# AN APPROACH TO TIME SERIES SMOOTHING AND FORECASTING USING THE EM ALGORITHM

BY R. H. SHUMWAY

*University of California, Davis*

AND D. S. STOFFER

*University of Pittsburgh*

**Abstract.** An approach to smoothing and forecasting for time series with missing observations is proposed. For an underlying state-space model, the EM algorithm is used in conjunction with the conventional Kalman smoothed estimators to derive a simple recursive procedure for estimating the parameters by maximum likelihood. An example is given which involves smoothing and forecasting an economic series using the maximum likelihood estimators for the parameters.

**Keywords and Phrases.** Missing data; Kalman filter; EM algorithm; forecasting; maximum likelihood.

## 1. INTRODUCTION

Many problems which arise in the analysis of data occurring in such diverse fields as air pollution, economics, or sociology require the investigator to work with incompletely specified noisy data. As examples one might mention pollution data where values can be missing on days when measurements were not made or economic data where several different sources are providing partially complete data relating to some given unobserved series of interest. General techniques are needed both for smoothing and interpolating sections of the record with missing values and for constructing reasonable forecasts for future values.

A number of approaches are possible which are based on a variety of modelling assumptions. A linear or non-linear regression model can be assumed for the mean, with an additive error superimposed (cf. Anderson (1971) or Hannan (1970)). The fitting of these models by estimating the parameters of the mean value function leads to smooth interpolations and forecasts which do not depend on observing the underlying data at all time points. For non-stationary series, one may use various exponential smoothing techniques such as those given by Holt (1957), Winters (1960), or Brown (1963) (cf. Makridakis (1978)). Such techniques generally require equally spaced data points and often assume that the parameters in the smoothing procedure can be specified in advance. One might also fit the autoregressive integrated moving average (ARIMA) model advocated by Box and Jenkins (1970) with the forecasts defined as conditional expectations. Smoothing of stationary series can also be approached using a symmetric moving average filter with an appropriate frequency response.

0143-9782/82/04 0253-12 $02.50/0
JOURNAL OF TIME SERIES ANALYSIS Vol. 3, No. 4

It has been suggested by Jones (1966), Morrison and Pike (1977), and others (cf. Kendall (1973)) that the general state-space or Kalman filter model (Kalman (1960), Kalman and Bucy (1961)) might provide an appropriate setting within which to parameterize smoothing and forecasting problems. In this case, the $p \times 1$ idealized vector series of interest $\boldsymbol{x}_t$ is not observed directly but only as a component in the random regression model

$$\boldsymbol{y}_t = M_t \boldsymbol{x}_t + \boldsymbol{v}_t, \qquad t = 1, 2, \ldots, n \tag{1}$$

where $M_t$ is a known $q \times p$ design matrix which expresses the pattern which converts the unobserved stochastic vector $\boldsymbol{x}_t$ into the $q \times 1$ observed series $\boldsymbol{y}_t$. The error or noise terms $\boldsymbol{v}_t$, $t = 1, \ldots, n$ are assumed to be zero-mean uncorrelated normally distributed noise vectors with common $q \times q$ covariance matrix $R$. The random series $\boldsymbol{x}_t$ is assumed to be of primary interest; it is modelled as a first order multivariate process of the form

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \qquad t = 1, \ldots, n, \tag{2}$$

where $\Phi$ is a $p \times p$ transition matrix describing the way the underlying series moves through successive time periods. The process $\boldsymbol{x}_t$ may be non-stationary since we do not make specializing assumptions about the roots of the characteristic equation of $\Phi$. The initial value $\boldsymbol{x}_0$ is assumed to be a normal random vector with mean vector $\boldsymbol{\mu}$ and $p \times p$ covariance matrix $\Sigma$. The $p \times 1$ noise terms $\boldsymbol{w}_t$, $t = 1, \ldots, n$ are zero-mean uncorrelated normal vectors with common covariance matrix $Q$.

The motivation for the model defined by (1) and (2) originates from a desire to account separately for uncertainties in the model as defined by the model error $\boldsymbol{w}_t$ and uncertainties in measurements made on the model as expressed by the measurement noise process $\boldsymbol{v}_t$. It might be helpful to envision (1) as a kind of random effects model for time series, where the effect vector $\boldsymbol{x}_t$ has a correlation structure over time imposed by the multivariate autoregressive model (2). In this context, it is a generalization of the ordinary autoregressive AR model which accounts for observation noise as well as model induced noise. One may regard the $M_t$ as fixed design matrices which define the way we observe the components of the vector $\boldsymbol{x}_t$. For example, in this paper, it provides a convenient method for dealing with the incomplete data problems introduced by missing observations.

