ARMA series . Additional elements. Kalman filter

School of Statistics for Astrophysics
Variability and Time Series Analysis
Autrans 6-11 October 2019

States space models

We are concerned in this chapter with linear gaussian state space models, also called DLM (dynamic linear models). We first describe the model in its simplest form.
A univariate series (X_t), the state variable, has a dynamic linear evolution in the form:

$$X_t = \phi X_{t-1} + W_t \tag{1}$$

with $\phi \in \mathbb{R}$, $W_t \sim i.i.d. \mathcal{N}(0, \sigma_W^2)$. (1) is called the **state equation**.
The state time series is not observed. Instead we observe a linear transformation of (X_t) :

$$Y_t = AX_t + V_t \tag{2}$$

with $V_t \sim i.i.d. \mathcal{N}(0, \sigma_V^2)$. (2) is called the **observation equation**.
We assume that the initial state $X_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and that X_0, (V_t) and (W_t) are independent.

Given observations $Y_{1:n} = \{Y_1, Y_2, \ldots, Y_n\}$, we are to estimate the parameters $\Theta = \{\mu_0, \Sigma_0, \sigma_W^2, \sigma_V^2, \phi, A\}$ and the state (X_t).

Henceforth we note $X_{t|s}$ the estimation of X_t given $Y_{1:s} = \{Y_1, Y_2, \ldots, Y_s\}$. The problem of obtaining $X_{t|s}$ is called:

- forecasting or prediction when $s < t$
- filtering when $s = t$
- smoothing when $s = n$

The above presentation is rather elementary. In a more realistic situation, we set the model as follows. The state equation has the form:

$$X_t = \Phi X_{t-1} + \Upsilon U_t + W_t \tag{3}$$

where X_t, W_t are p-dimensional vectors, Φ is a $p \times p$ matrix, U_t is a r-vector of exogeneous input series, Υ a $p \times r$ matrix (ΥU_t could be a regression part of the state equation), and $W_t \sim i.i.d.\mathcal{N}(0, Q)$, where the covariance matrix is $p \times p$.

Then the observation equation can be written:

$$Y_t = A_t X_t + \Gamma U_t + V_t \tag{4}$$

where $Y_t \in \mathbb{R}^q$, A_t is $q \times q$, Γ is a $q \times r$ matrix and $V_t \sim i.i.d.\mathcal{N}(0, R)$

Remarks

1. In a first approach we assume that X_0, (V_t) and (W_t) are uncorrelated. This assumption could be relaxed. Comparing (4) and (2), note that now A is allowed to depend on t. This gives more flexibility to the model.
2. Note also that q the dimension of the observation may be larger or smaller than the one of the state.

Example 1. Warming

We are given two different estimates of the temperature: The global mean land-ocean, and the only land's surface based measure.. We assume:

$$\begin{aligned} Y_{1t} &= X_t + V_{1t} \\ Y_{2t} &= X_t + V_{2t} \end{aligned}$$

That is

$$Y_t = \begin{pmatrix} Y_{1t} \\ Y_{2t} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} X_t + \begin{pmatrix} V_{1t} \\ V_{2t} \end{pmatrix}$$

Both variables Y_{1t} and Y_{2t} observe a common signal with different noises.
The state X_t can be modeled as:

$$X_t = \delta + X_{t-1} + W_t$$

that is a random walk with drift.
Here $p = 1$, $q = 2$, $\Upsilon = \delta$, $U_t \equiv 1$ and $\Gamma = 0$.

The following example is of particular interest due to missing data.

