Statistica Sinica **7**(1997), 973-991

MIXED MODEL REPRESENTATION OF STATE SPACE MODELS: NEW SMOOTHING RESULTS AND THEIR APPLICATION TO REML ESTIMATION

John V. Tsimikas and Johannes Ledolter

University of Massachusetts and University of Iowa

Abstract: Known results for the general linear mixed model and its special case, the variance components model, are applied to inference in state space models. New state and disturbance smoothing algorithms that accomodate fixed effects and diffuse initial conditions are developed. The algorithms are based on an augmented Kalman filter, and they avoid the backward recursions of standard smoothing algorithms.

The disturbance smoother is used to develop an EM algorithm for REML estimation of variance components in state space models. The EM algorithm for the structural time series model with polynomial trend and additive seasonality is illustrated in detail.

Key words and phrases: Diffuse prior distribution, EM algorithm, Kalman filter, mixed-model prediction, restricted maximum likelihood, state space model, variance components.

1. Introduction

In this paper we formulate the State Space Model (SSM) as a special case of the general linear mixed model. The states, or the disturbances driving the state transition equation, take the place of the random effects in the mixed model. Diffuse initial conditions on the states are included as "fixed" effects and are assigned a flat non-informative prior distribution.

A direct brute-force application of mixed-model theory to the state space model would require the inversion of the covariance matrix of the observations, which is of large dimension. The Kalman Filter (KF), which provides the Choleski decomposition of the inverse of the covariance matrix, is used to overcome this difficulty. In this paper we combine the Kalman Filter with results from mixed model theory, and present a general unifying framework that ties together various special algorithms that have been proposed for state space model inference. Estimation of time-invariant effects, states, signals, disturbances and future observations, as well as their mean square errors, can all be determined from a certain matrix that is obtained by applying the "sweep" operator to the end result of an augmented Kalman filter recursion.

In Section 2 of this paper we review needed general linear mixed-model results. In Section 3 these results are applied to the analysis of the general state space model. In Section 4 we develop recursive EM algorithms for the REML estimates of variance components in state space models. The estimation of structural time series models with polynomial trend and indicator-variable seasonality is discussed in detail. Two data sets illustrate the computations.

2. Review of General Linear Mixed Model Results

2.1. The general linear mixed model

Consider the general linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\pi} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{y}$ is a $T \times 1$ vector of observations, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown fixed effects, $\mathbf{X}$ is the $T \times p$ design matrix for the fixed effects with rank p, $\boldsymbol{\pi}$ is a $q \times 1$ vector of random effects with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}$, $\mathbf{Z}$ is the $T \times q$ design matrix for the random effects, and $\mathbf{e}$ is the random error vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$; in most applications $\mathbf{R} = \sigma^2\mathbf{I}_T$. The random effects $\boldsymbol{\pi}$ and the errors $\mathbf{e}$ are assumed uncorrelated. Hence the covariance matrix of $\mathbf{y}$ is given by $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$.

In variance components models $\mathbf{Z}$ is partitioned as $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_m]$ and $\boldsymbol{\pi}$ is partitioned as $\boldsymbol{\pi} = [\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_m']'$. Each $q_i \times 1$ vector $\boldsymbol{\pi}_i$ has mean $\mathbf{0}$ and covariance matrix $\mathrm{Var}(\boldsymbol{\pi}_i) = \sigma_i^2\mathbf{I}_{q_i}$, and $\mathrm{Cov}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j') = \mathbf{0}$ for all $i \neq j$. Hence $\mathbf{D}$ is a block diagonal matrix with $\sigma_i^2\mathbf{I}_{q_i}$ as its ith diagonal block. Then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\pi} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{m} \mathbf{Z}_i\boldsymbol{\pi}_i + \mathbf{e} \tag{2.2}$$

and $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R} = \sum_{i=1}^{m} \sigma_i^2\mathbf{Z}_i\mathbf{Z}_i' + \sigma^2\mathbf{I}_T$.

2.2. REML inference

REML estimation in the linear mixed model with normal random components, where $\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\pi}, \mathbf{R} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\pi}, \mathbf{R})$ and $\boldsymbol{\pi}|\mathbf{D} \sim N(\mathbf{0}, \mathbf{D})$, assigns $\boldsymbol{\beta}$ an improper flat prior distribution. Integrating the likelihood with respect to $\boldsymbol{\beta}$ and taking the logarithm, one obtains the *restricted loglikelihood*

$$l_R(\mathbf{R}, \mathbf{D}|\mathbf{y}) = \ln L(\mathbf{R}, \mathbf{D}|\mathbf{y}) = -\frac{T-p}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}\mathbf{y}'\mathbf{W}\mathbf{y}, \tag{2.3}$$

where

$$\mathbf{W} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}. \tag{2.4}$$

With a flat prior on $\boldsymbol{\beta}$, and for known covariance matrices $\mathbf{D}$ and $\mathbf{R}$, the posterior distribution of the effects $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$, given the data $\mathbf{y}$, is

$$\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\pi}\end{bmatrix}|\mathbf{y} \sim \mathbf{N}\Big(\begin{bmatrix}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\\ \mathbf{D}\mathbf{Z}'\mathbf{W}\mathbf{y}\end{bmatrix}, \begin{bmatrix}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} & -(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{D}\\ -\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} & \mathbf{D}-\mathbf{D}\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D}\end{bmatrix}\Big) \tag{2.5}$$

(see Sallas and Harville (1981).) Under square error loss the Bayes estimates of $\boldsymbol{\beta}$, $\boldsymbol{\pi}$, the *signal* $\mathbf{s} = E(\mathbf{y}|\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\pi}$ and the noise vector $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\pi}$ are obtained from (2.5) as posterior expectations; their mean square errors are obtained as posterior variances:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \text{ and } \mathrm{MSE}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{2.6}$$

$$\hat{\boldsymbol{\pi}} = \mathbf{D}\mathbf{Z}'\mathbf{W}\mathbf{y} \text{ and } \mathrm{MSE}(\hat{\boldsymbol{\pi}}) = \mathbf{D} - \mathbf{D}\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D} \tag{2.7}$$

$$\hat{\mathbf{s}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\pi}} = \mathbf{y} - \mathbf{R}\mathbf{W}\mathbf{y} \text{ and } \mathrm{MSE}(\hat{\mathbf{s}}) = \mathbf{R} - \mathbf{R}\mathbf{W}\mathbf{R} \tag{2.8}$$

$$\hat{\mathbf{e}} = \mathbf{R}\mathbf{W}\mathbf{y} \text{ and } \mathrm{MSE}(\hat{\mathbf{e}}) = \mathbf{R} - \mathbf{R}\mathbf{W}\mathbf{R}. \tag{2.9}$$

Without the normality assumption $\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$, whereas $\hat{\boldsymbol{\pi}}$, $\hat{\mathbf{s}}$ and $\hat{\mathbf{e}}$ are Best Linear Unbiased Predictors (BLUP); (see Searle et al. (1992), Chapter 7).

For the variance components model in (2.2) we have $\hat{\boldsymbol{\pi}}_i = \sigma_i^2\mathbf{Z}_i'\mathbf{W}\mathbf{y}$ and $\mathrm{MSE}(\hat{\boldsymbol{\pi}}_i) = \sigma_i^2\mathbf{I}_{q_i} - \sigma_i^4\mathbf{Z}_i'\mathbf{W}\mathbf{Z}_i$. The result

$$E(\frac{\boldsymbol{\pi}_i'\boldsymbol{\pi}_i}{q_i}|\mathbf{y}) = \sigma_i^2 + \frac{\sigma_i^4[\mathbf{y}'\mathbf{W}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{W}\mathbf{y} - \mathrm{tr}(\mathbf{Z}_i'\mathbf{W}\mathbf{Z}_i)]}{q_i}. \tag{2.10}$$

provides the basis for the EM algorithm for REML estimation of $[\sigma^2, \sigma_1^2, \ldots, \sigma_m^2]$. The estimates of the ith variance component ($i = 1, \ldots, m$) and of σ^2 at iteration $r+1$ are obtained from the equations

$$\sigma_i^2(r+1) = \sigma_i^2(r) + \frac{\sigma_i^4(r)[\mathbf{y}'\mathbf{W}(r)\mathbf{Z}_i\mathbf{Z}_i'\mathbf{W}(r)\mathbf{y} - \mathrm{tr}(\mathbf{Z}_i'\mathbf{W}(r)\mathbf{Z}_i)]}{q_i}$$

$$\sigma^2(r+1) = \sigma^2(r) + \frac{\sigma^4(r)[\mathbf{y}'\mathbf{W}(r)\mathbf{W}(r)\mathbf{y} - \mathrm{tr}(\mathbf{W}(r))]}{T-p}; \tag{2.11}$$

our notation $\mathbf{W}(r)$ reflects the fact that $\mathbf{W}$ depends on the previous estimates $\sigma_i^2(r)$.