The primary aim of a smoothing or forecasting procedure is to estimate the unobserved series $\boldsymbol{x}_t$ for $t = 1, 2, \ldots, n$ (smoothing) and for $t = n+1, n+2, \ldots$ (forecasting) using the observed series $y_1, y_2, \ldots, y_n$. If one knows the values for the parameters $\boldsymbol{\mu}$, $\Sigma$, $\Phi$, $Q$, and $R$ the conventional Kalman smoothing estimators can be calculated as conditional expectations and will have minimum mean square error. This is equivalent to regarding the process $\boldsymbol{x}_t$ as a random parameter vector in the Bayesian sense which depends on the prior values assumed for the parameters.

Since the smoothed values in a Kalman filter estimator will depend on the initial values assumed for the above parameters, it is of interest to consider various ways in which they might be estimated. In most cases this has been accomplished by maximum likelihood techniques involving the use of scoring or Newton–Raphson techniques to solve the nonlinear equations which result from differentiating the log-likelihood function (cf. Gupta and Mehra (1974)). Several examples have been given, notably by Ledolter (1979) and Goodrich and Caines (1979), which demonstrate the feasibility of these methods for several specific cases. The maximum likelihood estimation of the parameters in an autoregressive moving average (ARMA) process by similar methods has been considered by Harvey and Phillips (1979) and by Jones (1980).

The likelihood methods applied in the above references typically have several unattractive features which can be circumvented using the EM (Expectation Maximization) algorithm described in Dempster, et al. (1977). First, the corrections in the successive iterations generally involve calculating the inverse of the matrix of second order partials which can be rather large if there are a significant number of parameters. Furthermore, the successive steps involved in a Newton–Raphson may not necessarily increase the size of the likelihood or one may encounter extremely large steps which actually decrease the likelihood. The EM steps, however, always increase the likelihood and one is guaranteed convergence to a stationary point for an exponential family (cf. Wu (1981)). One may find either a local or global maximum there or one may move indefinitely along a ridge (cf. Boyles (1981)), depending on the shape of the likelihood function. Frequently, the EM equations take on a simple heuristically appealing form in contrast to the highly non-linear appearance of the Newton–Raphson or scoring corrections. Of course, since the matrix of second partials is never computed in the EM procedure, it is not available for providing estimated standard errors; these partials can still be approximated by perturbing the likelihood function in the neighborhood of the maximum. Another disadvantage is that the EM algorithm may converge slowly in the latter stages of the iterative procedure; one may want to switch to another algorithm at this stage.

This paper interprets the tasks of smoothing or forecasting in a missing data context as basically the problem of estimating the random process $\boldsymbol{x}_t$ in the state-space model (1), (2). The conditional means provide a minimum mean squared error solution based on the observed data if the parameters $\boldsymbol{\mu}$, $\Sigma$, $\Phi$, $Q$, and $R$ are known. If the parameters are not specified in advance, they are estimated by maximum likelihood using the EM algorithm. This requires both the conventional recursive forms for the conditional means and covariances and a new recursion which is given in Appendix A. We show also that a very general pattern of missing observations can be tolerated and indicate a correction procedure for adjusting the estimators. An example is given which involves smoothing and forecasting an economic series which is only partially observed by two different sources over the time period of interest.

## 2. MAXIMUM LIKELIHOOD ESTIMATION USING THE EM ALGORITHM

In order to develop a procedure for estimating the parameters in the state-space model defined by (1) and (2), we note first that the joint log likelihood of the complete data $x_0, x_1, \dots, x_n, y_1, \dots, y_n$ can be written in the form

$$\begin{aligned}\log L \doteq & -\tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}(x_0-\mu)'\Sigma^{-1}(x_0-\mu) \\ & -\frac{n}{2}\log|Q| - \tfrac{1}{2}\sum_{t=1}^{n}(x_t-\Phi x_{t-1})'Q^{-1}(x_t-\Phi x_{t-1}) \\ & -\frac{n}{2}\log|R| - \tfrac{1}{2}\sum_{t=1}^{n}(y_t-M_t x_t)'R^{-1}(y_t-M_t x_t)\end{aligned} \tag{3}$$

where $\log L$ is to be maximized with respect to the parameters $\mu$, $\Sigma$, $\Phi$, $Q$, and $R$. Since the log likelihood given above depends on the unobserved data series $x_t$, $t = 0, 1, \dots, n$, we consider applying the EM algorithm conditionally with respect to the observed series $y_1, y_2, \dots, y_n$. That is, define the estimated parameters at the $(r+1)$st iterate as the values $\mu$, $\Sigma$, $\Phi$, $Q$, $R$ which maximize

$$G(\mu, \Sigma, \Phi, Q, R) = E_r(\log L | y_1, \dots, y_n) \tag{4}$$

where $E_r$ denotes the conditional expectation relative to a density containing the $r$th iterate values $\mu(r)$, $\Sigma(r)$, $\Phi(r)$, $Q(r)$, and $R(r)$. An iterative procedure defined as a sequence of such steps has been shown in Dempster, et al. (1977) to yield non-decreasing likelihoods, with the fixed point defined as a stationary point of the likelihood function.

