Journal of Econometrics 23 (1983) 385–400. North-Holland

ALTERNATIVE ALGORITHMS FOR THE ESTIMATION OF DYNAMIC FACTOR, MIMIC AND VARYING COEFFICIENT REGRESSION MODELS*

Mark W. WATSON
Harvard University, Cambridge, MA 02138, USA

Robert F. ENGLE
University of California, San Diego, La Jolla, CA 92093, USA

Received April 1980, final version received March 1983

This paper provides a general approach to the formulation and estimation of dynamic unobserved component models. After introducing the general model, two methods for estimating the unknown parameters are presented. Both are algorithms for maximizing the likelihood function. The first is based on the method of Scoring. The second is the EM algorithm, a derivative-free method. Each iteration of EM requires a Kalman filter and smoother followed by straightforward regression calculations. The paper suggests using the EM methods to quickly locate a neighborhood of the maximum. Scoring can then be used to pinpoint the maximum and calculate the information matrix.

1. Introduction

The use of unobservable variables in economics is widely accepted as a fruitful approach to describing economic phenomena. Early models treated seasonality or measurement error as unobserved components which must be extracted. Other models considered the business cycle or long swing as an unobserved variable which indirectly determined the behavior or observable series. The most successful application was permanent income which, though unmeasurable, explained observed regularities in the data. Recent macroeconomic models abound with variables such as expectations, the real rate of interest and the natural rate of unemployment which are unobservable but which presumably help to explain the process generating observed data. In labor economics ability, 'spunk' and heredity are treated as unobserved determinants of earnings and education.

There are many possibilities for extending such models to other disciplines. In all of these cases, the statistical model is formulated as if the data were

*We thank Andrew Harvey and two referees for valuable comments on an earlier draft of this paper and the National Science Foundation and Harvard Graduate Society for financial support. An earlier version of this paper circulated under the title 'The EM algorithm for dynamic factor and MIMIC models'.

available on the unobservable. A joint distribution of the observable variables is then derived which can serve as a likelihood function for estimating the unknown parameters. In some cases, such as the one described in this paper, it is possible to obtain estimates also of the values of the unobserved components.

Many of the models mentioned above have not been estimated or at least have not fully used the *a priori* restrictions resulting from careful specification of the unobserved components. Even when empirically estimated fully efficiently, the estimation procedures are sufficiently complicated and specialized that slight variations on the specification cannot easily be considered and diagonistic testing is almost unknown.

In this paper a general approach to the formulation and estimation of unobserved component models will be given based upon the state-space model of engineering. In section 2, this model will be presented and discussed. Sections 3 and 4 discuss two methods for maximizing the likelihood function; Scoring is discussed in section 3, while section 4 presents the EM algorithm. Section 5 gives an empirical example.

2. General formulation of the model

All models discussed above are special cases of the 'state-space' model used in engineering to represent a variety of physical processes. In fact, a wide range of models used in econometrics can be viewed as special cases of state-space models as will be shown below. An introduction and comparison between econometric and engineering applications is given in Mehra (1974). The advantage of viewing the models in this way is that general solution concepts are available based upon the likelihood principle and the Kalman filter recursive algorithm.

The state-space model consists of two sets of equations: 'transition' or 'process' equations and 'measurement' equations. The 'transition' equations describe the evolution of $j \times 1$ vector x_t of characteristics of a physical process in response to a $k \times 1$ vector z_t of weakly exogenous or lagged dependent variables and a $m \times 1$ vector v_t of disturbances. The state vector x_t is unobservable and hence corresponds to the unobserved components which are to be isolated. The 'measurement' equations describe the relation between the unobserved state x_t and a $p \times 1$ vector of measurements y_t. The predetermined variables z_t and another vector of disturbances e_t may also enter the measurement equation.

The model can be specified as

$$\underset{j\times 1}{x_t} = \underset{j\times j}{\phi_t}\ \underset{j\times 1}{x_{t-1}} + \underset{j\times k}{\gamma_t}\ \underset{k\times 1}{z_t} + \underset{j\times m}{G_t}\ \underset{m\times 1}{v_t}\,, \tag{1}$$

$$\underset{p\times 1}{y_t} = \underset{p\times j}{\alpha_t}\ \underset{j\times 1}{x_t} + \underset{p\times k}{\beta_t}\ \underset{k\times 1}{z_t} + \underset{p\times 1}{e_t}, \tag{2}$$

and

$$\begin{pmatrix} v_t \\ e_t \end{pmatrix} \sim \mathrm{NI}\left(0, \begin{pmatrix} Q_t & 0 \\ 0 & R_t \end{pmatrix}\right). \tag{3}$$

In most applications some or all of the parameters and covariance matrices are constant over time so that their time subscript can be suppressed.

Table 1 lists some standard econometric models which are easily written in state-space form. The second column of the table shows the relevant parameter restrictions in the state-space model. The final column lists the interpretation of x_t and any other special features. Unless specified otherwise, the parameters of the models (ϕ, γ, G, α, β, Q, and R) are time invariant.

