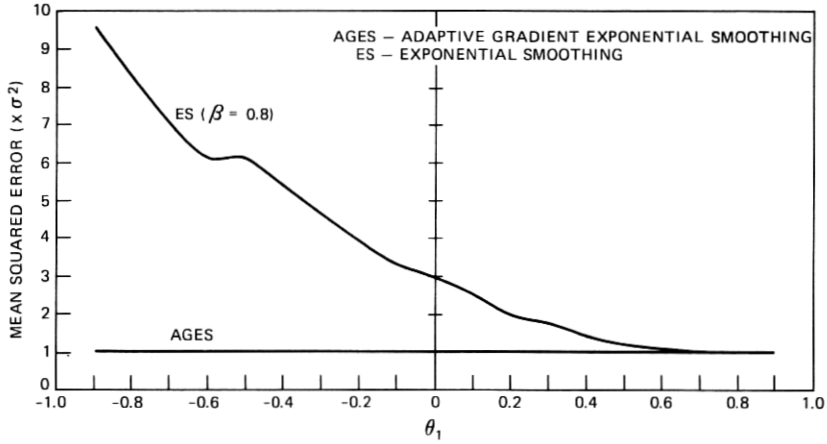


Fig. 3—Comparison of forecasting performance between ES ($\beta = 0.8$) and AGES.

and compared the results. In the second part we applied the AGES to real data that we took from Ref. 7.

Equation (16) was used to generate data of type S, LT, and LSM by the computer. The results of applying both ES and AGES on these data are presented in Figs. 3, 4, and 5 and in Table I. Each point on the curves of Fig. 3 corresponds to a complete run on a sequence of data generated with the particular choice of the θ_i. The resulting MSE for the ES and the AGES forecasts are presented and the comparison clearly indicates the superiority of the AGES algorithm. In addition, we observe that the AGES results, in almost all the runs, in a MSE very close to the minimum, σ^2.

In Fig. 5, we followed the variation of the $\theta_i(t)$ with time in a number of runs. The results clearly show that the $\hat{\theta}_i(t)$ converge to the θ_i from a variety of initial values; this indicates global convergence properties. Similar results are observed in Table I for data with seasonal multiplicative effects and linear trend. The $\hat{\theta}_i(t)$ clearly converge to the θ_i's, and the MSE, when AGES is applied, is again very close to the optimal

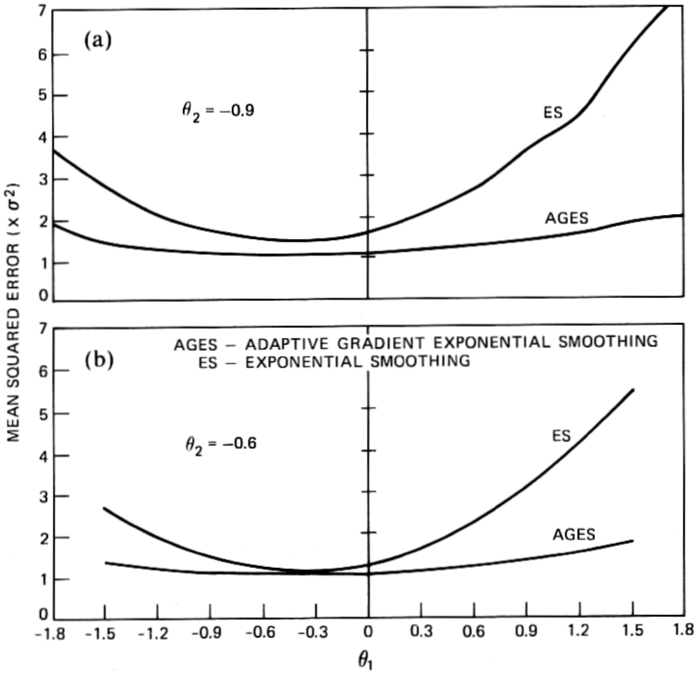


Fig. 4—Comparison of mean squared error in forecasting with ES ($\hat{\theta}_1 = \hat{\theta}_2 = -0.3$) and AGES as a function of the data-generating parameters θ_1 and θ_2. (a) $\theta_2 = -0.9$. (b) $\theta_2 = -0.6$.