In order to calculate the conditional expectation defined in (4), it is convenient to define the conditional mean

$$x_t^s = E(x_t | y_1, \dots, y_s) \tag{5}$$

and covariance functions

$$P_t^s = \operatorname{cov}(x_t | y_1, \dots, y_s) \tag{6}$$

and

$$P_{t,t-1}^s = \operatorname{cov}(x_t, x_{t-1} | y_1, \dots, y_s). \tag{7}$$

For example, the random vector $x_t^t$ is the usual Kalman filter estimator whereas $x_t^n$, $t = 0, 1, \dots, n$ is the minimum mean square error smoothed estimator of $x_t$ based on all of the observed data. The random vector $x_t^n$ for $t > n$ is the forecast value for the underlying series. A set of recursions for calculating $x_t^n$ and $P_t^n$ from standard Kalman filtering results (cf. Jazwinski (1970)) are given in Appendix A; we also give a new recursive method for computing the covariance $P_{t,t-1}^n$.

Now, taking conditional expectations in (3) yields

$$\begin{aligned}G(\mu, \Sigma, \Phi, Q, R) = & -\tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}\operatorname{tr}\{\Sigma^{-1}(P_0^n + (x_0^n-\mu)(x_0^n-\mu)')\} \\ & -\frac{n}{2}\log|Q| - \tfrac{1}{2}\operatorname{tr}\{Q^{-1}(C - B\Phi' - \Phi B' + \Phi A \Phi')\} \\ & -\frac{n}{2}\log|R| \\ & -\tfrac{1}{2}\operatorname{tr}\left\{R^{-1}\sum_{t=1}^{n}[(y_t - M_t x_t^n)(y_t - M_t x_t^n)' + M_t P_t^n M_t']\right\}\end{aligned} \tag{8}$$

where tr denotes the trace and

$$A = \sum_{t=1}^{n} (P_{t-1}^{n} + x_{t-1}^{n} x_{t-1}^{n\prime}). \tag{9}$$

$$B = \sum_{t=1}^{n} (P_{t,t-1}^{n} + x_{t}^{n} x_{t-1}^{n\prime}), \tag{10}$$

and

$$C = \sum_{t=1}^{n} (P_{t}^{n} + x_{t}^{n} x_{t}^{n\prime}). \tag{11}$$

The Kalman filter terms $x_t^n$, $P_t^n$, and $P_{t,t-1}^n$ are computed under the parameter values $\mu(r)$, $\Phi(r)$, $Q(r)$, $R(r)$ using the recursions in Appendix A. Furthermore, it is easy to see that the choices

$$\Phi(r+1) = BA^{-1}, \tag{12}$$

$$Q(r+1) = n^{-1}(C - BA^{-1}B'), \tag{13}$$

and

$$R(r+1) = n^{-1} \sum_{t=1}^{n} [(y_t - M_t x_t^n)(y_t - M_t x_t^n)' + M_t P_t^n M_t'] \tag{14}$$

maximize the last two lines in the likelihood function (8). The first term is analogous to a single replication of the multivariate normal likelihood so that one may take $\mu_0(r+1) = x_0^n$ and fix the value of $\Sigma$ at some resonable baseline level. The estimation of $\mu$ and $\Sigma$ for a replicated Kalman filter model is considered in Shumway, et al. (1981).

In certain cases one may want to constrain the elements of $\Phi$. For example, under the restriction

$$\Phi F = G, \tag{15}$$

with $F$ and $G$ specified $p \times s$ $(s \leq p)$ matrices, we obtain the constrained estimators

$$\Phi_c = \Phi - (\Phi F - G)(F'A^{-1}F)^{-1}F'A^{-1} \tag{16}$$

and

$$Q_c = Q + (\Phi F - G)(F'A^{-1}F)^{-1}(\Phi F - G)' \tag{17}$$

corresponding to (12) and (13), where the argument $(r+1)$ has been suppressed for notational convenience.