As shown in Schweppe (1965) or Harvey (1982) the likelihood function of the unknown parameters in (1)–(3) is easily formed. Let η_t denote the innovations in y_t [i.e., $y_t - \mathrm{E}(y_t | y_{t-1}, \ldots, y_1, z_t, \ldots, z_1)$] and let H_t denote the variance of η_t. The log likelihood can then be written as

$$L(\theta) = \text{constant} - \tfrac{1}{2} \sum_{t=1}^{T} (\log|H_t| + \eta_t' H_t^{-1} \eta_t),$$

where θ is the vector of unknown parameters. The innovations and their variances are easily calculated using the Kalman filter.

The Kalman filter requires a value of the mean and variance of x_0 as an initialization. Often these values arise naturally. For example, when $\gamma = 0$ and the x process is stationary, the filter is initialized with the unconditional mean and variance of x. When the x process is non-stationary, the likelihood function conditional on the initial state can be formed and the value of the initial state can be estimated as nuisance parameters. A method for estimating the initial state is presented in the next section.

3. Estimation by scoring

Given the data and the form of the likelihood function it is in principle a simple task to maximize the likelihood function with respect to the unknown parameters. Unfortunately in practice this maximization is not so simple as there are usually a large number of parameters and each evaluation of the likelihood function requires an appreciable number of calculations. We focus attention in this paper on two methods for maximizing the likelihood which we have found practical. The first is a generalization of the method of scoring discussed in Pagan (1980). This method uses only first derivatives and will produce asymptotically efficient estimates in one iteration from

Table 1
Some special cases of the state-space model.

Model			Restrictions	Comments
I.	*Univariate models*		$p=1$	
	(a)	Linear regression[a]	$\alpha=0$, or $Q=0$ and $\gamma=0$	x_t vanishes. β is coefficient vector. Kalman Filter produces recursive residuals.
	(b)	ARIMA model[b]	$\beta=0, \alpha=0$	x_t summarizes past info.
	(c)	Linear regression with[c] ARIMA disturbances	$\gamma=0$	$\alpha x_t + e_t$ is disturbance term.
	(d)	Time varying coefficient[d] regression		x_t is vector of time varying coefficients. α_t is vector of exogenous variables.
	(e)	Unobserved components[e]		x_t contains unobserved components (e.g. seasonal and non-seasonal components).
II.	*Multivariate models*		$p>1$	
	(a)	Multivariate regression	$\alpha=0$, or $Q=0$ and $\gamma=0$	Same as I(a).
	(b)	Multivariate ARIMA	$\beta=0, \gamma=0$	Same as I(b).
	(c)	Multivariate regression with ARIMA disturbances	$\gamma=0$	Same as I(c).
	(d)	Factor analysis	$\phi=0, \beta=0, \gamma=0$	x_t are factors. α contains factor loadings.
	(e)	Dynamic factor analysis[f]	$\beta=0, \gamma=0$	
	(f)	MIMIC[g]	$\phi=0$	x_t contains unobservables. y_t are indicators. z_t are causes.
	(g)	Dynamic MIMIC[h]		Same as II(e).

[a]See Brown, Durbin and Evans (1975) and Harvey and Collier (1977) for discussion of recursive residuals and their uses.
[b]See Hannan (1976).
[c]See Harvey and Phillips (1979).
[d]See Chow (1983) and the references therein.
[e]Examples can be found in Pagan (1975) and Engle (1979).
[f]See Geweke (1977).
[g]See Zellner (1970) and Goldberger (1972, 1977).
[h]See Engle and Watson (1981).

consistent initial parameter estimates. The second is based on the EM approach of Dempster, Laird and Rubin (1977). This is a derivative-free method and does not require any evaluations of the likelihood function. Each EM iteration involves a pass through a 'smoother', followed by familiar regression calculations, and is guaranteed to increase the value of the likelihood function.

The well-known method of scoring requires the gradient of $L(\theta)$ and an estimate of the information matrix. The required derivatives can be derived in a straightforward manner and one finds

$$\frac{\partial L}{\partial \theta_i} = \sum_t \frac{\partial L_t}{\partial \theta_i},$$

where

$$\frac{\partial L_t}{\partial \theta_i} = -\tfrac{1}{2}\operatorname{tr}\left(H_t^{-1}\frac{\partial H_t}{\partial \theta_i}\right) - \left(\frac{\partial \eta_t}{\partial \theta_i}\right)' H_t^{-1}\eta_t + \tfrac{1}{2}\eta_t' H_t^{-1}\frac{\partial H_t}{\partial \theta_i}H_t^{-1}\eta_t.$$

The derivatives $\partial H_t/\partial \theta_i$ and $\partial \eta_t/\partial \theta_i$ can be calculated recursively from the Kalman filter equations. Alternatively, numerical derivatives may be used. Engle and Watson (1981) show that the ijth element of the information matrix is given by

$$\mathcal{I}_{ij} = \sum_t \left\{ \tfrac{1}{2}\operatorname{tr}\left(H_t^{-1}\frac{\partial H_t}{\partial \theta_i}H_t^{-1}\frac{\partial H_t}{\partial \theta_j}\right) + \mathrm{E}\left(\left(\frac{\partial \eta_t}{\partial \theta_i}\right)H_t^{-1}\left(\frac{\partial \eta_t}{\partial \theta_j}\right)\right)\right\}. \tag{4}$$

Deleting the expected value operator yields an estimate which can be used for the iterations.