Mixed-model inference depends heavily on the efficient calculation of $\mathbf{W}$ in (2.4), as well as matrices of the form $\mathbf{Z}'\mathbf{W}\mathbf{y}$ and $\mathbf{Z}'\mathbf{W}\mathbf{Z}$. The very same matrices also arise in the first and second derivatives of the restricted loglikelihood

function, which play an important role in the Newton-Raphson, scoring, and EM algorithms for REML estimates, as well as in the asymptotic covariance matrix of the REML estimates. For the variance components model in (2.2) the derivatives with respect to the parameters $[\sigma^2, \sigma_1^2, \ldots, \sigma_m^2]$ include terms like $\mathbf{Z}_i^{'}\mathbf{W}\mathbf{y}$ and $\mathbf{Z}_i^{'}\mathbf{W}\mathbf{Z}_j$; (see Searle et al. (1992), page 252). These terms also arise in the Parametric Empirical Bayes approach of Kass and Steffey (1989), which improves the MSE approximation of the variance estimates when estimates are used in place of unknown variance components.

3. Mixed Model Inference in the State Space Model

3.1. The state space model

Consider the univariate SSM with both fixed and random time–invariant regressors

$$\begin{aligned} y_t &= \mathbf{x}_t^{'}\boldsymbol{\beta} + \mathbf{z}_t^{'}\boldsymbol{\gamma} + \mathbf{h}_t^{'}\boldsymbol{\alpha}_t + \nu_t \\ \boldsymbol{\alpha}_t &= \boldsymbol{\Phi}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{\xi}_t, \end{aligned} \tag{3.1}$$

where $t = 1, ..., T$. The vector $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{\gamma}$ is a $g \times 1$ vector of time–invariant random effects; $\mathbf{x}_t^{'}$ is the $1 \times p$ design vector for the fixed effects and $\mathbf{z}_t^{'}$ is the $1 \times g$ design vector for the random effects $\boldsymbol{\gamma}$; $\mathbf{h}_t^{'}$ is a $1 \times q$ vector and $\boldsymbol{\alpha_t}$ is the $q \times 1$ state vector at time t; ν_t is the observation noise, $\boldsymbol{\Phi}_t$ is a $q \times q$ transition matrix, and $\boldsymbol{\xi}_t$ is the disturbance term in the state transition equation. The ν_t's are uncorrelated and distributed with mean 0 and variance σ^2, the $\boldsymbol{\xi_t}$'s are uncorrelated and distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Xi}_t$, and disturbance and observation noise terms are assumed uncorrelated. The vector of time–invariant random effects has mean $\mathbf{0}$ and covariance matrix $\mathbf{B}_{11}$. The initial state $\boldsymbol{\alpha}_0$ may contain both a diffuse and a random part; let us partition $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{10}^{'}, \boldsymbol{\alpha}_{20}^{'})^{'}$, where the $q_1 \times 1$ vector $\boldsymbol{\alpha}_{10}$ has a diffuse prior and the $q_2 \times 1$ vector $\boldsymbol{\alpha}_{20}$ has a proper prior distribution with mean $\mathbf{0}$ and $q_2 \times q_2$ covariance matrix $\mathbf{B}_{22}$. Our model also allows for correlation between the random part of the initial state and the time–invariant random effects; assume that $\mathrm{cov}(\boldsymbol{\gamma}, \boldsymbol{\alpha}_{20}) = \mathbf{B}_{12}$.

The initial state conditions are moved into the observation equation. Defining $\boldsymbol{\alpha}_t^{\dagger} = \boldsymbol{\alpha}_t - \prod_{i=0}^{t-1} \boldsymbol{\Phi}_{t-i}\boldsymbol{\alpha}_0$, we express the SSM as

$$\begin{aligned} y_t &= \mathbf{x}_t^{'}\boldsymbol{\beta} + (\mathbf{h}_t^{'}\prod_{i=0}^{t-1}\boldsymbol{\Phi}_{t-i})\boldsymbol{\alpha}_0 + \mathbf{z}_t^{'}\boldsymbol{\gamma} + \mathbf{h}_t^{'}\boldsymbol{\alpha}_t^{\dagger} + \nu_t \\ \boldsymbol{\alpha}_t^{\dagger} &= \boldsymbol{\Phi}_t\boldsymbol{\alpha}_{t-1}^{\dagger} + \boldsymbol{\xi}_t. \end{aligned} \tag{3.2}$$

The state transition equation retains its original form, but the new initial state $\boldsymbol{\alpha}_0^\dagger$ is now fixed at the value $\mathbf{0}$. Harvey (1989), page 139, refers to this as the *Generalized Least Squares* (GLS) transformation.

Partitioning $\mathbf{b}_t^{'} = \mathbf{h}_t^{'} \prod_{i=0}^{t-1} \boldsymbol{\Phi}_{t-i} = (\mathbf{b}_{1t}^{'} : \mathbf{b}_{2t}^{'})$, according to the partition of the initial state $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{10}^{'}, \boldsymbol{\alpha}_{20}^{'})^{'}$, we can write the model as

$$\begin{aligned} y_t &= \mathbf{x}_t^{*'} \boldsymbol{\beta}^* + \mathbf{z}_t^{*'} \boldsymbol{\gamma}^* + \mathbf{h}_t^{'} \boldsymbol{\alpha}_t^\dagger + \nu_t \\ \boldsymbol{\alpha}_t^\dagger &= \boldsymbol{\Phi}_t \boldsymbol{\alpha}_{t-1}^\dagger + \boldsymbol{\xi}_t, \end{aligned} \tag{3.3}$$

where $\mathbf{x}_t^{*'} = (\mathbf{x}_t^{'} : \mathbf{b}_{1t}^{'})$ is the $1 \times (p+q_1)$ design vector at time t for the $(p+q_1) \times 1$ vector of "fixed" effects $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^{'}, \boldsymbol{\alpha}_{10}^{'})^{'}$ and $\mathbf{z}_t^{*'} = (\mathbf{z}_t^{'} : \mathbf{b}_{2t}^{'})$ is the $1 \times (g+q_2)$ design vector at time t for the $(g+q_2) \times 1$ vector of time–invariant random effects $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}^{'}, \boldsymbol{\alpha}_{20}^{'})^{'}$. In vector form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{H}\boldsymbol{\alpha}^\dagger + \boldsymbol{\nu}, \tag{3.4}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{*'} \\ \mathbf{x}_2^{*'} \\ \vdots \\ \mathbf{x}_T^{*'} \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^{*'} \\ \mathbf{z}_2^{*'} \\ \vdots \\ \mathbf{z}_T^{*'} \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \mathbf{h}_1^{'} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{h}_2^{'} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{h}_T^{'} \end{bmatrix},$$

$$\boldsymbol{\alpha}^\dagger = \begin{bmatrix} \boldsymbol{\alpha}_1^\dagger \\ \boldsymbol{\alpha}_2^\dagger \\ \vdots \\ \boldsymbol{\alpha}_T^\dagger \end{bmatrix}, \quad \boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_T \end{bmatrix}.$$

The SSM is now expressed in terms of a linear mixed model. The "fixed" effect vector $\boldsymbol{\beta}^*$ is assigned a flat prior distribution. The vector of the time–invariant random effects $\boldsymbol{\gamma}^*$ has a $N(\mathbf{0}, \mathbf{B})$ prior distribution; the covariance matrix $\mathbf{B}$ consists of $\mathbf{B}_{11}$ and $\mathbf{B}_{22}$ on its diagonal, and covariance component $\mathbf{B}_{12}$. The $Tq \times 1$ vector of modified states $\boldsymbol{\alpha}^\dagger$ has mean $\mathbf{0}$ and covariance matrix $\mathbf{D}$, which can be obtained recursively from the state transition equation in (3.1). Let $\mathbf{D}_{s,t} = \mathrm{cov}(\boldsymbol{\alpha}_s^\dagger, \boldsymbol{\alpha}_t^\dagger)$ and $\mathbf{D}_{0,0} = \mathbf{0}$. Then

$$\mathbf{D}_{s,t} = \begin{cases} \boldsymbol{\Phi}_t \mathbf{D}_{t-1,t-1} \boldsymbol{\Phi}_t^{'} + \boldsymbol{\Xi}_t, & \text{if } s = t, \\ \mathbf{D}_{s,s} \prod_{i=s+1}^{t} \boldsymbol{\Phi}_i^{'}, & \text{if } s < t. \end{cases} \tag{3.5}$$

Furthermore, let

$$\boldsymbol{\Lambda} = \mathbf{H}\mathbf{D}\mathbf{H}^{'} + \sigma^2 \mathbf{I} \quad \text{and} \quad \mathbf{V} = \mathrm{var}(\mathbf{y}) = \mathbf{Z}\mathbf{B}\mathbf{Z}^{'} + \boldsymbol{\Lambda}. \tag{3.6}$$

Simple matrix algebra shows that

$$\mathbf{V}^{-1} = \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{Z}(\mathbf{Z}^{'}\mathbf{\Lambda}^{-1}\mathbf{Z} + \mathbf{B}^{-1})^{-1}\mathbf{Z}^{'}\mathbf{\Lambda}^{-1}$$
$$\text{and} \quad |\mathbf{V}| = |\mathbf{Z}\mathbf{B}\mathbf{Z}^{'} + \mathbf{\Lambda}| = (|\mathbf{\Lambda}|/|\mathbf{B}^{-1}|)|\mathbf{B}^{-1} + \mathbf{Z}^{'}\mathbf{\Lambda}^{-1}\mathbf{Z}|, \tag{3.7}$$

implying $\ln|\mathbf{V}| = \ln|\mathbf{\Lambda}| + \ln|\mathbf{B}| + \ln|\mathbf{B}^{-1} + \mathbf{Z}^{'}\mathbf{\Lambda}^{-1}\mathbf{Z}|$. These results are needed later when we evaluate the restricted maximum likelihood function.