The value of the log likelihood function can be calculated at each stage using the 'innovations' form (cf. Gupta and Mehra (1974))

$$\begin{aligned} \log L \doteq & -\tfrac{1}{2} \sum_{t=1}^{n} \log |M_t P_t^{t-1} M_t' + R_t| \\ & -\tfrac{1}{2} \sum_{t=1}^{n} (y_t - M_t x_t^{t-1})'(M_t P_t^{t-1} M_t' + R_t)^{-1}(y_t - M_t x_t^{t-1}). \end{aligned} \tag{18}$$

A recommended procedure for the EM computations is as follows:

1. Calculate $\boldsymbol{x}_t^n$, $P_t^n$, $P_{t,t-1}^n$ using equations (A3)–(A12) in Appendix A with the initial estimators $\boldsymbol{\mu}(0)$, $\Phi(0)$, $Q(0)$, and $R(0)$.
2. Estimate $\boldsymbol{\mu}_0(1)=\boldsymbol{x}_0^n$ and use equations (12), (13), and (14) to get $\Phi(1)$, $Q(1)$, and $R(1)$ respectively.
3. Repeat 1 and 2 above until the estimates and the log likelihood function (18) are stable.

It should be noted that conventional maximum likelihood estimation procedures as in Gupta and Mehra (1974), Goodrich and Caines (1979), or Jones (1980), use equation (18) in conjunction with conventional non-linear methods such as Newton–Raphson or scoring. Such methods may have one or more of the disadvantages mentioned in Section 1. In contrast, the sequence of estimators defined in equations (12), (13), and (14) are very simple to apply since they are essentially just multivariate regression calculations. The price to be paid for this simplicity is in the additional computational effort needed to calculate the smoothed estimators $\boldsymbol{x}_t^n$, $P_t^n$, and $P_{t,t-1}^n$ required for (12), (13), and (14). This requires that one apply the backward recursions (A8)–(A12), whereas the conventional 'innovations' likelihood (18) only requires the forward recursions (A3)–(A7). Of course, the usual method for calculating the derivatives of the log likelihood function (18) is to set up a system of recursions for the derivatives of $x_t^{t-1}$ and $P_t^{t-1}$ which introduces a comparable amount of computing effort. Furthermore, the end result of interest is the smoothed series $\boldsymbol{x}_t^n$ and its covariance matrix $P_t^n$, which will need to be computed eventually anyway.

While the preceding material treats a missing data problem in the sense that the process $\boldsymbol{x}_t$ of interest in the state-space model (1) and (2) is unobserved, a more severe pattern of missing observations can be tolerated. The next section indicates how to approach the estimation and forecasting problem when elements of the observation vector $\boldsymbol{y}_t$ are missing for certain values of $t$.

## 3. MISSING OBSERVATION MODIFICATIONS

Suppose that at a given step, we define the partition of the $q\times 1$ observation vector $\boldsymbol{y}_t=(\boldsymbol{y}_2^{(1)\prime}, \boldsymbol{y}_t^{(2)\prime})$, where $\boldsymbol{y}_t^{(1)}$ is the $q_1\times 1$ observed portion and $\boldsymbol{y}_t^{(2)}$ is the $q_2\times 1$ unobserved portion. The overall complete data observation equations may now be expressed in the partitioned form

$$\begin{pmatrix}\boldsymbol{y}_t^{(1)}\\ \boldsymbol{y}_t^{(2)}\end{pmatrix}=\begin{pmatrix}M_t^{(1)}\\ M_t^{(2)}\end{pmatrix}\boldsymbol{x}_t+\begin{pmatrix}\boldsymbol{v}_t^{(1)}\\ \boldsymbol{v}_t^{(2)}\end{pmatrix} \tag{16}$$

where $M_t^{(1)}$ and $M_t^{(2)}$ are $q_1\times p$ and $q_2\times p$ matrices and

$$\operatorname{cov}\begin{pmatrix}\boldsymbol{v}_t^{(1)}\\ \boldsymbol{v}_t^{(2)}\end{pmatrix}=\begin{pmatrix}R_{11} & R_{12}\\ R_{21} & R_{22}\end{pmatrix}. \tag{17}$$

For the case where unobserved and observed components have uncorrelated errors ($R_{12}=0$ in (20)) only a very simple modification is needed in the estimation and forecasting equations.

By a somewhat lengthy extension of the usual orthogonality arguments (cf. Stoffer (1982)) one may establish that the equations in Appendix A hold for the missing data case given above if one makes the replacements $y_t' = (y_t^{(1)\prime}, \mathbf{0}')$ and $M_t' = (M_t^{(1)\prime}, 0')$, where 0 denotes a $q_2 \times p$ matrix of zeros. That is, if $y_t$ is incomplete, the filtered and smoothed estimators can be calculated from the usual equations by entering zeros in the observation vector $y_t$ where data is missing and by zeroing out the corresponding row of the design matrix $M_t$. This leads to the smoothed estimators $x_t^{(n)}$ and the covariance functions $P_t^{(n)}$, $P_{t,t-1}^{(n)}$ in the missing data case.