In many applications there may be a large number of weakly exogenous variables, so the β and γ may contain many unknown parameters. A straightforward application of the method of scoring is feasible, but computational gains can be achieved by using a 'zig-zag' procedure. Consider, for example, the DYMIMIC model

$$y_t = \alpha x_t + \beta z_t + e_t, \tag{5}$$

$$x_t = \phi x_{t-1} + \gamma z_t + v_t. \tag{6}$$

If ϕ, α, R, and Q were known, then we could successively substitute (6) into (5) to obtain a multivariate regression model with a complicated, but known error structure. The unknown parameters in β and γ could then be efficiently estimated by generalized least squares. As has been discussed elsewhere [e.g. Harvey and Phillips (1979)] the Kalman filter is a useful computational device for carrying out generalized least squares. Estimation is carried out via the filter by first writing the model in a slightly different form:

$$y_t = [\alpha \quad \tilde{z}_t^p \quad 0]\begin{bmatrix} x_t \\ \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix} + e_t, \tag{7}$$

$$\begin{bmatrix} x_t \\ \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix} = \begin{bmatrix} \phi & 0 & \tilde{z}_t^j \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} v_t, \tag{8}$$

where

$$\tilde{z}_t^j = (z_t' \otimes I_j), \quad \tilde{z}_t^p = (z_t' \otimes I_p),$$

$$\tilde{\beta} = \text{vec}(\beta), \quad \tilde{\gamma} = \text{vec}(\gamma).$$

If the filter is initialized with a vague prior for the unknown elements in $\tilde{\beta}$ and $\tilde{\gamma}$, then the final filtered estimates, $\tilde{\beta}_{T|T}$ and $\tilde{\gamma}_{T|T}$, will be the generalized least squares estimates conditional on the values of α, ϕ, R, and Q.

The 'zig-zag' approach that is suggested and that we have found successful is to fix α, ϕ, R, and Q at their kth iteration values, α^k, ϕ^k, R^k, and Q^k, and then use the filter to determine β^k and γ^k. With β^k and γ^k fixed, the scoring algorithm is then used to find α^{k+1}, ϕ^{k+1}, R^{k+1}, and Q^{k+1}. The procedure is repeated until convergence.

One word of caution is in order. The information matrix will in general not be block-diagonal between the unknown parameters in (β, γ) and (α, ϕ, Q, R). The standard errors computed from the information matrix of (α, ϕ, Q, R) will therefore be incorrect. Once the parameter estimates have converged it is necessary to calculate the entire information matrix for all the unknowns in θ. This is the correct information matrix to use for inference purposes.

An analogue of this method can be used when the initial value, x_0, is assumed to be an unknown constant, rather than a random variable. This assumption is equivalent to carrying out the analysis conditional on the initial state, which is the correct procedure, for instance, if the x_t process is non-stationary. The generalized least squares estimate of x_0 is just the smoothed estimate, $x_{0|T}$, when the filter has been initialized with a vague prior for x_0.[1] Once $x_{0|T}$ has been estimated conditional on θ^k, the filter is initialized with $x_0 = x_{0|T}$ and $P_{0|0} = 0$ (since x_0 is a constant), the likelihood function, etc. is evaluated and θ^{k+1} is formed. The procedure continues until convergence.

While the Scoring Algorithm is attractive because it uses only first derivatives, produces an estimate of the information matrix, and is one step asymptotically efficient from consistent initial estimates, we have found that it may be slow to converge, particularly when starting with poor initial estimates. As each iteration is reasonably expensive this is a serious drawback. Also the method often yields negative values for the variances in

[1]This is equivalent to the method proposed by Rosenberg (1973).

Q during the iterations, and special penalty functions must be employed or the parameters must be transformed to avoid this problem. The EM algorithm, adapted to the problem, avoids these problems.

4. Estimation by the EM algorithm

The EM algorithm is a method for maximizing a likelihood function in the presence of missing observations. It consists of two steps: an estimation and a maximization step which are iterated to convergence. The maximization step calculates the maximum likelihood estimates of all the unknown parameters conditional on a full data set. The estimation step constructs estimates of the sufficient statistics of the problem conditional on the observed data and the parameters. Essentially, the missing observations are estimated based on the parameter values at one step of the iteration and then the likelihood function is maximized assuming that this is the full observable data set in the other.

For data from exponential families, it is particularly easy to implement, and has been used for the static MIMIC model by Chen (1981). In a time series context Watson (1981) has shown how it can be used to obtain exact maximum likelihood estimates for the moving average model. In the DYMIMIC model the maximization step conditional on the data x, y, z is easily accomplished by regression. The estimation step is discussed in detail below where it is shown that careful calculation of sample moments of 'smoothed' values of x will produce the appropriate estimates of the sufficient statistics.