Example 2. Blood data
We are interested in biomedical markers in the blood after a bone marrow transplant. The variables are Log(White blood count) (WBC), Log(Platelet) (PLT) and hematocrit (HTC). Missing data are approximatively 40%. The state equation can be written

$$X_t = \begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{3,t-1} \end{pmatrix} + \begin{pmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{pmatrix}$$

and we assume that $Y_t = A_t X_t + V_t$ with $A_t = I$ (blood taken this day) or $A_t = 0$ (no blood taken).
R: plot(blood,type="0",pch=19,xlab="day",main=" ")

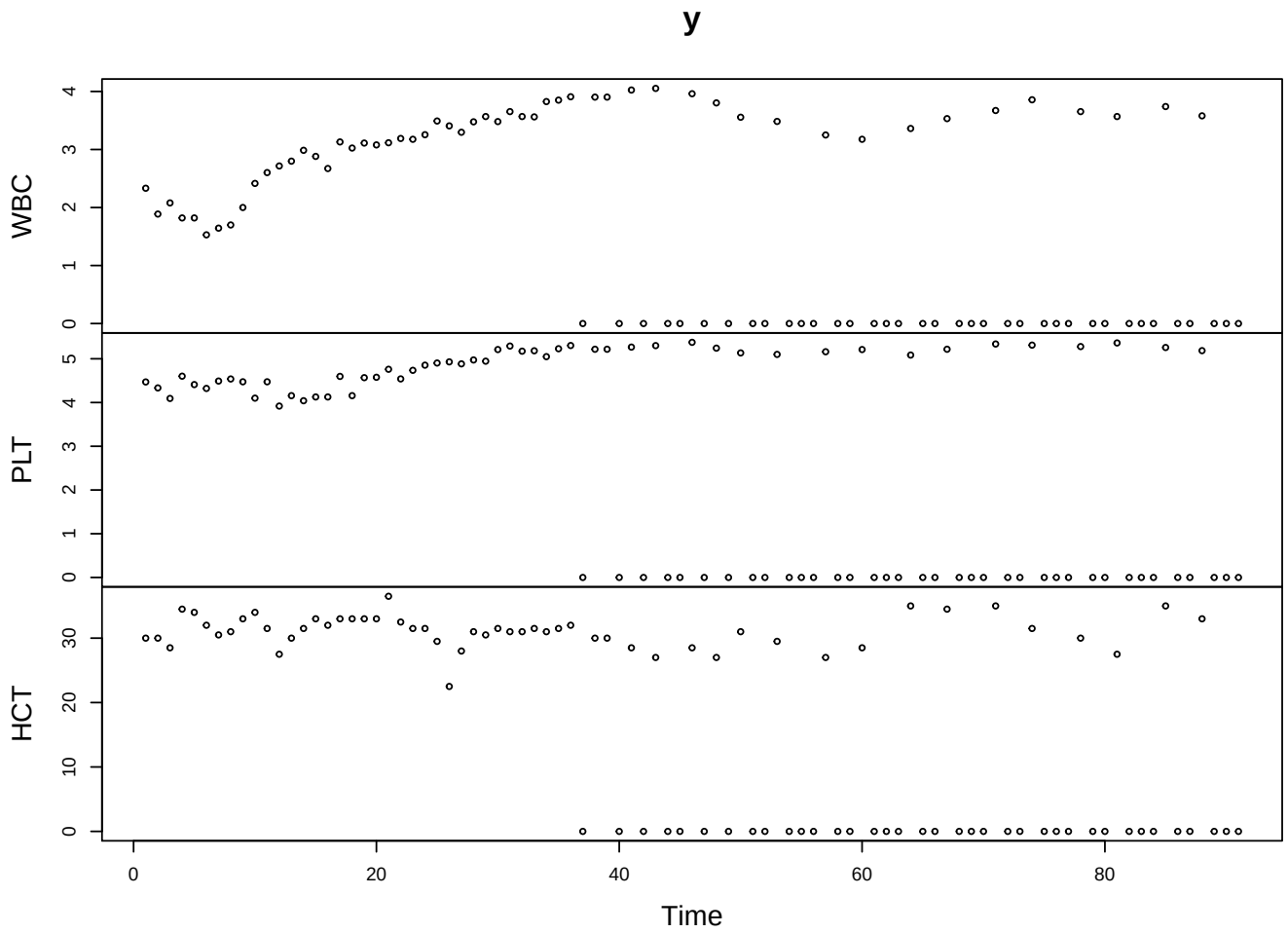

Figure 1: Blood data with missing variables

Example 3. State space representation of a VAR(2) model
Consider Z_t a p-dimensional VAR(2) model:

$$Z_t = \Phi_1 Z_{t-1} \Phi_2 Z_{t-2} W_t$$

where Z_t is a $p \times 1$ vector, $W_t \sim i.i.d. \mathcal{N}(0, \Sigma_p)$.
Set :

$$X_t = \begin{pmatrix} Z_t \\ Z_{t-1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{pmatrix} \begin{pmatrix} Z_{t-1} \\ Z_{t-2} \end{pmatrix} + \begin{pmatrix} W_t \\ 0 \end{pmatrix}$$

and the observation equation is

$$\begin{aligned} Y_t &= (A_t|0) X_t + V_t \\ &= A_t Z_t + V_t \end{aligned}$$

where A_t is $q \times p$ and Y_t is $q \times 1$.

This representation seems not too parsimonious, but in number of situations, appropriate choices of ϕ and A_t may lead to structures with fewer parameters than in a standard multivariate series.

Exercise. AR(1) with observation noise
Consider X_t a univaraite stae variable such that:

$$X_t = \phi X_{t-1} + W_t$$

and Y_t defined by

$$Y_t = X_t + V_t$$

That is Y_t is a noised AR(1).

We assume $|\phi| < 1$, $V_t \sim i.i.d.\mathcal{N}(0, \sigma_V^2)$, $W_t \sim i.i.d.\mathcal{N}(0, \sigma_W^2)$ and $X_0 \sim \mathcal{N}(0, \frac{\sigma_W^2}{1-\phi^2})$.
(V_t), (W_t) and X_0 are independent.
Calculate the ACF of (Y_t). Is (Y_t) an AR(1) process? Compare this ACF with the one of a ARMA(1,1) process.

Example 4. Local level model
We observe $Y_t = \mathrm{t}rend + \mathrm{n}oise$:

$$Y_t = \mu_t + V_t$$

with $V_t \sim i.i.d.\mathcal{N}(0, \sigma_V^2)$, and we assume that μ_t is a random walk:

$$\mu_t = \mu_{t-1} + W_t$$

with $W_t \sim i.i.d.\mathcal{N}(0, \sigma_W^2)$.
Below we will use a sample of (Y_t) to demonstrate the running of the Kalman filter and the Kalman smoother.

In this presentation of the linear gaussian states space model, recall that we are essentially looking for $X_{t|s} = \mathbb{E}(X_t|Y_{1:s})$ and that $X_{t|s}$ is in fact a linear function of $Y_{1:s}$).
We are also interested in the estimation of the quadratic error
$P_{t|s} = \mathbb{E}\left[(X_t - X_{t|s})(X_t - X_{t|s})'\right]$
We give now the description of the Kalman filter for forecasting and filtering.

The Kalman filter

For the model defined in (3) and (4), with initial conditions given by $X_{0|0} = \mu_0$ and $P_{0|0} = \Sigma_0$, for $t = 1, \ldots, n$:

$$X_{t|t-1} = \Phi X_{t-1|t-1} + \Upsilon U_t \quad (5)$$

$$P_{t|t-1} = \Phi P_{t-1|t-1} \Phi' + Q \quad (6)$$

with :

$$X_{t|t} = X_{t|t-1} + K_t(Y_t - A_t X_{t|t-1} - \Gamma U_t) \quad (7)$$

$$P_{t|t} = \left[I - K_t A_t\right] P_{t|t-1} \quad (8)$$

$$K_t = P_{t|t-1} A_t' \left[A_t P_{t|t-1} A_t' + R\right]^{-1} \quad (9)$$

K_t is the gain function.