The maximum likelihood estimators as computed in the EM procedure require that one take the conditional expectation of (3) under the assumption that $y_t$ is incompletely observed. In this case, defining the incomplete data as $Y_n^{(1)} = (y_1^{(1)}, y_2^{(1)}, \ldots, y_n^{(1)})$, we need only be concerned with the third term of (3), since the first two terms will have expectations which depend only on $x_t^{(n)}$, $P_t^{(n)}$, and $P_{t,t-1}^{(n)}$. The expectation of the third term will depend on evaluating

$$\begin{aligned} E_r(y_t^{(2)} \mid Y_n^{(1)}) &= E_r[E_r(y_t^{(2)} \mid Y_n^{(1)}, x_t) \mid Y_n^{(1)})] \\ &= E_r[M_t^{(2)} x_t + R_{21} R_{11}^{-1}(y_t^{(1)} - M_t^{(1)} x_t) \mid Y_n^{(1)}] \\ &= M_t^{(2)} x_t^{(n)} + R_{21} R_{11}^{-1}(y_t^{(1)} - M_t^{(1)} x_t^{(n)}) \end{aligned} \tag{21}$$

and

$$E_r(y_t^{(2)} y_t^{(2)\prime} \mid Y_n^{(1)}, x_t) = R_{22} - R_{21} R_{11}^{-1} R_{12} + E_r(y_t^{(2)} \mid Y_n^{(1)}, x_t) E_r(y_t^{(2)} \mid Y_n^{(1)}, x_t)'. \tag{22}$$

Substituting these two expressions in the conditional expectation of (3) with the assumption that $R_{12}=0$ leads to using equation (14) with the following modifications. The missing parts of the $y_t$ vector and the corresponding rows of $M_t$ are replaced by zeros as before. In this case the term which appears in the sum (14) defining $R(r+1)$ will get no contribution from either of the first two terms and one simply adds in the missing data estimate for the covariance from the previous iterate. To be specific, let $y_t' = (y_t^{(1)\prime}, y_t^{(2)\prime})$ as before where the missing data part of the vector at time $t$ is denoted by $y_t^{(2)}$. Then, the contribution to $R(r+1)$ will be of the form

$$C_t = n^{-1}\left\{ \begin{pmatrix} y_t^{(1)} - M_t^{(1)} x_t^{(n)} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} y_t - M_t^{(1)} x_t^{(n)} \\ \mathbf{0} \end{pmatrix}' + \begin{pmatrix} M_t^{(1)} \\ 0 \end{pmatrix} P_t^{(n)} (M_t^{(1)\prime}, 0') + \begin{pmatrix} 0 & 0 \\ 0 & R_{22}(r) \end{pmatrix} \right\}$$

where $R_{22}(r)$ is the submatrix of the estimated covariance matrix at the $r$th step.

## 4. AN EXAMPLE

As one example of the kind of missing data which can be handled, we consider the health series representing total expenditures for physician services as measured by two different sources. Table I, taken from Meltzer, et al. (1980), shows the total physician expenditures by year as measured by the Social Security Administration (SSA) and the Health Care Financing Administration (HCFA).

TABLE I

PHYSICIAN EXPENDITURES (IN MILLIONS)