3.2. The augmented Kalman filter

A brute-force application of the results in Section 2 to the analysis of the SSM in (3.4) would require the inversion of the large $T \times T$ covariance matrix $\mathbf{V}$. We circumvent this by using the Kalman filter (KF) to perform the Cholesky decomposition of $\mathbf{\Lambda} = \mathbf{H}\mathbf{D}\mathbf{H}^{'} + \sigma^2\mathbf{I}$. More specifically, $\mathbf{\Lambda}^{-1} = \mathbf{M}^{'}\mathbf{F}^{-1}\mathbf{M}$, where $\mathbf{M}$ is the KF operator (a lower triangular matrix with ones in the diagonal) that produces the innovations in the simple state space model $y_t = \mathbf{h}_t^{'}\boldsymbol{\alpha}_t^{\dagger} + \nu_t$. $\mathbf{F} = \text{diag}(f_1, \ldots, f_T)$ is a diagonal matrix which contains the variances of these innovations; it, too, is output of the KF recursions. The matrix $\mathbf{M}$ can be obtained by running the KF recursions on the T columns of the $T \times T$ identity matrix $\mathbf{I}$.

For computational efficiency we expand the prediction and updating recursions of the KF algorithm to matrix recursions; that is, we *augment* the Kalman Filter (see De Jong (1991) and Jones (1993) for related discussion). We apply the KF estimating and updating procedure that originates from the model $y_t = \mathbf{h}_t^{'}\boldsymbol{\alpha}_t^{\dagger} + \nu_t$; $\boldsymbol{\alpha}_t^{\dagger} = \mathbf{\Phi}_t\boldsymbol{\alpha}_{t-1}^{\dagger} + \boldsymbol{\xi}_t$; $\boldsymbol{\alpha}_0^{\dagger} = \mathbf{0}$, to each column of the $T \times (p + g + q + 1 + T + qT)$ matrix $[\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}]$. The predictions and updates that the KF produces at time t when applied to the columns of this matrix are collected in "state" prediction and "state" update matrices, denoted by $\mathbf{A}_{t/t-1}$ and $\mathbf{A}_{t/t}$ respectively. Here

$$\mathbf{A}_{t/s} = \Big[\boldsymbol{\alpha}_{t/s}^{x_{(1)}}, \ldots, \boldsymbol{\alpha}_{t/s}^{x_{(p+q_1)}} : \boldsymbol{\alpha}_{t/s}^{z_{(1)}}, \ldots, \boldsymbol{\alpha}_{t/s}^{z_{(g+q_2)}} : \boldsymbol{\alpha}_{t/s}^{(y)} : \boldsymbol{\alpha}_{t/s}^{i_1}, \ldots, \boldsymbol{\alpha}_{t/s}^{i_T} : \boldsymbol{\alpha}_{t/s}^{h_{(1)}}, \ldots, \boldsymbol{\alpha}_{t/s}^{h_{(qT)}}\Big]; s = t-1, t, \tag{3.8}$$

where $\mathbf{x}_{(j)}$ is the jth column of $\mathbf{X}$, $\mathbf{z}_{(j)}$ is the jth column of $\mathbf{Z}$, $\mathbf{h}_{(j)}$ is the jth column of $\mathbf{H}$ and $\mathbf{i}_j$ is the indicator vector for time j. The KF implies the following recursions:

$$\text{(a)}\ \mathbf{A}_{t/t-1} = \mathbf{\Phi}_t\mathbf{A}_{t-1/t-1}; \quad \text{(b)}\ \mathbf{P}_{t/t-1} = \mathbf{\Phi}_t\mathbf{P}_{t-1/t-1}\mathbf{\Phi}_t^{'} + \mathbf{\Xi}_t;$$
$$\text{(c)}\ \mathbf{A}_{t/t} = \mathbf{A}_{t/t-1} + \frac{1}{f_t}\mathbf{P}_{t/t-1}\mathbf{h}_t\mathbf{E}_t^{'}; \quad \text{(d)}\ \mathbf{P}_{t/t} = \mathbf{P}_{t/t-1} - \frac{1}{f_t}\mathbf{P}_{t/t-1}\mathbf{h}_t\mathbf{h}_t^{'}\mathbf{P}_{t/t-1};$$
$$\text{(e)}\ f_t = \mathbf{h}_t^{'}\mathbf{P}_{t/t-1}\mathbf{h}_t + \sigma^2; \ \text{(f)}\ \det_t = \det_{t-1} + \ln f_t, \tag{3.9}$$

where $\mathbf{E}_t^{'} = \left[\mathbf{x}_t^{*'} : \mathbf{z}_t^{*'} : y_t : \mathbf{i}_t^{'} : \mathbf{h}(t)^{'}\right] - \mathbf{h}_t^{'}\mathbf{A}_{t/t-1}$ is the $1\times(p+g+q+1+T+qT)$ row vector that collects the "innovations" at time t for the rows of $[\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}]$. The vector $\mathbf{h}(t)^{'}$ is the tth row of the matrix $\mathbf{H}$. Notice that only equations (a) and (c) in (3.9) are augmented. The initial conditions for the recursions in (3.9) are $\mathbf{A}_{0/0} = \mathbf{0}$, $\mathbf{P}_{0/0} = \mathbf{0}$ and $\det_0 = 0$. Recursion (f) is used to calculate $\ln|\mathbf{\Lambda}| = \det_T$. Simplifications in the recursions occur because of the special nature of the matrices $\mathbf{I}$ and $\mathbf{H}$ (that is, $\boldsymbol{\alpha}_{t/t-1}^{i_j} = \boldsymbol{\alpha}_{t/t}^{i_j} = \mathbf{0}$ for $j > t$ and $\boldsymbol{\alpha}_{t/t-1}^{h_{(j)}} = \boldsymbol{\alpha}_{t/t}^{h_{(j)}} = \mathbf{0}$ for $j > tq$), and similarly for the corresponding elements in $\mathbf{E}_t$. They result in considerable computational time savings in the algorithm.

The results of the augmented Kalman filter are used in the matrix recursion

$$\mathbf{Q}_t([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}]) = \mathbf{Q}_{t-1}([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}]) + \frac{1}{f_t}\mathbf{E}_t\mathbf{E}_t^{'}, \tag{3.10}$$

with starting value $\mathbf{Q}_0([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}]) = \mathrm{diag}(\mathbf{0}, \mathbf{B}^{-1}, 0, \mathbf{0}, \mathbf{0})$. At the end of the recursions we obtain the $(p + g + q + 1 + T + qT)$ crossproduct matrix

$$\mathbf{Q}_T([\mathbf{X}:\mathbf{Z}:\mathbf{y}:\mathbf{I}:\mathbf{H}]) = \begin{bmatrix} \mathbf{X}'\mathbf{\Lambda}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{\Lambda}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{\Lambda}^{-1}\mathbf{y} & \mathbf{X}'\mathbf{\Lambda}^{-1} & \mathbf{X}'\mathbf{\Lambda}^{-1}\mathbf{H} \\ \mathbf{Z}'\mathbf{\Lambda}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{\Lambda}^{-1}\mathbf{Z}+\mathbf{B}^{-1} & \mathbf{Z}'\mathbf{\Lambda}^{-1}\mathbf{y} & \mathbf{Z}'\mathbf{\Lambda}^{-1} & \mathbf{Z}'\mathbf{\Lambda}^{-1}\mathbf{H} \\ \mathbf{y}'\mathbf{\Lambda}^{-1}\mathbf{X} & \mathbf{y}'\mathbf{\Lambda}^{-1}\mathbf{Z} & \mathbf{y}'\mathbf{\Lambda}^{-1}\mathbf{y} & \mathbf{y}'\mathbf{\Lambda}^{-1} & \mathbf{y}'\mathbf{\Lambda}^{-1}\mathbf{H} \\ \mathbf{\Lambda}^{-1}\mathbf{X} & \mathbf{\Lambda}^{-1}\mathbf{Z} & \mathbf{\Lambda}^{-1}\mathbf{y} & \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{H} \\ \mathbf{H}'\mathbf{\Lambda}^{-1}\mathbf{X} & \mathbf{H}'\mathbf{\Lambda}^{-1}\mathbf{Z} & \mathbf{H}'\mathbf{\Lambda}^{-1}\mathbf{y} & \mathbf{H}'\mathbf{\Lambda}^{-1} & \mathbf{H}'\mathbf{\Lambda}^{-1}\mathbf{H} \end{bmatrix}. \tag{3.11}$$