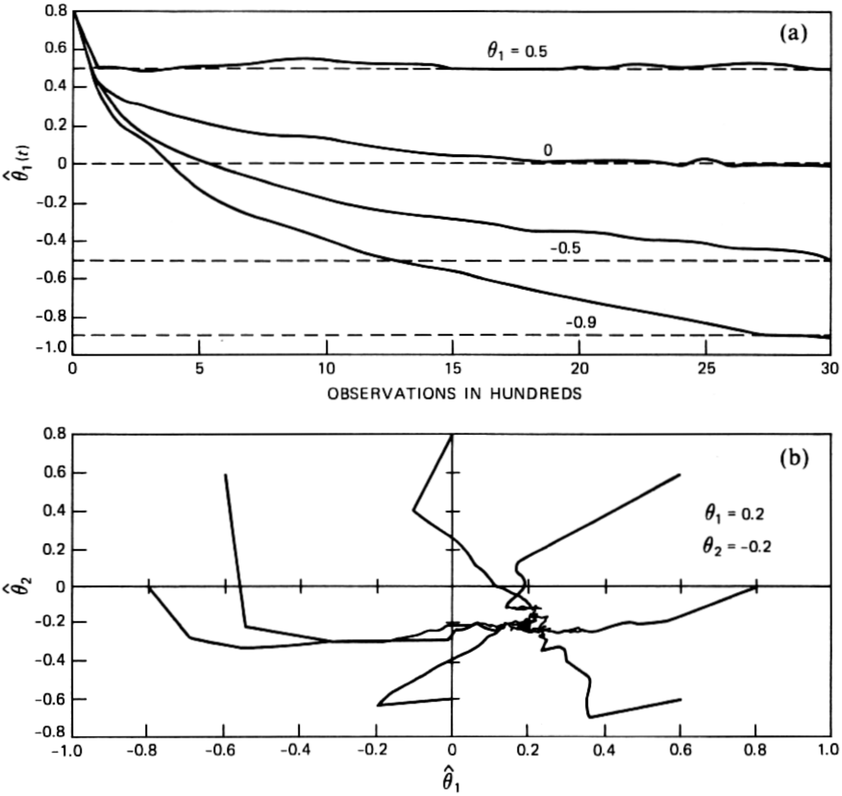


Fig. 5—Convergence of: (a) $\hat{\theta}_1(t)$ to the optimal value θ_1 in the AGES method. (b) $\hat{\theta}_1(t)$ and $\hat{\theta}_2(t)$ from various initial conditions to θ_1 and θ_2, using AGES on data with linear trend.

value, σ^2. From Ref. 7 we took data of the simple kind (no linear trend or seasonal effects): The IBM common stock closing prices, daily, from May 17, 1961 through November 2, 1962. On the data we applied both ES and AGES and the results are presented in Fig. 6. Each point on the curves corresponds to a run on the same data, each time with a different coefficient (for the ES) and different initial condition (for the AGES). The further the coefficient used in the ES is from θ_1 (which in this case is equal to -0.1, as indicated in Ref. 7), the better the performance is for AGES.

Further experiments were conducted on monthly international airline passengers data.[7] These data, as Fig. 7 indicates, are with linear trend and multiplicative seasonal effects. We applied the AGES algorithm (with $M = 3$) and the results are presented in Fig. 8. In Ref. 7 it is claimed that sometimes rather than work with the actual data it

Table I—Comparison of MSE in forecasting with ES ($\hat{\theta}_1 = -0.2$, $\hat{\theta}_2 = 0.5$, $\hat{\theta}_3 = 0.4$) and AGES