To define the EM algorithm for the DYMIMIC model we first write the model as a system of multivariate regressions and present the estimator that would be used if data on x were observable. We then show how to construct the expected value of the sufficient statistics in the estimator, conditional on the observed data and the current estimate of the parameters.

We begin by rewriting (5) and (6) as

$$y_t = [\tilde{x}_t^p \; \tilde{z}_t^p] \begin{bmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{bmatrix} + e_t,$$

$$x_t = [\tilde{x}_{t-1}^j \; \tilde{z}_t^j] \begin{bmatrix} \tilde{\phi} \\ \tilde{\gamma} \end{bmatrix} + v_t,$$

where

$$\tilde{x}_t^p = [x_t' \otimes I_p], \quad \tilde{x}_{t-1}^j = [x_{t-1}' \otimes I_j],$$

$$\tilde{\alpha} = \text{vec}(\alpha), \qquad \tilde{\phi} = \text{vec}(\phi).$$

and $\tilde{z}_t^p$, $\tilde{z}_t^j$, $\tilde{\beta}$, and $\tilde{\gamma}$ are defined in (7) and (8). It is often the case that the parameters are estimated subject to restrictions (e.g. certain parameters may be restricted to be zero, etc.). For expositional purposes we consider restrictions of the form

$$R_1\begin{pmatrix}\tilde{\alpha}\\ \tilde{\beta}\end{pmatrix}=0, \tag{9}$$

$$R_2\begin{pmatrix}\tilde{\phi}\\ \tilde{\gamma}\end{pmatrix}=0, \tag{10}$$

where R_1 is $l_1\times(jp+kp)$ and R_2 is $l_2\times(j^2+jk)$. [Non-homogeneous linear restrictions and linear restrictions across eqs. (5) and (6) are straightforward generalizations and will not be discussed.] This implies that we can find matrices D_1 and D_2 of dimension $(jp+kp)\times(jp+kp-l_1)$ and $(j^2+jk)\times(j^2+jk-l_2)$, such that

$$\begin{pmatrix}\tilde{\alpha}\\ \tilde{\beta}\end{pmatrix}=D_1\delta_1 \quad \text{and} \quad \begin{pmatrix}\tilde{\phi}\\ \tilde{\gamma}\end{pmatrix}=D_2\delta_2$$

satisfy (9) and (10) for all vectors δ_1 and δ_2. Imposing the constraints we can then write the model (5) and (6) in terms of the unrestricted parameters, δ_1 and δ_2 as

$$y_t=[\tilde{x}_t^p \quad \tilde{z}_t^p]D_1\delta_1+e_t,$$

$$x_t=[\tilde{x}_{t-1}^j \quad \tilde{z}_t^j]D_2\delta_2+v_t.$$

If the x_t data were observed we would then find the efficient estimates of δ_1, δ_2, R, and Q as solutions to the SUR equations,[2]

$$\hat{\delta}_1=[D_1'\hat{A}_1D_1]^{-1}D_1'\hat{B}_1, \tag{11a}$$

$$\hat{\delta}_2=[D_2'\hat{A}_2D_2]D_2'\hat{B}_2, \tag{11b}$$

$$\hat{Q}=(1/T)\sum\hat{v}_t\hat{v}_t', \tag{12a}$$

[2]We are avoiding complications caused by the initial value x_0 by assuming that it is a fixed, known quantity. When x_0 is unknown, approximate maximum likelihood estimates can be found by setting x_0 equal to its unconditional expected value or its expected value conditional on the data. These are analogous to the conditional and unconditional least squares estimation methods for ARIMA models outlined in Box and Jenkins (1976). Both are approximate since they neglect the Jacobian term of the initial observation in the likelihood function. The method to be outlined in the text can easily be adapted for this method.

$$\hat{R} = (1/T)\sum \hat{e}_t \hat{e}_t', \tag{12b}$$

where

$$\hat{A}_1 = \begin{bmatrix} (X'X \otimes \hat{R}^{-1}) & (X'Z \otimes \hat{R}^{-1}) \\ (Z'X \otimes \hat{R}^{-1}) & (Z'Z \otimes \hat{R}^{-1}) \end{bmatrix}, \qquad \hat{B}_1 = \begin{bmatrix} (X'Y \otimes \hat{R}^{-1}) \\ (Z'Y \otimes \hat{R}^{-1}) \end{bmatrix},$$

$$\hat{A}_2 = \begin{bmatrix} (X'_{-1}X_{-1} \otimes \hat{Q}^{-1}) & (X'_{-1}Z \otimes \hat{Q}^{-1}) \\ (Z'X_{-1} \otimes \hat{Q}^{-1}) & (Z'Z \otimes \hat{Q}^{-1}) \end{bmatrix}, \quad \hat{B}_2 = \begin{bmatrix} (X'_{-1}X \otimes \hat{Q}^{-1}) \\ (Z'X \otimes \hat{Q}^{-1}) \end{bmatrix},$$

with

$$W' = (w_1, w_2, \ldots, w_T),$$

$$W'_{-1} = (w_0, w_1, \ldots, w_{T-1}), \quad \text{for any vector } W,$$

$$\hat{e}_t = y_t - [\tilde{x}_t^p \quad \tilde{z}_t^p]\, D_1 \hat{\delta}_1,$$

$$\hat{v}_t = x_t - [\tilde{x}_{t-1}^j \quad \tilde{z}_t^j] D_2 \hat{\delta}_2.$$

The solution can be found by iterating between eqs. (11) and (12).