The necessary quantities for inference in the SSM (that is, for likelihood evaluation, signal extraction and smoothing) are obtained by *sweeping* the matrix $\mathbf{Q}_T([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}])$ on its first $p + g + q$ rows, which correspond to time–invariant (fixed and random) effects. Sweeping a matrix $\mathbf{C}$ on its kth row results in a matrix $\mathbf{C}^*$ with elements: $c_{kk}^* = 1/c_{kk}$, $c_{ik}^* = -c_{ik}/c_{kk}$ for $i \neq k$, $c_{kj}^* = c_{kj}/c_{kk}$ for $j \neq k$, and $c_{ij}^* = c_{ij} - (c_{ik}c_{kj}/c_{kk})$ for $i \neq k$, $j \neq k$. The diagonal element c_{kk} is referred to as the pivot of the sweep. The sweep operator is reversible and commutative. (See Goodnight (1979) for a tutorial on the sweep operator.) Let us denote the resulting matrix by

$$\mathbf{Q}_T^*([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}]) = \mathrm{SWP}[1, .., p + g + q]\mathbf{Q}_T([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}]). \tag{3.12}$$

The successive sweep operations are easy to program. However, for the purpose of showing that the resulting matrix Q_T^* contains all needed quantities, we carry out the sweep in two stages. The first sweep is on the rows that correspond to the time–invariant random effects; that is, on the second "row" of matrices in

(3.11). This results in

$$\begin{bmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & -\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} & \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} & \mathbf{X}'\mathbf{V}^{-1} & \mathbf{X}'\mathbf{V}^{-1}\mathbf{H} \\ \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{B}-\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} & \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y} & \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1} & \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{H} \\ \mathbf{y}'\mathbf{V}^{-1}\mathbf{X} & -\mathbf{y}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} & \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} & \mathbf{y}'\mathbf{V}^{-1} & \mathbf{y}'\mathbf{V}^{-1}\mathbf{H} \\ \mathbf{V}^{-1}\mathbf{X} & -\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} & \mathbf{V}^{-1}\mathbf{y} & \mathbf{V}^{-1} & \mathbf{V}^{-1}\mathbf{H} \\ \mathbf{H}'\mathbf{V}^{-1}\mathbf{X} & -\mathbf{H}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} & \mathbf{H}'\mathbf{V}^{-1}\mathbf{y} & \mathbf{H}'\mathbf{V}^{-1} & \mathbf{H}'\mathbf{V}^{-1}\mathbf{H} \end{bmatrix}. \tag{3.13}$$

The matrices in the second "row" and "column" are obtained after using the identities

$$\begin{aligned} (\mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{Z}+\mathbf{B}^{-1})^{-1} &= \mathbf{B}-\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} \\ (\mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{Z}+\mathbf{B}^{-1})^{-1}\mathbf{Z}'\boldsymbol{\Lambda}^{-1} &= (\mathbf{B}-\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B})\mathbf{Z}'\boldsymbol{\Lambda}^{-1} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}. \end{aligned} \tag{3.14}$$

A by-product of this first sweep is $|\mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{Z}+\mathbf{B}^{-1}| = \prod_{i=p+q_1+1}^{p+q_1+g+q_2} \mathrm{pv}_i$, where pv_i denotes the ith pivot of the sweep. Next we sweep the matrix in (3.13) on the rows that correspond to the "fixed" effects; that is the first "row" of matrices in (3.13), and obtain

$$\mathbf{Q}_T^*([\mathbf{X}:\mathbf{Z}:\mathbf{y}:\mathbf{I}:\mathbf{H}]) = \begin{bmatrix} (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} & -\mathbf{U}\mathbf{Z}\mathbf{B} & \hat{\boldsymbol{\beta}}^* & \mathbf{U} & \mathbf{U}\mathbf{H} \\ -\mathbf{B}\mathbf{Z}'\mathbf{U}' & \mathbf{B}-\mathbf{B}\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{B} & \hat{\gamma}^* & \mathbf{B}\mathbf{Z}'\mathbf{W} & \mathbf{B}\mathbf{Z}'\mathbf{W}\mathbf{H} \\ -\hat{\boldsymbol{\beta}}^{*\prime} & \hat{\gamma}^{*\prime} & \mathbf{y}'\mathbf{W}\mathbf{y} & \mathbf{y}'\mathbf{W} & \mathbf{y}'\mathbf{W}\mathbf{H} \\ -\mathbf{U}' & -\mathbf{W}\mathbf{Z}\mathbf{B} & \mathbf{W}\mathbf{y} & \mathbf{W} & \mathbf{W}\mathbf{H} \\ -\mathbf{H}'\mathbf{U}' & -\mathbf{H}'\mathbf{W}\mathbf{Z}\mathbf{B} & \mathbf{H}'\mathbf{W}\mathbf{y} & \mathbf{H}'\mathbf{W} & \mathbf{H}'\mathbf{W}\mathbf{H} \end{bmatrix} \tag{3.15}$$

where $\mathbf{W} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$ and $\mathbf{U} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$.

3.3. Inference in the SSM

The entries in the matrix (3.15) hold the key to all inference in the state space model. Selected submatrices of $\mathbf{Q}_T^*$, together with results from Section 2, lead to the following results:

(1) *Computation of the restricted loglikelihood* follows immediately from equation (2.3) and is achieved by using the elements in $\mathbf{Q}_T^*([\mathbf{X}:\mathbf{Z}:\mathbf{y}]) = \mathrm{SWP}[1,..,p+g+q]\mathbf{Q}_T([\mathbf{X}:\mathbf{Z}:\mathbf{y}])$, the identity $|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| = \prod_{i=1}^{p+q_1} \mathrm{pv}_i$ (a by–product of the second sweep), the result in (3.7), and the recursion (f) in (3.9). The restricted loglikelihood is calculated as

$$l_R = \mathrm{const} - \frac{1}{2}\mathrm{det}_T - \frac{1}{2}\sum_{i=1}^{p+g+q} \ln \mathrm{pv}_i - \frac{1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathbf{y}'\mathbf{W}\mathbf{y} \tag{3.16}$$

(2) *Estimation of effects and their mean square errors* follows from equations (2.6) and (2.7). The matrix $\mathbf{Q}_T^*([\mathbf{X}:\mathbf{Z}:\mathbf{y}])$ contains the necessary quantities

needed to obtain the BLUE of the fixed effects β^* and the BLUP of the time-invariant effects γ^* in model (3.4), together with their mean square errors: $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, $\hat{\gamma}^* = \mathbf{B}\mathbf{Z}'\mathbf{W}\mathbf{y}$, $\mathrm{MSE}(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, $\mathrm{MSE}(\hat{\gamma}^*) = \mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{B}$.

(3) *Signal extraction and noise estimation* follows from the mixed-model results in (2.8) and (2.9) and setting $\mathbf{R} = \sigma^2\mathbf{I}$. The needed quantities are given in $\mathbf{Q}_T^*([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I}]) = \mathrm{SWP}[1, \ldots, p+g+q]\mathbf{Q}_T(\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I})$. The signal and noise vector estimates and their MSEs are given by

$$\begin{aligned} \hat{\mathbf{s}} &= \mathbf{y} - \sigma^2\mathbf{W}\mathbf{y}; \mathrm{MSE}(\hat{\mathbf{s}}) = \sigma^2\mathbf{I} - \sigma^4\mathbf{W} \\ \hat{\boldsymbol{\nu}} &= \sigma^2\mathbf{W}\mathbf{y}; \mathrm{MSE}(\hat{\boldsymbol{\nu}}) = \sigma^2\mathbf{I} - \sigma^4\mathbf{W}. \end{aligned} \tag{3.17}$$

Note that signal extraction is done directly, without first having to obtain the smoothed states.

(4) *State smoothing* follows from (2.7). The BLUP of the vector of modified states and its MSE are given by

$$\hat{\boldsymbol{\alpha}}^\dagger = \mathbf{D}\mathbf{H}'\mathbf{W}\mathbf{y} \quad \text{and} \quad \mathrm{MSE}(\hat{\boldsymbol{\alpha}}^\dagger) = \mathbf{D} - \mathbf{D}\mathbf{H}'\mathbf{W}\mathbf{H}\mathbf{D}. \tag{3.18}$$

The required matrices $\mathbf{H}'\mathbf{W}\mathbf{y}$ and $\mathbf{H}'\mathbf{W}\mathbf{H}$ are part of $\mathbf{Q}_T^*([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{H}])$ in (3.15), and the matrix $\mathbf{D}$ is defined in (3.5). An estimate of the unmodified state $\boldsymbol{\alpha}_t$ is easily obtained, as it is expressed as a linear combination of $\boldsymbol{\alpha}_t^\dagger$ and the initial condition $\boldsymbol{\alpha}_0$.