Data-Generating Coefficients (and Adaptive Coefficients)			MSE ($\times\sigma^2$)	
θ_1 ($\hat{\theta}_1$)*	θ_2 ($\hat{\theta}_2$)	θ_3 ($\hat{\theta}_3$)	AGES	ES
1.4 (1.39)	−1.3 (−1.29)	0.8 (0.75)	1.1972	11.7155
2.1 (1.97)	−1.95 (−1.79)	0.8 (0.68)	1.3831	20.5821
0.75 (0.69)	−0.6 (−0.56)	0.8 (0.78)	1.0749	5.1431
0.6 (0.6)	−0.75 (−0.76)	0.8 (0.77)	1.0600	5.0256
−0.75 (−0.73)	0.6 (0.6)	0.8 (0.79)	1.0663	1.3577
0.0 (−0.04)	0.0 (0.01)	0.0 (0.03)	1.0040	1.5737
1.0 (0.98)	−1.0 (−0.99)	1.0 (0.94)	1.2001	8.6414
−0.2 (−0.16)	0.5 (0.49)	0.4 (0.4)	1.0397	1.0137
−0.1 (−0.09)	0.25 (0.26)	0.4 (0.41)	0.9973	1.0839
1.2 (1.19)	−0.9 (−0.81)	0.4 (0.36)	1.1044	7.1069
1.8 (1.74)	−1.35 (−1.24)	0.4 (0.34)	1.2192	13.0215
0.3 (0.31)	−0.75 (−0.75)	0.4 (0.38)	1.0574	3.8630
1.0 (0.96)	−0.5 (−0.48)	0.0 (−0.04)	1.0605	4.2218
1.5 (1.42)	−0.75 (−0.66)	0.0 (−0.05)	1.1165	6.9455
0.75 (0.77)	0.0 (0.03)	0.0 (0.03)	1.0212	2.4687
0.0 (−0.03)	−0.75 (−0.74)	0.0 (0.02)	1.0314	3.5907
0.2 (0.19)	0.5 (0.53)	−0.4 (−0.43)	1.0186	1.7692
1.2 (1.19)	−0.15 (−0.14)	−0.4 (−0.39)	1.1115	4.0338
−1.8 (−1.7)	−1.35 (−1.22)	−0.4 (−0.36)	1.2077	13.8494
−0.3 (−0.3)	−0.75 (−0.76)	−0.4 (−0.39)	1.0438	4.6957
−0.75 (−0.79)	−0.3 (−0.28)	−0.4 (−0.38)	1.0062	3.9623
0.4 (0.43)	0.5 (0.49)	−0.8 (−0.78)	1.0461	2.6126
−0.7 (−0.73)	−0.65 (−0.62)	−0.8 (−0.8)	1.0670	6.1628
−0.6 (−0.61)	−0.75 (−0.75)	−0.8 (−0.79)	1.0815	6.8677
−0.75 (−0.74)	−0.6 (−0.56)	−0.8 (−0.74)	1.1065	7.0182
−0.5 (−0.52)	−0.4 (−0.38)	−0.8 (−0.81)	1.0759	5.1676

* The values to which $\hat{\theta}_i(t)$ converge are given in parentheses.

is more convenient to work with the logarithm of the data. As we argue in Appendix B, these logarithms, as data, have linear trend and *additive* seasonal effects (see Fig. 8). Hence, on the logarithms we applied AGES for linear trend and additive seasonal effects and the results are presented in Fig. 8a ($M = 3$). We used the same data (the logarithms) to see whether the performance improves with larger M. AGES was applied with $M = 13$ and the results, as presented in Fig. 8b, clearly indicate that for this data $M = 3$ was sufficient.

V. CONCLUSIONS

In this paper we have introduced a new forecasting technique, Adaptive Gradient Exponential Smoothing (AGES), which is based on Exponential Smoothing (ES). We have elaborated on the optimality properties in the MSE sense of the ES. For certain types of data, the ES can result in optimal performance provided some coefficients are known. In general, these coefficients are unavailable, and the AGES shows strong indications of converging to these unknown coefficients and providing optimal performance.