These estimates cannot be formed because the moment matrices $X'X$, $X'Z$, $X'Y$, $X'_{-1}X_{-1}$, $X'_{-1}Z$, and $X_{-1}X$ are not known. Letting θ be the vector of unknown parameters, the EM algorithm forms θ^{k+1} as the solution to (11) and (12) using the expected value of the moment matrices, conditional on θ^k and the observed data.

The conditional expected values of the moment matrices can be formed by using a 'smoothing' algorithm. The smoothing algorithm is similar in form to a Kalman filter and recursively calculates

$$x_{t|T} = \mathrm{E}(x_t | y_T, y_{T-1}, \ldots, y_1, Z_T, Z_{T-1}, \ldots, Z_1),$$

and

$$P_{t|T} = \mathrm{var}(x_t | y_T, y_{T-1}, \ldots, y_1, Z_T, Z_{T-1}, \ldots, Z_1).$$

Various smoothing algorithms exist and the reader is referred to Anderson and Moore (1979) for a detailed discussion of the algorithms.

The conditional expected value of the moment matrices can now easily be derived. Since $x_{t|T}$ is the conditional expected value of x_t given the observed data, the error $(x_t - x_{t|T})$ is uncorrelated with any observed data. This implies that

$$\mathrm{E}(X'Z | \text{data}, \theta^k) = \hat{X}'_k Z, \tag{13}$$

where

$$\hat{X}_k' = (x_{1|T}^k, x_{2|T}^k, \ldots, x_{T|T}^k),$$

and $x_{t|T}^k$ is the smoothed estimate of x_t using θ^k.

Similarly,

$$\mathrm{E}(X'Y|\text{data}, \theta^k) = \hat{X}_k' Y, \tag{14}$$

and

$$\mathrm{E}(X'_{-1}Z|\text{data}, \theta^k) = \hat{X}'_{-1,k} Z. \tag{15}$$

To find the conditional expected value of $X'X$, note that we can decompose x_t into two uncorrelated parts

$$x_t = x_{t|T}^k + (x_t - x_{t|T}^k),$$

so that

$$\mathrm{E}[x_t x_t'|\text{data}, \theta^k] = x_{t|T}^k x_{t|T}'^k + P_{t|T}^k,$$

where $P_{t|T}^k$ is the conditional variance of x_t. Letting

$$P^k = \sum_{t=1}^{T} P_{t|T}^k \quad \text{and} \quad P_{-1}^k = \sum_{t=0}^{T-1} P_{t|T}^k,$$

$$\mathrm{E}[X'X|\text{data}, \theta^k] = \hat{X}_k' \hat{X}_k + P^k. \tag{16}$$

Similarly,

$$\mathrm{E}[X'_{-1}X_{-1}|\text{data}, \theta^k] = \hat{X}'_{-1,k} \hat{X}_{-1,k} + P_{-1}^k. \tag{17}$$

The only remaining term is $X'_{-1}X$. Its conditional expected value can be derived by using the decomposition 4.5 and noting that $(x_t - x_{t|T})$ is uncorrelated with $x_{t-1|T}$, since $x_{t-1|T}$ is a linear combination of the observed data. This implies that

$$\begin{aligned} &\mathrm{E}(X'_{-1}X|\text{data}, \theta^k) = \hat{X}'_{-1,k} \hat{X}_k + C_1^k, \\ &C_1^k = \sum \mathrm{E}(x_{t-1} - x_{t-1|T}^k)(x_t - x_{t|T}^k)'. \end{aligned} \tag{18}$$

All of the terms necessary for forming the conditional expected value of the moment matrices are produced by the smoother except

$$\mathrm{E}(x_{t-1} - x_{t-1|T}^k)(x_t - x_{t|T}^k)'.$$

This will also be calculated by the smoother if the state vector is augmented to include x_{t-1}.

When forming the estimates $\hat{Q}$ and $\hat{R}$ from eq. (12) the algorithm must take account of the fact that $\hat{v}_t$ and $\hat{e}_t$ are not observed. During the $(k+1)$st maximization step eq. (12) is evaluated using the sufficient statistics given by eqs. (13)–(18). To form $\hat{R}$ we note that

$$\begin{aligned}\hat{e}_t &= (y_t - \hat{\alpha}x^k_{t|T} - \hat{\beta}z_t) - \hat{\alpha}(x_t - x^k_{t|T}) \\ &\equiv e^k_{t|T} - \hat{\alpha}(x_t - x^k_{t|T}),\end{aligned}$$

where the two terms on the right-hand side are uncorrelated, so that

$$\mathrm{E}[\hat{e}_t\hat{e}_t' | \text{data}, \theta^k] = e^k_{t|T}e^{k\prime}_{t|T} + \hat{\alpha}P^k_{t|T}\hat{\alpha}'.$$