Comments

Our algorithm provides a smoother that does *not* require backward recursions. This is different from traditional smoothing approaches which require both forward and backward recursions (see De Jong (1991)). Furthermore, our algorithm provides the complete MSE matrices of the vector of smoothed states and of the vector of signals (that is, variances as well as covariances of estimated states and signals at different times). The only drawback of our approach is the large storage requirement that depends on both the number of time points T and the dimension of the state vector q. For example, in the local linear trend model discussed in Section 4.2 the state vector is of dimension 2, and with 100 observations $\mathbf{Q}_T^*([\mathbf{X} : \mathbf{y} : \mathbf{I} : \mathbf{H}])$ is a 303×303 matrix. However, considerable savings are obtained during the matrix recursions by utilizing the simplifications that arise from the special structure of $\mathbf{I}$ and $\mathbf{H}$. Also, if the goal is likelihood evaluation and signal extraction (which implies that $\mathbf{H}$ does not have to be included in the augmented recursions), the required $\mathbf{Q}_T^*([\mathbf{X} : \mathbf{y} : \mathbf{I}])$ is a 103×103 matrix. While the number of pivots in the sweep is small (as $p + g + q = 2$

is small), the number of elements that are changed during each sweep is of the order T^2. However, if one's interest is in MSE($\hat{\mathbf{s}}_t$) only and not in the covariances among estimated signals, then the calculations simplify considerably, as it is not necessary to sweep the off–diagonal elements of $\mathbf{\Lambda}^{-1}$. Finally, if all one needs is the evaluation of the likelihood, $\mathbf{I}$ can be left out as well and $\mathbf{Q}_T^*([\mathbf{X} : \mathbf{y}])$ is a 3×3 matrix. This discussion shows that while the dimension of the matrix may be large, computational savings can be realized by processing only the relevant components.

3.4. The disturbance smoother for direct estimation of disturbance terms

Consider the state space model in (3.3), but now express this model in terms of the disturbances $\boldsymbol{\xi}$. That is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{K}\boldsymbol{\xi} + \boldsymbol{\nu}, \tag{3.19}$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{h}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{h}_2'\mathbf{\Phi}_2 & \mathbf{h}_2' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_T' \prod_{i=0}^{T-2} \mathbf{\Phi}_{T-i} & \mathbf{h}_T' \prod_{i=0}^{T-3} \mathbf{\Phi}_{T-i} & \cdots & \mathbf{h}_T' \end{bmatrix} \quad \text{and} \quad \boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_T \end{bmatrix}. \tag{3.20}$$

The matrix $\mathbf{K}$ is of dimension $T \times qT$. The disturbance vector $\boldsymbol{\xi}$ has mean $\mathbf{0}$ and covariance matrix $\tilde{\mathbf{\Xi}} = \text{diag}(\mathbf{\Xi}_1, \ldots, \mathbf{\Xi}_T)$. Equation (2.7) provides the BLUP of the vector $\boldsymbol{\xi}$ and its associated MSE:

$$\hat{\boldsymbol{\xi}} = \tilde{\mathbf{\Xi}}\mathbf{K}'\mathbf{W}\mathbf{y}; \text{MSE}(\hat{\boldsymbol{\xi}}) = \tilde{\mathbf{\Xi}} - \tilde{\mathbf{\Xi}}\mathbf{K}'\mathbf{W}\mathbf{K}\tilde{\mathbf{\Xi}}. \tag{3.21}$$

In particular,

$$\text{MSE}(\hat{\boldsymbol{\xi}}_t) = \mathbf{\Xi}_t - \mathbf{\Xi}_t\mathbf{K}_t\mathbf{W}\mathbf{K}_t'\mathbf{\Xi}_t, \tag{3.22}$$

where $\mathbf{K}_t = [\mathbf{0}, \ldots, \mathbf{0}, \mathbf{h}_t, \mathbf{\Phi}_{t+1}'\mathbf{h}_{t+1}, (\mathbf{\Phi}_{t+1}'\mathbf{\Phi}_{t+2}')\mathbf{h}_{t+2}, \ldots, (\mathbf{\Phi}_{t+1}' \ldots \mathbf{\Phi}_T')\mathbf{h}_T]$ is a $q \times T$ matrix. The expressions that are needed for these calculations can be obtained by replacing the matrix $\mathbf{H}$ in the augmented KF recursions in Section 3.2 by the matrix $\mathbf{K}$. The resulting matrix

$$\mathbf{Q}_T^*([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{K}]) = \text{SWP}[1, .., p + g + q]\mathbf{Q}_T([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{K}]) \tag{3.23}$$

contains the needed quantities $\mathbf{K}'\mathbf{W}\mathbf{y}$ and $\mathbf{K}'\mathbf{W}\mathbf{K}$. The disturbance smoother is computationally simpler than the state smoother in the previous section, due to the fact that $\tilde{\mathbf{\Xi}}$ is a block diagonal matrix. Furthermore, if we are not interested

in the covariances between disturbances, the sweep in (3.23) needs to be carried out only on the Tq^2 elements of the diagonal blocks in $\mathbf{K}'\mathbf{\Lambda}^{-1}\mathbf{K}$.

3.5. Forecasting and missing observations

Missing observations can be treated as parameters with a diffuse prior distribution. Suppose model (3.19) has $T-T_0$ missing observations; without loss of generality, assume that the missing observations are the last $T-T_0$ observations. Missing values are incorporated as

$$\mathbf{y}^* = \begin{bmatrix} \mathbf{y}_{\text{obs}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{\text{obs}} & \mathbf{0} \\ \mathbf{X}_{\text{mis}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^* \\ -\mathbf{y}_{\text{mis}} \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_{\text{obs}} \\ \mathbf{Z}_{\text{mis}} \end{bmatrix} \boldsymbol{\gamma}^* + \mathbf{K}\boldsymbol{\xi} + \boldsymbol{\nu}, \tag{3.24}$$

where $\mathbf{y}_{\text{obs}}$ is the $T_0 \times 1$ vector of observed responses, $\mathbf{0}$ is a $(T-T_0)\times 1$ vector of zeroes, $\mathbf{X}_{\text{obs}}$ and $\mathbf{Z}_{\text{obs}}$ are the design matrices that correspond to the observed part of the data, $\mathbf{X}_{\text{mis}}$ and $\mathbf{Z}_{\text{mis}}$ are the $(T-T_0)\times(p+q_1)$ and $(T-T_0)\times(g+q_2)$ design matrices that correspond to the missing data, and $\mathbf{I}$ is the $(T-T_0)\times(T-T_0)$ identity matrix. The above model can be expressed as

$$\mathbf{y}^* = \left[\mathbf{X}\middle|\mathbf{J}\right] \begin{bmatrix} \boldsymbol{\beta} \\ -\mathbf{y}_{\text{mis}} \end{bmatrix} + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{K}\boldsymbol{\xi} + \boldsymbol{\nu}. \tag{3.25}$$

We impute zeroes for the missing observations, augment the design matrix by a $T\times(T-T_0)$ matrix of indicators for the missing values, and augment the vector of "fixed" effects by parameters that represent the negative of the missing values.

The inference discussed in the previous sections applies directly. In fact, the matrix $\mathbf{J}$ in (3.25) is processed through the KF as part of the $\mathbf{I}_{T\times T}$ matrix. This implies that we run the augmented KF on the same matrices as before, except that $\mathbf{y}$ is replaced by $\mathbf{y}^*$. We then further sweep the matrix in (3.23) on the rows that correspond to the missing observations in the fourth "row" of matrices. Forecasts can be viewed as predictions of "missing" future observations. Hence the same method can be used for forecasting a vector of future observations and its MSE matrix.

4. Algorithms for REML Estimates in State Space Models

4.1 EM algorithm in the general state space model

In Section 3.3 we showed how to calculate the restricted loglikelihood function from the output of the augmented Kalman filter and the subsequent sweep operation. Recursive numerical optimization approaches can then be employed to calculate REML estimates.

In this section we develop an EM algorithm for obtaining the REML estimates of the variance components in the state space model. The EM algorithm

has the advantages of never decreasing the loglikelihood function during the course of the iterations, and of converging to feasible (non-negative) variance estimates. The disturbance smoother of Section 3.4 plays a central role in the EM computations.

We formulate the EM algorithm for the state space model in terms of its general mixed model representation in (3.19): $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{K}\boldsymbol{\xi} + \boldsymbol{\nu}$, with independent random vectors $\boldsymbol{\nu} \sim N(\mathbf{0}, \mathbf{R} = \sigma^2\mathbf{I}_T)$ and $\boldsymbol{\xi} \sim N(\mathbf{0}, \tilde{\boldsymbol{\Xi}} = \mathbf{I}_T \bigotimes \boldsymbol{\Xi})$. We assume that the transition matrices are known, and we let $\boldsymbol{\theta}$ be the vector of parameters, not including the p-dimensional vector of fixed effects.