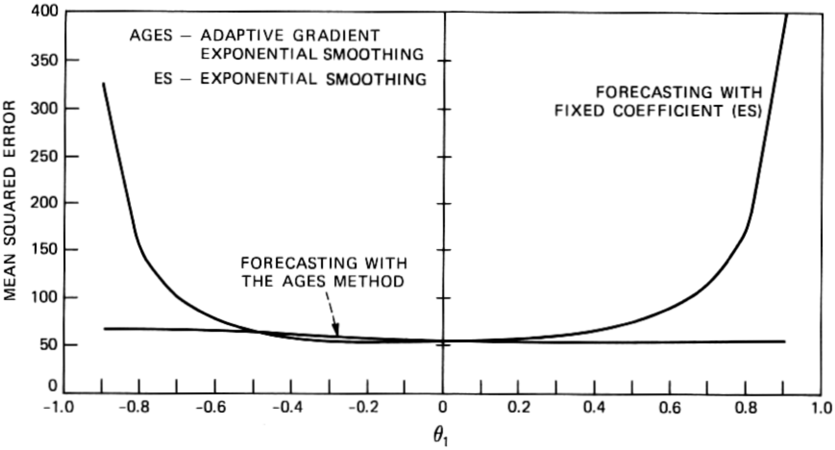


Fig. 6—Comparison of performance of forecasting with the ES (varying the coefficient in each run) and the AGES methods.

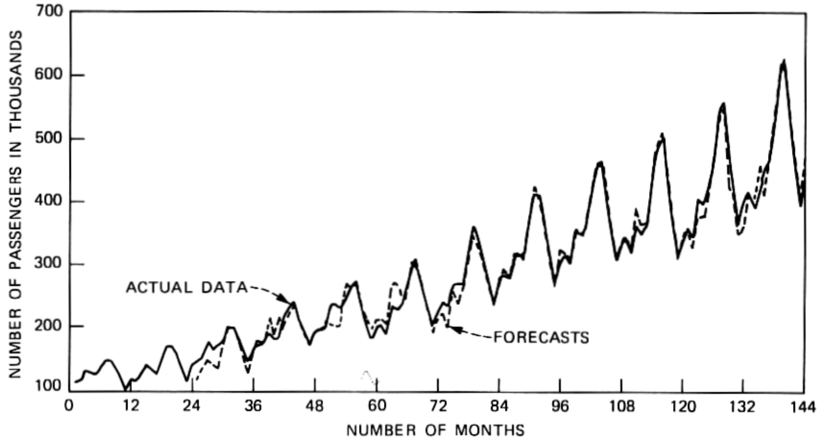


Fig. 7—Forecasting with AGES international airline passengers ($M = 3$). (Note that these data have linear trend and multiplicative seasonal effects.)

Clearly, more extensive experiments and practical use of the proposed forecasting technique, the AGES, are required. A user-friendly software package can be developed for implementation of this technique if sufficient interest is generated.

VI. ACKNOWLEDGMENT

The author wishes to thank the reviewers for their thoroughness. Their comments were very helpful in the revision of this paper.