| Year | [1]SSA $y_{1t}$ | [2]HCFA $y_{2t}$ | Initial $x_t^n$ | Initial $\sqrt{P_t^n}$ | MLE $x_t^n$ | MLE $\sqrt{P_t^n}$ |
|---|---|---|---|---|---|---|
| 1949 | 2 633 | — | 2 582 | 67 | 2 541 | 178 |
| 1950 | 2 747 | — | 2 726 | 66 | 2 711 | 185 |
| 1951 | 2 868 | — | 2 874 | 65 | 2 864 | 186 |
| 1952 | 3 042 | — | 3 055 | 65 | 3 045 | 186 |
| 1953 | 3 278 | — | 3 275 | 65 | 3 269 | 186 |
| 1954 | 3 574 | — | 3 521 | 65 | 3 519 | 186 |
| 1955 | 3 689 | — | 3 753 | 65 | 3 736 | 186 |
| 1956 | 4 067 | — | 4 075 | 65 | 4 063 | 186 |
| 1957 | 4 419 | — | 4 443 | 65 | 4 433 | 186 |
| 1958 | 4 910 | — | 4 873 | 65 | 4 876 | 186 |
| 1959 | 5 481 | — | 5 312 | 65 | 5 331 | 186 |
| 1960 | 5 684 | — | 5 647 | 65 | 5 644 | 186 |
| 1961 | 5 895 | — | 6 001 | 65 | 5 972 | 186 |
| 1962 | 6 498 | — | 6 504 | 65 | 6 477 | 186 |
| 1963 | 6 891 | — | 7 073 | 65 | 7 032 | 185 |
| 1964 | 8 065 | — | 7 871 | 64 | 7 866 | 179 |
| 1965 | 8 745 | 8 474 | 8 566 | 54 | 8 521 | 110 |
| 1966 | 9 156 | 9 175 | 9 261 | 53 | 9 198 | 108 |
| 1967 | 10 287 | 10 142 | 10 212 | 53 | 10 160 | 108 |
| 1968 | 11 099 | 11 104 | 11 250 | 53 | 11 159 | 108 |
| 1969 | 12 629 | 12 648 | 12 661 | 53 | 12 645 | 108 |
| 1970 | 14 306 | 14 340 | 14 228 | 53 | 14 289 | 108 |
| 1971 | 15 835 | 15 918 | 15 752 | 53 | 15 835 | 108 |
| 1972 | 16 916 | 17 162 | 17 194 | 53 | 17 171 | 108 |
| 1973 | 18 200 | 19 278 | 19 073 | 54 | 19 106 | 109 |
| 1974 | — | 21 568 | 21 733 | 64 | 21 675 | 119 |
| 1975 | — | 25 181 | 24 741 | 68 | 25 027 | 120 |
| 1976 | — | 27 931 | 27 573 | 80 | 27 932 | 129 |

[1] Social Security Administration, *Compendium of National Health Data*, January, 1976.
[2] Health Care Financing Administration, *Health Care Financing Review*, Summer, 1979.

Note that the HCFA series is missing for the first sixteen years, whereas the SSA series is not measured for the last three years. The problem is to merge these two separate sources of information $(y_{1t}, y_{2t})$ into an overall estimated expenditure series (smoothing) and then to forecast each series for a given number of years in the future.

In order to motivate the use of the model for this particular case, we assume, first of all, that the two agencies are trying to measure essentially the same expenditure series $x_t$. The extent to which each agency is successful will be reflected by the measurement errors $v_t = (v_{t1}, v_{t2})'$ and their variances and covariances in the matrix $R$. We assume that the measurement errors made by the two agencies are uncorrelated so that $R_{12} = 0$. The pattern for observing the data is fixed by noting that the design matrix $M_t$ can be taken as $(1, 1)'$ when both series are observed (1965–1973), $M_t = (1, 0)'$ when only the SSA series is observed (1949–1964), and $M_t = (0, 1)'$ when only the HFCA series is observed (1974–1976). The underlying series $x_t$ as observed in table I appears to be growing exponentially so that the model given by equation (2) is not an unreasonable one. The 'inflation factor' $\Phi$ may not be constant over time and, in fact, may not be necessarily a reflection of any common cause. The effects of population growth, level of medical care and pure price inflation are clearly present in varying degrees over the time period of interest. The assumption that the parameters vary over time can be investigated by obtaining locally smoothed estimates for $\Phi$, $R$, and $Q$ over a time band, since the Kalman filter recursions given in the appendix allow time varying parameters. In summary, our feeling is that smoothing the data optimally under a specific model is to be preferred to the simpler *ad hoc* exponential smoothing techniques which might be proposed for this situation. The difficulty with classical fixed exponential regression models is that the smoothed values cannot deviate from the pure exponential form. The estimation and smoothing procedure proposed here adjusts separately to the data ($R$) and model ($Q$) errors as well as the growth rate ($\Phi$).

In order to apply the EM procedure, initial values are required for the parameters and these were simply guessed by examining portions of the two completely observed series. It is a good idea to examine several different sets of starting values, since the EM algorithm may reach different kinds of stationary values corresponding, for example, to local rather than global maxima. The initial values of $Q$, $R_{11}$, and $R_{22}$ were taken to be 10 000 which implies a standard error of about 100. The inflation rate was assumed to be approximately 10%