Assigning a flat prior to the fixed effects $\boldsymbol{\beta}^*$ and integrating over $\boldsymbol{\beta}^*$ leads to the *complete data marginal likelihood* $L(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}^*, \boldsymbol{\xi})$, which is equal to

$$\text{const} \times \frac{1}{|\tilde{\boldsymbol{\Xi}}|^{\frac{1}{2}}|\mathbf{B}|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{\xi}'\tilde{\boldsymbol{\Xi}}^{-1}\boldsymbol{\xi} - \frac{1}{2}\boldsymbol{\gamma}^{*\prime}\mathbf{B}^{-1}\boldsymbol{\gamma}^*} \int \frac{1}{|\mathbf{R}|^{\frac{1}{2}}} e^{-\frac{1}{2}[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*-\mathbf{Z}\boldsymbol{\gamma}^*-\mathbf{K}\boldsymbol{\xi}]'\mathbf{R}^{-1}[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*-\mathbf{Z}\boldsymbol{\gamma}^*-\mathbf{K}\boldsymbol{\xi}]} d\boldsymbol{\beta}^*. \tag{4.1}$$

Let $\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the "hat" matrix in linear regression. By taking the logarithm of the above equation and substituting $\mathbf{R} = \sigma^2\mathbf{I}$, we obtain the *complete data restricted loglikelihood* (up to a constant)

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}^*, \boldsymbol{\xi}) = &-\frac{T-p}{2}\ln\sigma^2 - \frac{1}{2}\ln|\mathbf{X}'\mathbf{X}| - \frac{1}{2\sigma^2}[\mathbf{y}-\mathbf{Z}\boldsymbol{\gamma}^*-\mathbf{K}\boldsymbol{\xi}]'\mathbf{G}[\mathbf{y}-\mathbf{Z}\boldsymbol{\gamma}^*-\mathbf{K}\boldsymbol{\xi}] \\ &- \frac{1}{2}\ln|\tilde{\boldsymbol{\Xi}}| - \frac{1}{2}\boldsymbol{\xi}'\tilde{\boldsymbol{\Xi}}^{-1}\boldsymbol{\xi} - \frac{1}{2}\ln|\mathbf{B}| - \frac{1}{2}\boldsymbol{\gamma}^{*\prime}\mathbf{B}^{-1}\boldsymbol{\gamma}^*. \end{aligned} \tag{4.2}$$

The $(r+1)$th iteration of the EM algorithm involves the following two steps. In the *expectation step* we obtain $E[l(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}^*, \boldsymbol{\xi})|\mathbf{y}, \boldsymbol{\theta}(r)]$, where $\boldsymbol{\theta}(r)$ is the rth iterate of the parameter vector. In the *maximization step* we maximize this conditional expectation with respect to $\boldsymbol{\theta}$; the maximizer is the $(r+1)$th iterate $\boldsymbol{\theta}(r+1)$. These steps are repeated until a suitable convergence criterion is satisfied.

The E–step of the EM algorithm results in

$$\begin{aligned} &E[l(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}^*, \boldsymbol{\xi})|\mathbf{y}, \boldsymbol{\theta}(r)] \\ =&-\frac{T-p}{2}\ln\sigma^2 - \frac{1}{2}\ln|\mathbf{X}'\mathbf{X}| - \frac{1}{2\sigma^2}E\{[\mathbf{y}-\mathbf{Z}\boldsymbol{\gamma}^*-\mathbf{K}\boldsymbol{\xi}]'\mathbf{G}[\mathbf{y}-\mathbf{Z}\boldsymbol{\gamma}^*-\mathbf{K}\boldsymbol{\xi}]|\mathbf{y}, \boldsymbol{\theta}(r)\} \\ &-\frac{1}{2}\ln|\tilde{\boldsymbol{\Xi}}| - \frac{1}{2}E\{\boldsymbol{\xi}'\tilde{\boldsymbol{\Xi}}^{-1}\boldsymbol{\xi}|\mathbf{y}, \boldsymbol{\theta}(r)\} - \frac{1}{2}\ln|\mathbf{B}| - \frac{1}{2}E\{\boldsymbol{\gamma}^{*\prime}\mathbf{B}^{-1}\boldsymbol{\gamma}^*|\mathbf{y}, \boldsymbol{\theta}(r)\}. \end{aligned} \tag{4.3}$$

The $(r+1)$th iterate for the noise variance becomes

$$\sigma^2(r+1) = \frac{[\hat{\boldsymbol{\nu}}(r)]'\mathbf{G}[\hat{\boldsymbol{\nu}}(r)] + \text{tr}[\mathbf{G}\text{MSE}(\hat{\boldsymbol{\nu}}(r))]}{T-p}. \tag{4.4}$$

The quantities $\hat{\boldsymbol{\nu}}(r) = \mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\gamma}}^*(r) - \mathbf{K}\hat{\boldsymbol{\xi}}(r) = \sigma^2(r)\mathbf{W}(r)\mathbf{y}$ and $\mathrm{MSE}(\hat{\boldsymbol{\nu}}(r)) = \sigma^2(r)\mathbf{I} - \sigma^4(r)\mathbf{W}(r)$, where $\mathbf{W}(r) = \mathbf{V}^{-1}(r) - \mathbf{V}^{-1}(r)\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}(r)\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(r)$ and $\mathbf{V}(r) = \mathbf{K}\tilde{\boldsymbol{\Xi}}(r)\mathbf{K}' + \mathbf{Z}\mathbf{B}\mathbf{Z}' + \sigma^2(r)\mathbf{I}$, can be obtained from the disturbance smoother in Section 3.4. Simple algebra shows that

$$[\hat{\boldsymbol{\nu}}(r)]'\mathbf{G}[\hat{\boldsymbol{\nu}}(r)] = \sigma^4(r)\mathbf{y}'\mathbf{W}(r)\mathbf{W}(r)\mathbf{y} \quad \text{and}$$
$$\mathrm{tr}[\mathbf{G}\mathrm{MSE}(\hat{\boldsymbol{\nu}}(r))] = (T-p)\sigma^2(r) - \sigma^4(r)\mathrm{tr}\mathbf{W}(r)$$

and hence

$$\sigma^2(r+1) = \sigma^2(r) + \sigma^4(r)\frac{[(\mathbf{W}(r)\mathbf{y})'(\mathbf{W}(r)\mathbf{y}) - \mathrm{tr}\,\mathbf{W}(r)]}{T-p}. \tag{4.5}$$

The result $E\{\boldsymbol{\xi}'\tilde{\boldsymbol{\Xi}}^{-1}\boldsymbol{\xi}|\mathbf{y}, \boldsymbol{\theta}(r)\} = \sum_{t=1}^{T}\hat{\boldsymbol{\xi}}_t'(r)\ \boldsymbol{\Xi}^{-1}\ \hat{\boldsymbol{\xi}}_t(r) + T\mathrm{tr}(\boldsymbol{\Xi}^{-1}\boldsymbol{\Xi}(r)) - \sum_{t=1}^{T}\mathrm{tr}(\boldsymbol{\Xi}^{-1}\boldsymbol{\Xi}(r)\mathbf{K}_t\mathbf{W}(r)\mathbf{K}_t'\boldsymbol{\Xi}(r))$, which follows from properties of the normal distribution and the mean square error result in (3.22), is used in the derivation of the EM iterates for the covariance matrix $\boldsymbol{\Xi}$. Taking the derivative with respect to $\boldsymbol{\Xi}$ of the expectation of the restricted complete loglikelihood in (4.3) conditional on the observed data, evaluating the derivative at $\boldsymbol{\theta}(r)$, and setting it to $\mathbf{0}$, results in

$$-\frac{T}{2}\boldsymbol{\Xi}^{-1} + \frac{1}{2}\boldsymbol{\Xi}^{-1}\left(\sum_{t=1}^{T}\hat{\boldsymbol{\xi}}_t(r)\hat{\boldsymbol{\xi}}_t'(r)\right)\boldsymbol{\Xi}^{-1} + \frac{T}{2}\boldsymbol{\Xi}^{-1}\boldsymbol{\Xi}(r)\boldsymbol{\Xi}^{-1}$$
$$-\frac{1}{2}\boldsymbol{\Xi}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{\Xi}(r)\mathbf{K}_t\mathbf{W}(r)\mathbf{K}_t'\boldsymbol{\Xi}(r)\Big)\boldsymbol{\Xi}^{-1} = \mathbf{0}. \tag{4.6}$$

This leads to the updating equation

$$\boldsymbol{\Xi}(r+1) = \boldsymbol{\Xi}(r) + \frac{\sum_{t=1}^{T}\hat{\boldsymbol{\xi}}_t(r)\hat{\boldsymbol{\xi}}_t'(r) - \boldsymbol{\Xi}(r)(\sum_{t=1}^{T}\mathbf{K}_t\mathbf{W}(r)\mathbf{K}_t')\boldsymbol{\Xi}(r)}{T}. \tag{4.7}$$

All quantities needed for computation are stored in the matrix $\mathbf{Q}_T^*([\mathbf{X} : \mathbf{Z} : \mathbf{y} : \mathbf{I} : \mathbf{K}])$ (see Section 3.4).