We then form $\hat{R}$ during the $(k+1)$st maximization step as[3]

$$\hat{R} = (1/T)\left(\sum e^k_{t|T}e^{k\prime}_{t|T} + \hat{\alpha}P^k\hat{\alpha}'\right). \tag{19}$$

Similarly we can write $\hat{v}_t$ as the sum of two uncorrelated terms

$$\hat{v}_t = (x^k_{t|T} - \hat{\phi}x^k_{t-1|T} - \hat{\gamma}z_t) + [(x_t - x^k_{t|T}) - \hat{\phi}(x_{t-1} - x^k_{t-1|T})].$$

If we denote the first term on the right-hand side by $v^k_{t|T}$, then $\hat{Q}$ at the $(k+1)$st maximization step is

$$\hat{Q} = (1/T)\{\textstyle\sum}(v^k_{t|T}v^{k\prime}_{t|T}) + P^k + \hat{\phi}P^k_{-1}\hat{\phi}' - \hat{\phi}C^k_1 - C^{k\prime}_1\hat{\phi}'\}. \tag{20}$$

The EM algorithm is now completely defined. The estimation step is given by (11) and (12) which give parameter values at step $k+1$ based on moment matrices estimated in step k. Eqs. (13)–(20) define all the moments needed in (11) and (12). Each of these moments can be constructed from the output of one pass through the Kalman smoother. Since this is a recursive algorithm, the required moment matrices can be constructed as the algorithm proceeds, thus avoiding the storage of the data on $\hat{X}$, and the matrices $P_{t|T}$. There is no need to iterate between (11) and (12) during each step of EM. One can merely construct (11) at step $k+1$ using R_k and Q_k. Eq. (12) can then be used to form R_{k+1} and Q_{k+1}. When the E and M steps are iterated to

[3]Alternatively, note that $e^k_{t|T} = \hat{e}_t + \hat{\alpha}(x_t - x^k_{t|T})$, so that $e^k_{t|T}y_t' = \mathrm{E}(\hat{e}_t y_t' | \text{data}, \theta^k) = \mathrm{E}(\hat{e}_t\hat{e}_t' | \text{data}, \theta^k)$. $\hat{R}$ could then be formed using

$$\hat{R} = (1/T)\textstyle\sum e^k_{t|T}y_t'. \tag{19'}$$

convergence the final parameter estimates will satisfy both (11) and (12) by construction.

One step of the EM algorithm therefore involves two SUR calculations and one pass through the Kalman smoother. In contrast, one step of any first derivative method would require at least one pass through the Kalman filter per parameter, as well as construction and inversion of the information matrix.

The algorithm can easily be generalized. For example, non-linear parameter restrictions in the regression models (5) and (6) can easily be incorporated. These restrictions need only be linearized to form restrictions of the form (9) and (10) or their generalizations. The regression calculations (11) and (12) can then be carried out, the restrictions linearized around these new estimates, and so on, until convergence.

A typical case would be autoregressive errors where the Cochran–Orcutt formulation allows calculation of the maximum likelihood parameter estimates under the common factor restrictions. Moving average errors could also be handled using an analogous non-linear least squares approach. The models are still linear in the unobserved data, so that the sufficient statistics are easily estimated using the Kalman smoother. A second extension is to time varying parameters. In this case α becomes α_t which is the observable weakly exogenous data postulated to have a time varying coefficient. Making this substitution, the algorithm is defined in exactly an analogous form.

The EM algorithm has many attractive features. Foremost among these are its computational simplicity and its convergence properties. In our experience, the method finds estimates in the region of the maximum reasonably quickly even from poor initial guesses. The method also has the desirable property that it constructs R and Q to be positive semidefinite matrices and therefore eliminates the need for arbitrary penalty functions to bound the parameter space.

The algorithm also has certain undesirable features. While it does move to a region close to maximum reasonably quickly, it does not have quadratic convergence properties. Once it is close to the maximum it may take quite a few iterations to pinpoint the maximum. It does not produce an estimate of the information matrix or the Score which are useful for inference. Underidentification may also go undetected, as the EM algorithm will merely move to some point on the ridge of the likelihood function.

The most practical method seems to be a mix of EM and Scoring. EM can be used to quickly move the parameters to a neighborhood of the maximum. Scoring can then be used to pinpoint the maximum and calculate an estimate of the information matrix. Any local identification problems will become apparent when the Scoring algorithm attempts to invert the information matrix. The Scoring algorithm can also be used in a straightforward manner to calculate Lagrange Multiplier statistics.

5. Economic application

In this section, the techniques described above will be applied to the estimation of a common factor in wage rate data from several industries in one metropolitan area. For further details, see Engle and Watson (1981). The determination of a wage rate in Los Angeles is assumed to depend upon factors specific to the industry nationally, a factor specific to Los Angeles and common to all sectors, and factors which are specific to both the industry and region. The objective is to obtain a series on the metropolitan wage rate in Los Angeles and thereby observe whether wages are rising or falling relative to the U.S. as a whole.