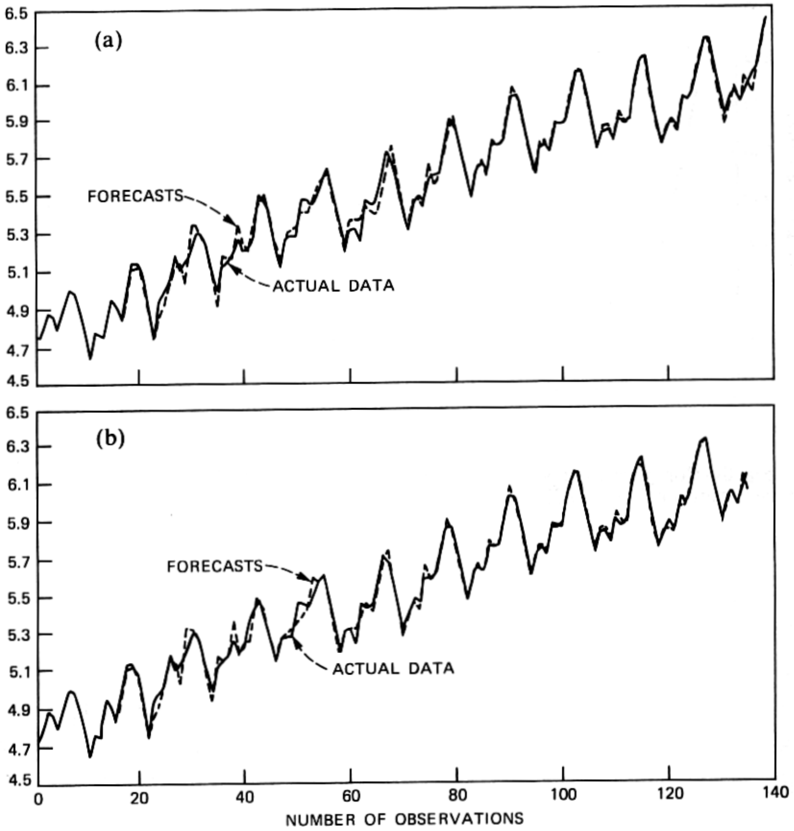


Fig. 8—Forecasting with AGES the logarithm of the data in Fig. 7 for: (a) $M = 3$. (b) $M = 13$. (Note that the logarithm of the data in Fig. 7 has the form of data with additive seasonal effects and linear trend.)

REFERENCES

1. C. D. Pack and B. A. Whitaker, "Kalman Filter Models for Network Forecasting," B.S.T.J., *61,* No. 1 (January 1982), pp. 1–14.
2. J. P. Moreland, "A Robust Sequential Projection for Traffic Load Forecasting," B.S.T.J., *61,* No. 1 (January 1982), pp. 15–38.
3. C. R. Szelag, "A Short-Term Forecasting Algorithm for Trunk Demand Servicing," B.S.T.J., *61,* No. 1 (January 1982), pp. 67–96.
4. P. R. Winters, "Forecasting Sales, by Exponentially Weighted Moving Averages," Management Sciences, *6,* No. 3 (April 1960), pp. 324–42.
5. R. G. Brown, "Smoothing, Forecasting and Prediction of Discrete Time Series," Englewood Cliffs, NJ: Prentice-Hall, 1962.
6. J. F. Muth, "Optimal Properties of Exponentially Weighted Forecasts of Time Series With Permanent and Transitory Components," J. Am. Statis. Assn., *55* (June 1960), pp. 299–306.
7. G. P. Box and G. M. Jenkins, "Time Series Analysis, Forecasting and Control," San Francisco, CA: Holden-Day, 1970.
8. W. M. Chow, "Adaptive Control of the Exponential Smoothing Constant," J. Indust. Eng., *16,* No. 5 (October 1965), pp. 314–7.

9. D. W. Trigg and A. G. Leach, “Exponential Smoothing With an Adaptive Response Rate,” Oper. Res. Quart., *18,* No. 1 (March 1967), pp. 53–60.
10. D. C. Whybark, “A Comparison of Adaptive Forecasting Techniques,” Logist. Transport. Rev., *8* (1973), pp. 13–26.
11. S. D. Roberts and R. Reed, “The Development of a Self-Adaptive Forecasting Technique,” AIIE Trans., *1* (December 1969), pp. 314–22.
12. S. Ekern, “Adaptive Exponential Smoothing Revisited,” J. Oper. Res. Soc., *32,* No. 9 (September 1981), pp. 775–82.

APPENDIX A

Necessity of Conditions 1 and 2 for the Convergence of e(t) *to* ϵ(t) *in Equation (17)*

Condition 1 is clearly necessary (as well as sufficient) for the convergence of $E\{e^2(t)\}$ to a finite value. We want to show that Condition 2 is necessary for $E\{e^2(t)\}$ to converge to σ^2.