Note that, for $t = 1, \ldots, n$, the innovations are obtained in the algorithm:

$$\varepsilon_t = Y_t - Y_{t|t-1} = Y_t - A_t X_{t|t-1} - \Gamma U_t \quad (10)$$

and

$$\Sigma_t \equiv \mathbb{C}\mathrm{ov}(\varepsilon_t) = \mathbb{C}\mathrm{ov}(A_t(X_t - X_{t|t-1}) + V_t) = A_t P_{t|t-1} A_t' + R \quad (11)$$

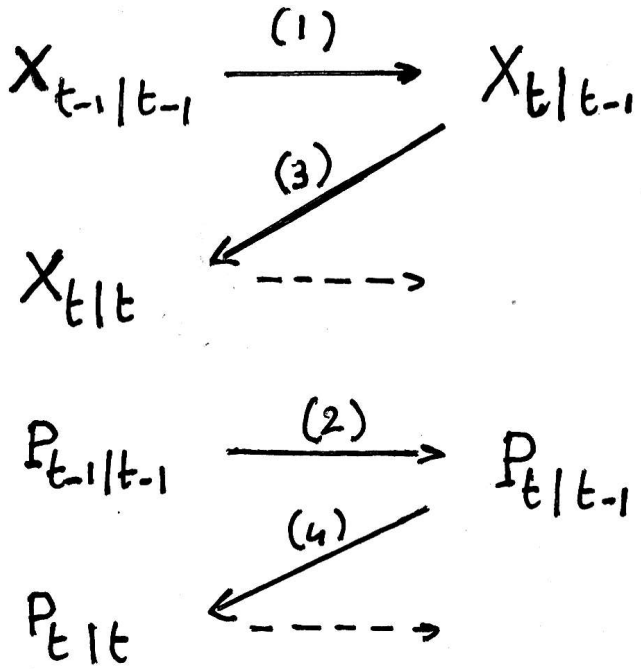

Figure 2: Principle of the Kalman filter

Time varying

An important propert is that the result above, with appropriate substitutions, hold true when any or all the parameters are time-dependent, $\phi = \phi_t$, $\Upsilon = \Upsilon_t$, $Q = Q_t$ in the state equation, or $\Gamma = \Gamma_t$, $R = R_t$ in the observation equation, or the dimension itself $q = q_t$.
In particular this property allows to deal with missing data.

The Kalman smoother

We consider now the problem of estimating X_t, for $t \leq n$, given observations $Y_1, \ldots, Y_n$, that is defining the *Kalman smoother* $X_{t|n}$. Note that in fact , $\{X_{t|n}; t = 1, \ldots, n\}$ is smoother than the forecasts $\{X_{t|t-1}; t = 1, \ldots, n\}$ or the filter $\{X_{t|t}; t = 1, \ldots, n\}$.

The Kalman smoother

For the model defined in (3) and (4), with initial conditions $X_{n|n}$ and $P_{n|n}$ obtained in the Kalman filter, for for $t = n, n-1 \ldots, 1$:

$$X_{t-1|n} = X_{t-1|t-1} + J_{t-1}(X_{t|n} - X_{t|t-1}) \quad (12)$$

$$P_{t-1|n} = P_{t-1|t-1} + J_{t-1}(P_{t|n} - P_{t|t-1})J'_{t-1} \quad (13)$$

where

$$J_{t-1} = P_{t-1|t-1}\Phi' P^{-1}_{t|t-1}$$

Exercise. Running the Kalman filter and smoother on a sample of the local level model

Fill in the table given below, using firstly the KF and then the KS.

t	y_t	μ_t	$\mu_{t\|t-1}$	$P_{t\|t-1}$	$\mu_{t\|t}$	$P_{t\|t}$	$\mu_{t\|n}$	$P_{t\|n}$
0					0.0	1.0		
1	-0.05							
2	-1.90							
3	-1.90							
4	1.77							
5	-0.22							
6	0.30							
7	2.00							
8	2.45							
9	1.92							
10	3.75							