The econometric specification is quite simple. Let the log of the wage rate in industry i and year t be w_{it} in Los Angeles and n_{it} in the U.S. The log of the unobserved metropolitan wage rate is m_t, and a_{it} are auto-correlated, AR(1) Gaussian disturbances. All data are constructed with mean zero. For each of five sections, the model is

$$w_{it}=\alpha_i m_t+\beta_i n_{it}+a_{it},$$

$$a_{it}=\rho_i a_{it-1}+e_{it}.$$

The metropolitan wage rate was assumed to follow an AR(2) process. Letting e_t be the 5×1 disturbance vector and α the vector of loadings, this dynamic factor analysis model can be written in state space form as

$$\begin{pmatrix} m_t \\ m_{t-1} \\ m_{t-2} \end{pmatrix}=\begin{pmatrix} \phi_1 & \phi_2 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} m_{t-1} \\ m_{t-2} \\ m_{t-3} \end{pmatrix}+\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}v_t,$$

$$\tilde{W}_t=(\alpha \quad \tilde{\alpha} \quad 0)\begin{pmatrix} m_t \\ m_{t-1} \\ m_{t-2} \end{pmatrix}+\begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_5 \end{pmatrix}\tilde{N}_t+e_t,$$

$$Q=(\sigma_m^2), \qquad R=\begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_5^2 \end{pmatrix}$$

where

$$\tilde{w}_{it}=w_{it}-\rho_i w_{it-1}, \qquad \tilde{\alpha}_i=-\rho_i\alpha_i,$$

and

$$\tilde{n}_{it}=n_{it}-\rho_i n_{it-1} \quad \text{for} \quad i=1,\ldots,5, \quad t=2,\ldots,T.$$

To retain the initial observation we write the measurement equations for $t=1$ as

$$\Psi_i w_{i1} = (\Psi_i \alpha_i \quad 0 \quad 0)\begin{pmatrix} m_1 \\ m_0 \\ m_{-1} \end{pmatrix} + \beta_i \Psi_i n_{i1} + e_{i1},$$

where

$$\Psi_i = (1-\rho_i^2)^{\frac{1}{2}}.$$

Table 2 shows the values of the evolution of the parameter estimates from the EM algorithm. The function value shown is the exact value of the likelihood function; the Jacobian term is included. While fifty iterations were required for convergence, each iteration was very inexpensive. Indeed all fifty iterations required less than one CPU minute on a DEC VAX-11/780 computer.

Table 2
Convergence using EM algorithm.

Parameter	Iteration 0	1	5	10	20	30	40	50
ϕ_1	0.900	1.231	1.428	1.492	1.557	1.582	1.558	1.535
ϕ_2	−0.100	−0.362	−0.566	−0.632	−0.688	−0.688	−0.698	−0.634
$\sigma_m^2 \times 10^4$ [a]	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
α_1	1.000	1.417	1.201	1.051	0.901	0.783	0.572	0.456
α_2	1.000	0.763	0.616	0.521	0.468	0.486	0.525	0.544
α_3	1.000	0.728	0.572	0.475	0.419	0.427	0.452	0.455
α_4	1.000	0.543	0.378	0.334	0.283	0.227	0.125	0.072
α_5	1.000	1.066	0.829	0.709	0.607	0.540	0.450	0.402
β_1	1.000	1.102	1.091	1.081	1.072	1.068	1.068	1.070
β_2	1.000	0.935	0.928	0.922	0.917	0.915	0.913	0.912
β_3	1.000	0.894	0.886	0.881	0.876	0.874	0.871	0.870
β_4	1.000	1.045	1.037	1.034	1.031	1.030	1.033	1.035
β_5	1.000	0.981	0.971	0.965	0.958	0.953	0.949	0.948
ρ_1	0.600	0.724	0.725	0.702	0.681	0.711	0.775	0.807
ρ_2	0.600	0.729	0.714	0.723	0.713	0.658	0.397	0.019
ρ_3	0.600	0.684	0.691	0.698	0.690	0.649	0.485	0.263
ρ_4	0.600	0.589	0.554	0.546	0.544	0.564	0.619	0.646
ρ_5	0.600	0.444	0.354	0.333	0.324	0.357	0.507	0.582
$\sigma_1^2 \times 10^4$	1.000	1.086	0.821	0.759	0.813	1.057	1.710	1.997
$\sigma_2^2 \times 10^4$	1.000	0.913	0.911	0.923	0.909	0.811	0.521	0.349
$\sigma_3^2 \times 10^4$	1.000	0.560	0.548	0.575	0.574	0.497	0.281	0.190
$\sigma_4^2 \times 10^4$	1.000	1.115	1.165	1.147	1.159	1.226	1.358	1.399
$\sigma_5^2 \times 10^4$	1.000	1.034	0.939	0.949	0.964	1.020	1.309	1.475
Function value	−155.35	−69.48	−66.94	−66.00	−65.71	−65.64	−63.96	−62.86

[a] σ_m^2 was normalized to identify the factor loadings.