TABLE II
SUMMARY OF SUCCESSIVE EM ITERATES FOR THE MAXIMUM LIKELIHOOD ESTIMATORS

| | | | Iteration | | | |
|---|---|---|---|---|---|---|
| $r$ | $\mu(r)$ | $\varphi(r)$ | $Q(r)$ | $R_{11}(r)$ | $R_{22}(r)$ | $-2 \log L$ |
| 1 | 2500 | 1.100 | 10 000 | 10 000 | 10 000 | 885 |
| 2 | 2417 | 1.114 | 49 837 | 41 583 | 24 105 | 680 |
| 3 | 2396 | 1.116 | 78 153 | 54 666 | 25 486 | 675 |
| 4 | 2383 | 1.116 | 93 513 | 59 958 | 25 580 | 675 |
| 5 | 2374 | 1.116 | 100 571 | 62 483 | 25 384 | 674 |
| 10 | 2342 | 1.116 | 105 152 | 65 725 | 23 920 | 674 |
| 20 | 2279 | 1.116 | 104 814 | 67 760 | 20 971 | 672 |
| 40 | 2277 | 1.116 | 105 115 | 68 636 | 19 394 | 671 |
| 50 | 2276 | 1.116 | 105 097 | 68 663 | 19 354 | 671 |
| 75 | 2277 | 1.116 | 105 115 | 68 675 | 19 329 | 671 |

per year which implies $\varphi = 1.10$. The result of applying these initial values in equations (A3)–(A10) are shown as the initial smoothed estimators. The results of applying equations (9)–(14) are shown in table II, and we note considerable changes over the first five iterations and then a very slow convergence to the final values. The last column shows $-2 \log L$ as computed using the innovations form of the likelihood in equation (18). The final estimators reallocate the errors to the state-space part of the model (where the standard error rises from 100 to 324) and to the observation error in the SSA series (100 to 262). The inflation rate rises to approximately 12% ($\varphi = 1.116$). The change produces larger uncertainties in the estimated smoothed values produced by the maximum likelihood estimators as displayed in the last column of table I.

The computations were performed on a TRS-80, Model III microcomputer with 48K bytes of internal memory and required approximately three minutes per iteration. A copy of the program, written in Basic, is available from the first author.

As a final comment, note that equations (A3)–(A7) can be used to produce forecasts as shown in table III. In this case, note the large uncertainties and generally higher values associated with the forecast function evaluated at the maximum likelihood estimators.

TABLE III
FIVE-YEAR FORECAST FOR PHYSICIAN EXPENDITURES

| | Initial | | MLE | |
|---|---|---|---|---|
| Year | $x_t^n$ | $\sqrt{P_t^n}$ | $x_t^n$ | $\sqrt{P_t^n}$ |
| 1977 | 30 330 | 133 | 31 178 | 355 |
| 1978 | 33 363 | 177 | 34 801 | 512 |
| 1979 | 36 670 | 219 | 38 846 | 657 |
| 1980 | 40 369 | 261 | 43 361 | 802 |
| 1981 | 44 406 | 304 | 48 400 | 952 |

ACKNOWLEDGEMENTS

The authors are indebted to Cyrus Baghelai of Applied Management Sciences, Silver Spring, Maryland, for furnishing the data on physician expenditures.

## APPENDIX

The Kalman smoother estimator

$$x_t^n = E(x_t | y_1, y_2, \ldots, y_n) \tag{A1}$$

for the model defined by equations (1) and (2) is obtained by minimizing the mean square error

$$P_t^n = E[(x_t - x_t^n)(x_t - x_t^n)' | y_1, \ldots, y_n] \tag{A2}$$

and can be calculated recursively using the following equations which are taken from Jazwinski

(1970), pp. 201, 217. (We allow $\Phi_t$, $R_t$, and $Q_t$ in model (1), (2) to vary with time.) For $t = 1, \ldots, n$

$$x_t^{t-1} = \Phi_t x_{t-1}^{t-1} \tag{A3}$$

$$P_t^{t-1} = \Phi_t P_{t-1}^{t-1} \Phi_t' + Q_t \tag{A4}$$

$$K_t = P_t^{t-1} M_t' (M_t P_t^{t-1} M_t' + R_t)^{-1} \tag{A5}$$

$$x_t^t = x_t^{t-1} + K_t(y_t - M_t x_t^{t-1}) \tag{A6}$$

$$P_t^t = P_t^{t-1} - K_t M_t P_t^{t-1} \tag{A7}$$

where we take $x_0^0 = \mu$ and $P_0^0 = \Sigma$. In order to calculate $x_t^n$ and $P_t^n$ one performs the set of backward recursions $t = n, n-1, \ldots, 1$ on the equations

$$J_{t-1} = P_{t-1}^{t-1} \Phi_t' (P_t^{t-1})^{-1} \tag{A8}$$

$$x_{t-1}^n = x_{t-1}^{t-1} + J_{t-1}(x_t^n - \Phi_t x_{t-1}^{t-1}) \tag{A9}$$

$$P_{t-1}^n = P_{t-1}^{t-1} + J_{t-1}(P_t^n - P_t^{t-1}) J_{t-1}'. \tag{A10}$$

We note that equation (10) in the text requires the covariance $P_{t,t-1}^n$ which can be calculated using the backward recursions

$$P_{t-1,t-2}^n = P_{t-1}^{t-1} J_{t-2}' + J_{t-1}(P_{t,t-1}^n - \Phi_t P_{t-1}^{t-1}) J_{t-2}' \tag{A11}$$

for $t = n, n-1, \ldots, 2$ where

$$P_{n,n-1}^n = (I - K_n M_n) \Phi_n P_{n-1}^{n-1}. \tag{A12}$$

The derivation of these relations is somewhat lengthy and can be found in Shumway and Stoffer (1981).