Let

$$\gamma_{ee}(\tau) = E\{e(t)\cdot e(t-\tau)\} \tag{30}$$

and

$$\gamma_{ee}(\tau) = E\{e(t)\epsilon(t-\tau)\}. \tag{31}$$

Clearly,

$$\gamma_{ee}(\tau) = \gamma_{ee}(-\tau) \tag{32}$$

and, from eq. (17) and the definition of $\epsilon(t)$

$$\gamma_{e\epsilon}(-\tau) = 0. \tag{33}$$

With these definitions it follows from eq. (17), after transients die, that

$$\begin{aligned}
&\gamma_{e\epsilon}(0) = \sigma^2\\
&\gamma_{e\epsilon}(1) - \hat{\theta}_1\gamma_{ee}(0) = -\theta_1\sigma^2\\
&\gamma_{e\epsilon}(2) - \hat{\theta}_1\gamma_{e\epsilon}(1) - \hat{\theta}_2\gamma_{e\epsilon}(0) = -\theta_2\sigma^2\\
&\vdots\\
&\gamma_{e\epsilon}(M) - \hat{\theta}_1\gamma_{e\epsilon}(M-1) - \cdots - \hat{\theta}_M\gamma_{e\epsilon} = -\theta_M\sigma^2,
\end{aligned}$$

or in matrix form

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0\\ -\theta_1 & 1 & 0 & \cdots & 0\\ -\theta_2 & -\theta_1 & 1 & \cdots & 0\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ -\theta_M & -\theta_{M-1} & \cdot & \cdots & 1 \end{bmatrix} \begin{bmatrix} \gamma_{e\epsilon}(0)\\ \gamma_{e\epsilon}(1)\\ \cdot\\ \cdot\\ \cdot\\ \gamma_{e\epsilon}(M) \end{bmatrix} = -\sigma^2 \begin{bmatrix} -1\\ \theta_1\\ \theta_2\\ \cdot\\ \cdot\\ \cdot\\ \theta_M \end{bmatrix}. \tag{34}$$

Also,

$$\begin{aligned}
&\gamma_{ee}(0) - \hat{\theta}_1\gamma_{ee}(1) - \hat{\theta}_2\gamma_{ee}(2) - \cdots - \hat{\theta}_M\gamma_{ee}(M)\\
&\qquad = \gamma_{e\epsilon}(0) - \theta_1\gamma_{e\epsilon}(1) - \cdots - \theta_M\gamma_{e\epsilon}(M)\\
&\gamma_{ee}(1) - \hat{\theta}_1\gamma_{ee}(0) - \hat{\theta}_2\gamma_{ee}(1) - \cdots - \hat{\theta}_M\gamma_{ee}(M-1)\\
&\qquad = -\theta_1\gamma_{e\epsilon}(0) - \cdots - \theta_M\gamma_{e\epsilon}(M-1)\\
&\gamma_{ee}(2) - \hat{\theta}_1\gamma_{ee}(1) - \hat{\theta}_2\gamma_{ee}(0) - \cdots - \hat{\theta}_M\gamma_{ee}(M-2)\\
&\qquad = -\theta_2\gamma_{ee}(0) - \cdots - \theta_M\gamma_{e\epsilon}(M-2)\\
&\vdots\\
&\gamma_{ee}(M) - \hat{\theta}_1\gamma_{ee}(M-1) - \hat{\theta}_2\gamma_{ee}(M-2) - \cdots - \hat{\theta}_M\gamma_{ee}(0)\\
&\qquad = -\theta_M\gamma_{e\epsilon}(0)
\end{aligned}$$

or again in a matrix form

$$\left\{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0\\ -\hat{\theta}_1 & 1 & 0 & & 0\\ -\hat{\theta}_2 & -\hat{\theta}_1 & 1 & & 0\\ \cdot & \cdot & & & \cdot\\ \cdot & \cdot & & & \cdot\\ \cdot & \cdot & & & \cdot\\ -\hat{\theta}_M & -\hat{\theta}_{M-1} & \cdots & & 1 \end{bmatrix}\right.$$