Table 3

Dynamic factor analysis, where $m_t = \phi_1 m_{t-1} + \phi_2 m_{t-2} + v_t$, $w_{it} = \alpha_i m_t + \beta_i n_{it} + a_{it}$, $a_{it} = \rho_i a_{it-1} + e_{it}$, for sectors $i = 1, \ldots, 5$ (standard errors are in parentheses).

	α	β	ρ	$\sigma^2 \times 10^4$	SE
Contract construction	0.456	1.070	0.807	1.997	0.014
	(0.256)	(0.035)	(0.129)	(0.597)	
Durable manufactures	0.544	0.912	0.019	0.349	0.006
	(0.169)	(0.032)	(0.344)	(0.052)	
Non-durable manufactures	0.455	0.870	0.263	0.190	0.004
	(0.142)	(0.027)	(0.289)	(0.089)	
Wholesale trade	0.072	1.035	0.646	1.399	0.011
	(0.146)	(0.019)	(0.186)	(0.405)	
Retail trade	0.402	0.948	0.582	1.475	0.012
	(0.174)	(0.027)	(0.166)	(0.438)	

	ϕ_1	ϕ_2	$\alpha^2 \times 10^4$
Metropolitan component	1.535	−0.634	1.000
	(0.246)	(0.241)	

Table 3 presents the final estimates and the asymptotic standard errors as calculated by the Scoring algorithm. These parameter estimates are superior to the estimates presented in Engle and Watson (1981). Those estimates were produced using only the Scoring algorithm using one-sided numerical derivatives for the score and information matrix. The likelihood function takes on a value of −66.42 using these estimates, while the EM estimates produced a value of −62.86. EM is also cost effective for this model. The cost of fifty EM iterations was roughly the same as one iteration on Scoring. This is the result of the large number of parameters in the model. For simple time varying parameter models, with few unknown parameters, we have found Scoring and EM to be roughly equivalent in terms of cost and performance.

References

Anderson, B.D.O. and J.B. Moore, 1979, Optimal filtering (Prentice Hall, Englewood Cliffs, NJ).

Box, G.E.P. and G.M. Jenkins, 1976, Time series analysis: Forecasting and control, Rev. ed. (Holden-Day, San Francisco, CA).

Brown, R.L., J. Durbin and J.M. Evans, 1975, Techniques for testing the constancy of regression relationships over time, with comments, Journal of the Royal Statistical Society B 37, 149–192.

Chen, C., 1981, The EM approach to the multiple indicators and multiple causes model via the estimation of the latent variable, Journal of the American Statistical Association 76, 704–708.

Chow, G.C., 1982, Random and changing coefficient models, in: M. Intriligator and Z. Griliches, eds., Handbook of econometrics (North-Holland, Amsterdam).

Cooley, T.F. and E.C. Prescott, 1973, Tests of an adaptive regression model, Review of Economics and Statistics 55, 248–256.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39, 1–38.

Engle, R.F. and M.W. Watson, 1981, A one factor multivariate time series model of metropolitan wage rates, Journal of the American Statistical Association 76, 774–781.

Geweke, J., 1977, The dynamic factor analysis of economic time-series models, in: D.J. Aigner and A.S. Goldberger, eds., Latent variables in socio-economic models (North-Holland, Amsterdam).

Goldberger, Arthur S., 1972, Structural equation methods in the social sciences, Econometrica 40, 979–1001.

Goldberger, Arthur S., 1977, Maximum-likelihood estimation of regressions containing unobservable independent variables, in: D.J. Aigner and A.S. Goldberger, eds., Latent variables in socio-economic models (North-Holland, Amsterdam).

Hannan, E.J., 1976, The identification and parametrization of ARMAX and state space forms, Econometrica 44, 713–724.

Harvey, A.C., 1981, Time series models (Halsted Press, New York).

Harvey, A. and P. Collier, 1977, Testing for functional misspecification in regression analysis, Journal of Econometrics 6, 103–120.

Harvey, A. and G.D.A. Phillips, 1979, The estimation of regression models with autoregressive-moving average disturbances, Biometrika 66, 49–58.

Lehman, B., 1981, Ph.D. dissertation (Dept. of Economics, University of Chicago, Chicago, IL).

Mehra, R.K., 1974, Identification in control and economics: Similarities and differences, Annals of Economic and Social Measurement 3, 21–47.

Pagan, A., 1975, A note on the extraction of components from time series, Econometrica 43, 163–168.

Pagan, A.R., 1980, Some identification and estimation results for regression models with stochastically varying coefficients, Journal of Econometrics 13, 341–364.

Rosenberg, B., 1973, The analysis of a cross section of time series by stochastically convergent parameter regression, Annals of Economic and Social Measurement 2, 399–428.

Schweppe, F., 1965, Evaluation of likelihood functions for Gaussian signals, IEEE Transactions on Information Theory 11, 61–70.

Watson, M.W., 1981, Maximum likelihood estimation of a moving average process via the EM algorithm, Mimeo.

Zellner, A., 1970, Estimation of regression relationships containing unobservable variables, International Economic Review 11, 441–454.