4.2. Estimation of variances in structural time series models

The EM iterations in Section 4.1 assume an arbitrary covariance matrix for the disturbances $\boldsymbol{\xi}_t$. In structural time series model this covariance matrix is diagonal, and the discussion in this section shows how to handle this situation. Structural time series models represent a flexible class for modeling and forecasting time series observations, and these models are described in detail by Harvey (1989). The structural time series model with local polynomial trend of order g

and indicator variable seasonality of order s can be written in state space form as follows. The observation equation is $y_t = \mathbf{h}'\boldsymbol{\alpha}_t + \nu_t$, where the vector $\mathbf{h}$ is a $(g+1)+(s-1) = g+s$ vector of zeros, except for ones in rows 1 and $g+2$ and $\boldsymbol{\alpha}_t = [\mu_t^1, \mu_t^2, \ldots, \mu_t^{g+1}, \tau_t, \tau_{t-1}, \ldots, \tau_{t-s+2}]'$ is a state vector of dimension $(g+1)+(s-1)$; the first $g+1$ components represent the trend coefficients, while the remaining $(s-1)$ components correspond to seasonal indicators. The state transition equation is given by $\boldsymbol{\alpha}_t = \boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\xi}_t$, where the $(g+s)\times(g+s)$ transition matrix is

$$\boldsymbol{\Phi} = \left[\begin{array}{c|c} \boldsymbol{\Psi} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Omega} \end{array}\right]. \tag{4.8}$$

The nonzero elements of $\boldsymbol{\Psi}$ are given by $\Psi_{ij} = 1/(j-i)!$, for $j = i, \ldots, g+1$; $i = 1, \ldots, g+1$. The nonzero elements of $\boldsymbol{\Omega}$ are given by $\Omega_{ij} = -1$, for $i = 1$; $j = 1, \ldots, s-1$ and $\Omega_{i,i-1} = 1$ for $i = 2, \ldots, s-1$. The $g+2$ nonzero random components in the disturbance vector $\boldsymbol{\xi}_t = \left[\xi_t^1, \xi_t^2, \ldots, \xi_t^{g+1}, \xi_t^{g+2}, 0, \ldots, 0\right]$ are mutually independent with zero means and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{g+1}^2, \sigma_{g+2}^2$; in addition, they are independent across time. (See Harvey (1989) and Abraham and Ledolter (1986) for further discussion.)

The variance components $\sigma^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_{g+1}^2, \sigma_{g+2}^2$ need to be estimated from the data. Without prior information for the initial values of the trend and seasonal coefficients, we assign a diffuse prior to $\boldsymbol{\alpha}_0 = \left[\mu_0^1, \mu_0^2, \ldots, \mu_0^g, \mu_0^{g+1}; \tau_0, \ldots, \tau_{-s+2}\right]'$. We incorporate the initial conditions into the vector of fixed effects ($\boldsymbol{\beta} = \boldsymbol{\alpha}_0$) and write the structural time series model in the following state space form:

$$\begin{aligned} y_t &= \mathbf{h}'\boldsymbol{\Phi}^t\boldsymbol{\alpha}_0 + \mathbf{h}'\boldsymbol{\alpha}_t^\dagger + \nu_t = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{h}'\boldsymbol{\alpha}_t^\dagger + \nu_t \\ \boldsymbol{\alpha}_t^\dagger &= \boldsymbol{\Phi}\boldsymbol{\alpha}_t^\dagger + \boldsymbol{\xi}_t, \end{aligned} \tag{4.9}$$

with the modified initial condition $\boldsymbol{\alpha}_t^\dagger$ fixed at $\mathbf{0}$. Equation (4.8) implies $\boldsymbol{\Phi}^t = \text{diag}(\boldsymbol{\Psi}^t, \boldsymbol{\Omega}^t)$. After some tedious algebra we can express the model in (4.9) as a variance components model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{L}_1\boldsymbol{\xi}^1 + \mathbf{L}_2\boldsymbol{\xi}^2 + \cdots + \mathbf{L}_{g+1}\boldsymbol{\xi}^{g+1} + \mathbf{L}_{g+2}\boldsymbol{\xi}^{g+2} + \boldsymbol{\nu}. \tag{4.10}$$

The rows of the $T\times(g+s)$ matrix $\mathbf{X}$ are given by $\mathbf{x}_t' = \left[1, t, \frac{t^2}{2!}, \ldots, \frac{t^g}{g!}; \mathbf{l}_t'\right]$, where $\mathbf{l}_t'$ is a row vector of length $s-1$, with $\mathbf{l}_t' = (-1, -1, \ldots, -1, -1)$ for $t = ks+1$; $\mathbf{l}_t' = (0, 0, \ldots, 0, 1)$ for $t = ks+2$; $\mathbf{l}_t' = (0, 0, \ldots, 1, 0)$ for $t = ks+3$;...; $\mathbf{l}_t' = (1, 0, \ldots, 0, 0)$ for $t = ks+s$. The matrices $\mathbf{L}_i$ are lower triangular and are obtained by expressing the state vector in terms of the disturbance vectors $\boldsymbol{\xi}^i = (\xi_1^i, \xi_2^i, \ldots, \xi_T^i)'$, $i = 1, \ldots, g+2$; this is achieved by repeated substitution

in the state transition equation (4.9). The lower triangular parts of the matrices $\mathbf{L}_{j+1}$, $j = 0, \ldots, g$, are given by

$$\mathbf{L}_{j+1}(t,k) = \frac{(t-k)^j}{j!} \; ; \; t = 1, \ldots, T \; ; \; k = 1, \ldots, t. \tag{4.11}$$

The nonzero elements of the lower triangular incidence matrix for the seasonal disturbances $\mathbf{L}_{g+2}$ are

$$\begin{aligned} &\mathbf{L}_{g+2}(t, t-ks) = +1 \; ; \; k = 0, \ldots, \Big[\frac{t-1}{s}\Big] \\ \text{and} \quad &\mathbf{L}_{g+2}(t, t-ks-1) = -1; k = 0, \ldots, \Big[\frac{t-2}{s}\Big]. \end{aligned} \tag{4.12}$$

The model in (4.10) is expressed as a variance components model of the form (2.2) with $m = g+2$, $\mathbf{Z}_i = \mathbf{L}_i$, $\boldsymbol{\pi}_i = \boldsymbol{\xi}^i$, and $\mathbf{e} = \boldsymbol{\nu}$. Hence we can apply the result in (2.11) to obtain the EM iterates for the REML estimates of the variance components:

$$\begin{aligned} \sigma^2(r+1) &= \sigma^2(r) + \sigma^4(r)\frac{[\mathbf{W}(r)\mathbf{y}]'[\mathbf{W}(r)\mathbf{y}] - \mathrm{tr}[\mathbf{W}(r)]}{T-g-s} \\ \sigma_i^2(r+1) &= \sigma_i^2(r) + \sigma_i^4(r)\frac{[\mathbf{L}_i'\mathbf{W}(r)\mathbf{y}]'[\mathbf{L}_i'\mathbf{W}(r)\mathbf{y}] - \mathrm{tr}[\mathbf{L}_i'\mathbf{W}(r)\mathbf{L}_i]}{T}; \\ &\qquad\qquad i = 1, \ldots, g+2, \end{aligned} \tag{4.13}$$

where $\mathbf{W}(r) = \mathbf{V}^{-1}(r) – \mathbf{V}^{-1}(r)\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}(r)\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(r)$ with $\mathbf{V}(r) = \sigma^2(r)\mathbf{I} + \sum_{i=1}^{g+2}\sigma_i^2(r)\mathbf{L}_i\mathbf{L}_i'$.

The EM algorithm requires the calculation of the matrices $\mathbf{W}\mathbf{y}$, $\mathbf{L}_i'\mathbf{W}\mathbf{y}$ and the diagonal elements of $\mathbf{L}_i'\mathbf{W}\mathbf{L}_i$ $(i = 1, \ldots, g+2)$. These quantities are output of the augmented KF and sweep algorithm in Section 3.2. The KF recursions are applied to the columns of the matrix $[\mathbf{X} : \mathbf{y} : \mathbf{I} : \mathbf{L}_1 : \mathbf{L}_2 : \ldots : \mathbf{L}_{g+2}]$ and the resulting cross–product matrix is swept on its first $g+s$ rows. The result is stored in $\mathbf{Q}_T^*(\mathbf{X} : \mathbf{y} : \mathbf{I} : \mathbf{L}_1 : \mathbf{L}_2 : \ldots : \mathbf{L}_{g+2})$. The fact that the $\mathbf{L}_i$ matrices are lower triangular simplifies the computational burden.

The first and second derivatives of the restricted loglikelihood are given by

$$\begin{aligned} &\text{(i)} \quad \frac{dl_R}{d\sigma_i^2} = -\frac{1}{2}\mathrm{tr}(\mathbf{L}_i'\mathbf{W}\mathbf{L}_i) + \frac{1}{2}\mathbf{y}'\mathbf{W}\mathbf{L}_i\mathbf{L}_i'\mathbf{W}\mathbf{y}, \\ &\text{(ii)} \quad \frac{d^2l_R}{d\sigma_i^2 d\sigma_j^2} = \frac{1}{2}\mathrm{ssq}(\mathbf{L}_i'\mathbf{W}\mathbf{L}_j) - \mathbf{y}'\mathbf{W}\mathbf{L}_j\mathbf{L}_j'\mathbf{W}\mathbf{L}_i\mathbf{L}_i'\mathbf{W}\mathbf{y}, \\ &\text{(iii)} \quad -E\Big(\frac{d^2l_R}{d\sigma_i^2 d\sigma_j^2}\Big) = \frac{1}{2}\mathrm{ssq}(\mathbf{L}_i'\mathbf{W}\mathbf{L}_j), \end{aligned} \tag{4.14}$$

where $i = 0, \ldots, g+2$, $j = 0, \ldots, g+2$. In the above expression $\mathbf{L}_0 = \mathbf{I}$, $\sigma_0^2 = \sigma^2$ and ssq($\mathbf{A}$) denotes the sum of squares of the elements of the matrix $\mathbf{A}$ (see Searle at al. (1992), page 252). Expression (iii) in (4.14) provides the ijth element of the *information matrix* $\mathbf{I}(\boldsymbol{\theta})$ and expression (ii) gives the ijth element of the Hessian $\mathbf{H}(\boldsymbol{\theta})$. These matrices, together with the vector of first derivatives whose elements are given in (i) above, define the Newton–Raphson and the scoring algorithm for obtaining REML estimates:

$$\begin{aligned} \text{Newton–Raphson} : \boldsymbol{\theta}(r+1) &= \boldsymbol{\theta}(r) - \mathbf{H}^{-1}(\boldsymbol{\theta}(r))\frac{dl_R}{d\boldsymbol{\theta}}|_{\boldsymbol{\theta}(r)} \\ \text{Scoring} : \boldsymbol{\theta}(r+1) &= \boldsymbol{\theta}(r) + \mathbf{I}^{-1}(\boldsymbol{\theta}(r))\frac{dl_R}{d\boldsymbol{\theta}}|_{\boldsymbol{\theta}(r)}. \end{aligned} \tag{4.15}$$

The *asymptotic variance covariance matrix* of the REML estimates $\hat{\boldsymbol{\theta}} = [\hat{\sigma}^2, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_{g+2}^2]'$ is estimated as $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$.