$$\left. - \begin{bmatrix} 0 & \hat{\theta}_1 & \hat{\theta}_2 & \cdots & \hat{\theta}_{M-2} & \hat{\theta}_{M-1} & \hat{\theta}_M\\ 0 & \hat{\theta}_2 & \hat{\theta}_3 & \cdots & \hat{\theta}_{M-1} & \hat{\theta}_M & 0\\ 0 & \hat{\theta}_3 & \hat{\theta}_4 & \cdots & \hat{\theta}_M & 0 & 0\\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & \hat{\theta}_M & 0 & \cdots & 0 & 0 & 0\\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}\right\} \begin{bmatrix} \gamma_{ee}(0)\\ \gamma_{ee}(1)\\ \gamma_{ee}(2)\\ \cdot\\ \cdot\\ \cdot\\ \gamma_{ee}(M)\\ \gamma_{ee}(M) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -\theta_1 & -\theta_2 & \cdots & -\theta_M\\ -\theta_1 & -\theta_2 & -\theta_3 & & 0\\ -\theta_2 & -\theta_3 & -\theta_4 & & 0\\ \cdot & \cdot & \cdot & & \cdot\\ \cdot & \cdot & \cdot & & \cdot\\ \cdot & \cdot & \cdot & & \cdot\\ \theta_M & 0 & 0 & & 0 \end{bmatrix} \begin{bmatrix} \gamma_{ee}(0)\\ \gamma_{ee}(1)\\ \gamma_{ee}(2)\\ \cdot\\ \cdot\\ \cdot\\ \gamma_{ee}(3) \end{bmatrix}. \qquad (35)$$

Now, if we claim that $e(t)$ converges to $\epsilon(t)$, it means that

$$\gamma_{e\epsilon}(\tau) = \delta(\tau)\sigma^2 = \begin{cases} \sigma^2 & \text{for} \quad \tau = 0 \\ 0 & \text{for} \quad \tau \neq 0 \end{cases}$$

and

$$\gamma_{ee}(\tau) = \delta(\tau)\sigma^2.$$

Then, substituting this in (34) or (35) results in

$$\hat{\theta}_j = \theta_j \qquad \text{for} \quad j = 1, 2, \cdots, M,$$

which is Condition 2. Hence this condition is necessary as claimed.

APPENDIX B

Possible Transformation of Multiplicative Seasonal Effects Into Additive Seasonal Effects

Suppose we are given data of the noise-free form

$$\left.\begin{aligned} y(t) &= (a + bt)c(t) \\ c(t + L) &= c(t) \end{aligned}\right\}, \tag{36}$$

which is with linear trend and multiplicative seasonal effects.

Let

$$z(t) = \text{Log}[y(t)]. \tag{37}$$

Then substitution of (36) gives (if we assume $bt \ll a$, which is true in most real data):

$$\begin{aligned} z(t) &= \log a + \log\left(1 + \frac{b}{a}t\right) + \log c(t) \\ &\approx \log a + \frac{b}{a}t + \log c(t) \\ &\approx \tilde{a} + \tilde{b}t + \tilde{c}(t), \end{aligned} \tag{38}$$

where

$$\tilde{a} = \log a$$

$$\tilde{b} = \frac{b}{a}$$

$$\tilde{c}(t) = \log c(t).$$

Hence $z(t)$ clearly has the form of data with linear trend and *additive* seasonal effects.

AUTHOR

Arie Feuer, B.Sc. (Mechanical Engineering), 1967, M.Sc. (Mechanical Engineering), 1973, Technion, Israel; Ph.D. (Control Systems Engineering), 1978, Yale University; Bell Laboratories, 1978—. Since joining Bell Laboratories Mr. Feuer has been involved in telephone network measurement planning and implementations. He is actively involved in research in control system theory, signal processing, and adaptive systems, and currently looks into their possible applications for network measurements and operations.