## REFERENCES


ANDERSON, T. W. (1971) *The Statistical Analysis of Time Series.* Wiley: New York.

ANDERSON, T. W. (1978) Repeated Measurements on Autoregressive Processes. *J. Amer. Statist. Ass.* 73, 271–278.

BOX, G. E. P. and G. M. JENKINS (1970) *Time Series Analysis, Forecasting, and Control.* Holden Day: San Francisco.

BOYLES, R. A. (1980) Convergence Results for the EM Algorithm. Technical Report No. 13. Division of Statistics, University of California, Davis.

BROWN, R. G. (1963) *Smoothing, Forecasting, and Prediction of Discrete Time Series.* Prentice Hall: New Jersey.

DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Statist. Soc. B* 39, 1–38.

GOODRICH, R. L. and P. E. CAINES (1979) Linear System Identification from Nonstationary Cross-Sectional Data, *IEEE Trans. Aut. Cont. AC-24* 403–411.

GUPTA, N. K. and R. K. MEHRA (1974) Computational Aspects of Maximum Likelihood Estimation and Reduction in Sensitivity Function Calculations. *IEEE Trans. Aut. Cont. AC-19* 774-783.

HANNAN, E. J. (1970) *Multiple Time Series.* Wiley: New York.

HARVEY, A. C. and G. D. A. PHILLIPS (1979) Maximum Likelihood Estimation of Regression Models with Autoregressive-Moving Average Disturbances. *Biometrika* 66, 49–58.

HOLT, C. C. (1957) Forecasting Seasonal and Trends by Exponentially Weighted Moving Averages. Carnegie Inst. of Technology, Pittsburgh, Pennsylvania.

JAZWINSKI, A. H. (1970) *Stochastic Processes and Filtering Theory.* Academic Press: New York.

JONES, R. H. (1966) Exponential Smoothing for Multivariate Time Series. *J. R. Statist. Soc. B* 1, 241–251.

JONES, R. H. (1980) Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. *Technometrics* 22, 389–395.

KALMAN, R. E. (1960) A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME J. of Basic Eng.* 8, 35–45.

KALMAN, R. E. and R. S. BUCY (1961) New Results in Linear Filtering and Prediction Theory. *Trans. ASME J. of Basic Eng.* 83, 95–108.

KENDALL, M. G. (1973) *Time Series*. Hafner Press: New York.
LEDOLTER, J. (1979) A Recursive Approach to Parameter Estimation in Regression and Time Series Problems. *Comm. Statist. Theor. Meth.* A8, 1227–1245.
MAKRIDAKIS, S. and S. WHEELWRIGHT (1978) *Interactive Forecasting*. Holden Day: San Francisco.
MELTZER, A., C. GOODMAN, K. LANGWELL, J. COSLER, C. BAGHELAI and J. BOBULA, (1980) Develop Physician and Physician Extender Data Bases. G-155, Final Report, Applied Management Sciences, Inc., Silver Spring, MD 20910.
MORRISON, G. W. and D. H. PIKE (1977) Kalman Filtering Applied to Statistical Forecasting, *Management Science* 23, 768–774.
SHUMWAY, R. H., D. E. OLSEN and L. J. LEVY (1981) Estimation and Tests of Hypotheses for the Initial Mean and Covariance in the Kalman Filter Model. *Commun. Statist.-Theor. Meth.* A10(16), 1625–1641.
SHUMWAY, R. H. and D. S. STOFFER (1981) Time Series Smoothing and Forecasting Using the EM Algorithm. Technical Report No. 27, Division of Statistics, University of California, Davis.
STOFFER, D. S. (1982) Estimation of Parameters in a Linear Dynamic System with Missing Observations. Ph.D. Dissertation, University of California, Davis.
WINTERS, P. R. (1960) Forecasting sales by exponentially weighted moving averages. *Management Science* April, 324–342.
WU, C. F. (1981) On the Convergence of the EM Algorithm. Technical Report No. 642, Department of Statistics, University of Wisconsin, Madison.