The EM algorithm in (4.13) requires neither the off-diagonal elements of the $\mathbf{L}_i'\mathbf{W}\mathbf{L}_i$ matrices nor matrices of the form $\mathbf{L}_i'\mathbf{W}\mathbf{L}_j$ with $i \neq j$, as do the scoring and Newton–Raphson algorithms in (4.15). Hence each EM iteration requires considerably fewer computations (sweeping operations) than the other two algorithms. The EM algorithm has two other advantages: it increases the likelihood at each iteration, and the iterates it produces are always nonnegative. On the other hand, the EM algorithm exhibits only linear convergence and is especially slow when one or more of the variance components are close to 0.

All quantities needed to implement the Newton–Raphson, the scoring and the EM algorithms are stored in $\mathbf{Q}_T^*(\mathbf{X} : \mathbf{y} : \mathbf{I} : \mathbf{L}_1 : \mathbf{L}_2 : \ldots : \mathbf{L}_{g+2})$. Hence, it is straightforward to alternate between the three algorithms whenever convergence problems occur.

4.3. Examples

We illustrate our results by applying our algorithms to (1) the Chicago purse snatching data, taken from Harvey (1989), and to (2) the monthly housing starts of privately owned single–family structures, taken from Abraham and Ledolter (1983). We assume convergence if from one iteration to the next the maximum relative change over all parameter estimates is smaller than .001%.

Harvey (1989), page 218 fits a local linear trend (that is, the model in (4.9) without the seasonal components) to the Chicago purse snatching data. This data set consists of the reported purse snatchings in the Hyde park neighborhood of Chicago from January 1968 through September 1973 (the observations are 28 days apart, $T = 71$). We notice slow convergence of the parameter estimates in the local linear trend model, as the slope variance is close to zero.

Following Harvey's suggestion, we set σ_2^2 equal to zero and fit a model with a deterministic slope and a level that follows a random walk. For this model the EM algorithm converges without difficulties. The final estimates are almost identical to those obtained by Harvey. Our estimates of the noise variance and the level variance are 23.9368 and 6.1738, versus 23.93 and 6.174 found by Harvey. It took 114 iterations for our algorithm to converge, although reasonable estimates were obtained long before that point. The starting values for the iterations were naively set to 10.0 for both variance components. Our estimated asymptotic standard errors of the parameter estimates (s.e.) are close to those obtained by Harvey's frequency domain methods. A summary of the convergence history for this data set is given in Table 1.

Table 1. EM for purse snatching data

Iter	σ^2	σ_1^2	σ_2^2	loglike
1	14.6468	11.4637	0	-166.409
2	17.5013	11.5317	0	-165.525
⋮	⋮	⋮	⋮	⋮
22	23.2408	7.0072	0	-164.671
⋮	⋮	⋮	⋮	⋮
100	23.9356	6.1751	0	-164.648
⋮	⋮	⋮	⋮	⋮
113	23.9367	6.1738	0	-164.648
114	23.9368	6.1738	0	-164.648
s.e.	5.613	3.47	n.a.	

We fit a basic structural model with monthly seasonality to the housing starts of privately owned single–family structures from January 1965 to December 1975 ($T = 132$). We use a combination of the scoring and EM algorithm to obtain REML estimates of the four variance components. Since the scoring iterations can yield negative estimates, we simply replace a negative iterate by an estimate that we obtain by dividing the previous iterate by 10. We turn to the EM algorithm once we are unable to further increase the restricted loglikelihood through scoring. The starting values were set at 10000 for each variance component. Table 2 shows the history of convergence for this data set.

The variance components for the slope and the seasonal coefficients are essentially zero. This implies a deterministic slope and a deterministic seasonal pattern. Figure 1 shows the observations and the estimated signal, as well as the forecasts and their associated upper and lower 95% forecast limits for the next year (using the results in Section 3.5).

Table 2. Convergence for housing starts data

Algo	Iter	σ^2	σ_1^2	σ_2^2	σ_3^2	loglike
SCO	1	13990809.735	11224596.074	930728.738	1000	-1121.0089554012
SCO	2	11986772.993	17079492.173	186802.785	100	-1118.8007195076
⋮	⋮	⋮	⋮	⋮		
SCO	9	10548086.619	21591514.717	1.7964531	1E-05	-1117.6430397670
EM	10	10221795.294	20911861.022	1.7964499	1E-07	-1117.5907009885
⋮	⋮	⋮	⋮	⋮		
EM	61	9942260.949	20884456.366	1.7963800	1E-07	-1117.5849238671
EM	62	9942167.686	20884584.483	1.7963784	1E-07	-1117.5849238578
	s.e.	4070165.312	6004292.148	17331.296	242871.962	

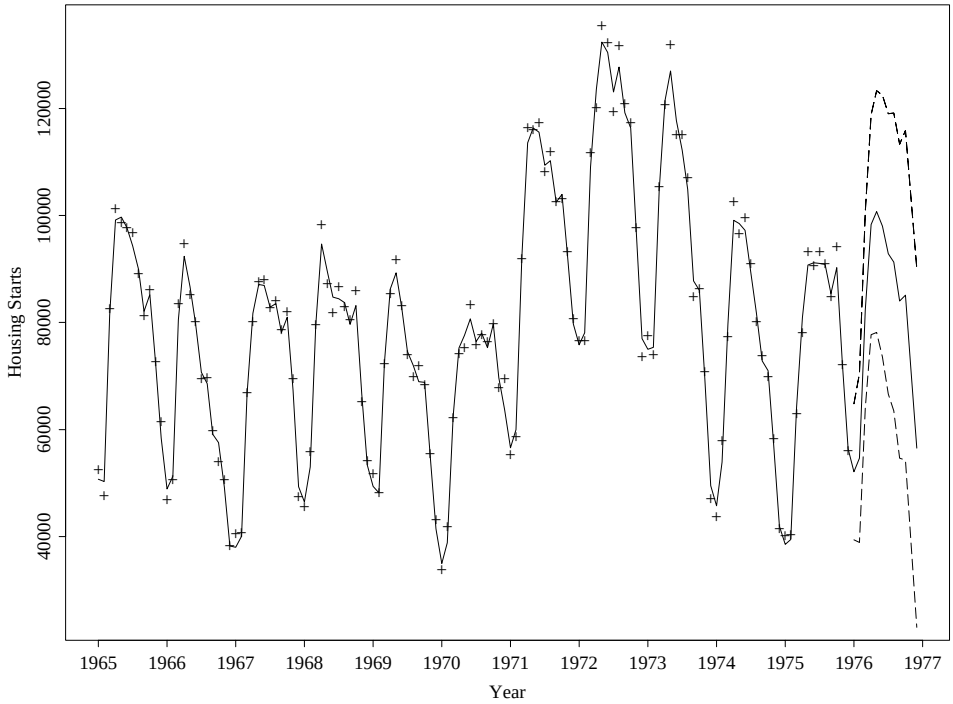

Figure 1. Housing starts: Data (+), signal estimates and monthly forecasts (–) and 95% prediction limits (- -)

References

Abraham, B. and Ledolter, J. (1983). *Statistical Methods for Forecasting.* Wiley, New York.

Abraham, B. and Ledolter, J. (1986). Forecast functions implied by ARIMA models and other related forecast procedures. *Internat. Statist. Rev.* **54**, 51-66.

De Jong, P. (1991). The diffuse Kalman filter. *Ann. Statist.* **19**, 1073-1083.

Goodnight, J. H. (1979). A tutorial on the SWEEP operator. *Amer. Statist.* **33**, 149-158.

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

Jones, R. H. (1993). *Longitudinal Data with Serial Correlation: A State Space Approach.* Chapman Hall, London.

Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (Parametric empirical Bayes models). *J. Amer. Statist. Assoc.* **84**, 717-726.

Sallas, W. M. and Harville, D. A. (1981). Best linear recursive estimation for mixed linear models. *J. Amer. Statist. Assoc.* **76**, 860-869.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components.* John Wiley, New York.

Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, U.S.A.

E-mail: tsimikas@math.umass.edu

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, U.S.A.

E-mail: ledolter@stat.uiowa.edu

(Received October 1995; accepted February 1997)