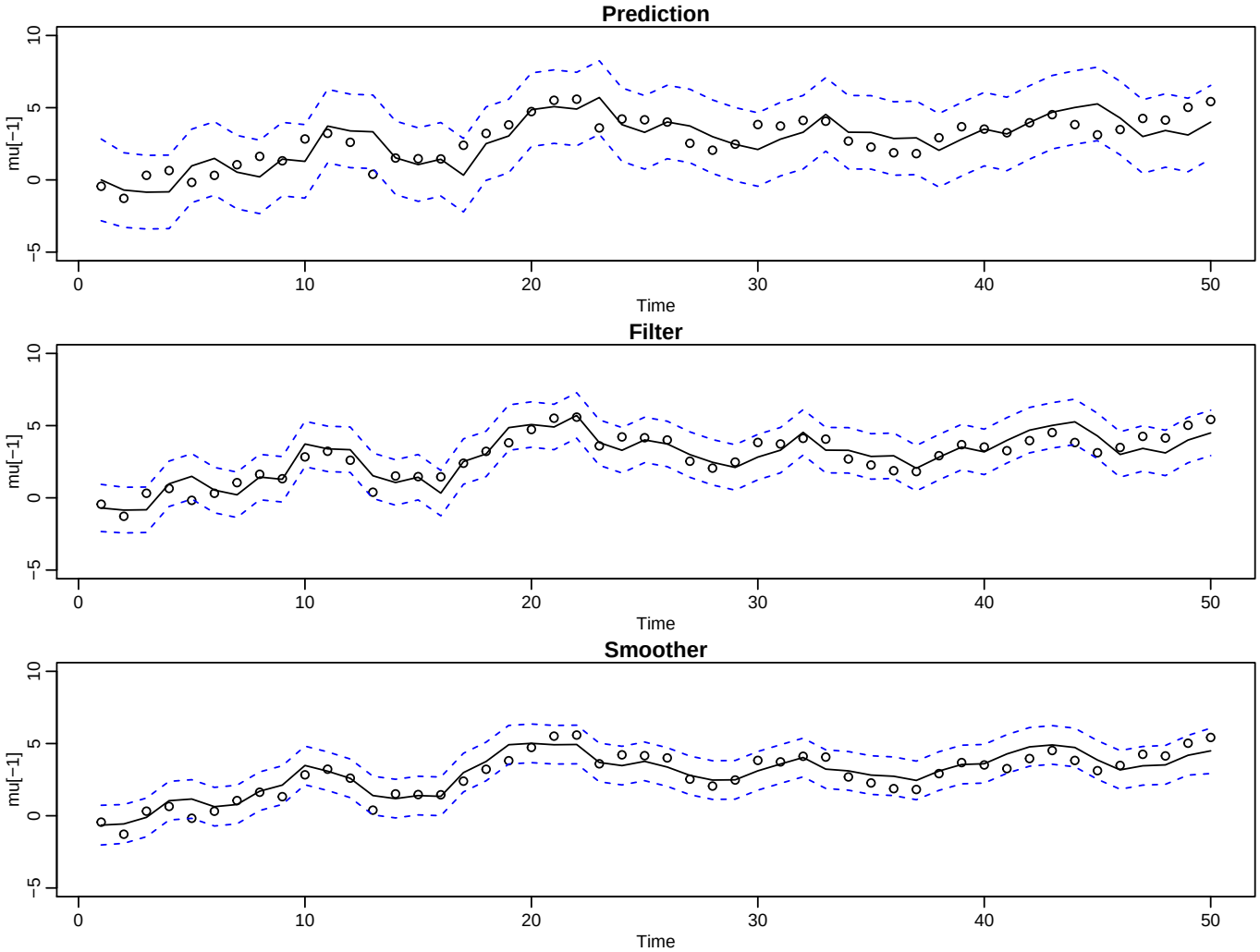

Figure 3: The simulated local level model

Maximum likelihood estimation

Recall that we must estimate $\theta = \{\mu_0, \Sigma_0, \Phi, Q, R\}$. Two methods come from similar approaches developped in other contexts:

- one based on Newton-Raphson
- the other one based on the EM algorithm

a) Newton-Raphson

The maximum likelihood estimation is used under the assumtion that $X_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and (W_t) and (V_t) are each gaussian noise.

The likelihood is computed with the innovations defined by (10)

$$\varepsilon_t = Y_t - Y_{t|t-1} = Y_t - A_t X_{t|t-1} - \Gamma U_t \tag{14}$$

and their covariance matrix by:

$$\Sigma_t = A_t P_{t|t-1} A_t' + R \tag{15}$$

Hence we may write the likelihood $L(\theta; Y_{1:n})$:

$$-Ln\, L(\theta; Y_{1:n}) = \frac{1}{2} \sum_{t=1}^{n} Ln|\Sigma_t(\theta)| + \frac{1}{2} \varepsilon_t'(\theta) \Sigma_t(\theta)^{-1} \varepsilon_t(\theta) \tag{16}$$

which can be minimized by performing a Newton-Raphson procedure.

a) Algorithm EM.
The idea is that if we had in hand the values $X_0, \ldots, X_n$, we could write the complete likelihood of the involved variables:

$$L(\theta; X_{0:n}, Y_{1:n}) = p^{\mu_0 \Sigma_0}(X_0) \prod_{t=1}^{n} p^{\phi, Q}(X_t | X_{t-1}) \prod_{t=1}^{n} p^{R}(Y_t | X_t) \tag{17}$$

Hence starting with a guessed value of the parameter θ we estimate the "hidden" variables and so the complete likelihood.

We choose initial values for the parameter $\theta = \{\mu_0, \Sigma_0, \Phi, Q, R\}$. Then roughly speaking, at iteration k with $\theta^{(k-1)}$ from the previous iteration:

- Expectation step. We use the Kalman filter to get estimates of $X_{0:n}$, and then the estimate of the complete likelihood conditionaly on $Y_{1:n}$.
- Maximisation step. We maximize the estimated complete likelihood to get the new estimate $\theta^{(k)}$

We repeat the two steps to convergence.

Dealing with missing data in state space models

The principle consists in making use of the flexibility of the model, in particular the possibility of A to change with t.

Suppose that, at time t, Y_t can be partitioned in such a way that the first q_{1t} components are observed and the q_{2t} other ones are unobserved, $q_{1t} + q_{2t} = q$.

$$\begin{pmatrix} Y_t^{(1)} \\ Y_t^{(2)} \end{pmatrix} = \begin{pmatrix} A_t^{(1)} \\ A_t^{(2)} \end{pmatrix} X_t + \begin{pmatrix} V_t^{(1)} \\ V_t^{(2)} \end{pmatrix}$$

where $A_t^{(1)}$ is $q_{1t} \times p$ and $A_t^{(2)}$ is $q_{2t} \times p$, and

$$COV \begin{pmatrix} V_t^{(1)} \\ V_t^{(2)} \end{pmatrix} = \begin{pmatrix} R_{11t} & R_{12t} \\ R_{21t} & R_{22t} \end{pmatrix}$$

Since $Y_t^{(2)}$ is not observed, we may rewrite the observation equation in the following way

$$X_t = \Phi X_{t-1} + W_t \quad \text{and} \quad Y_t^{(1)} = A_t^{(1)} X_t + V_t^{(1)}$$

where the dimension of the observation equation is now q_{1t} at time t. The filter algorithm holds with appropriate substitution.

Note that if $q_{2t} = q$ (the complete vector Y_t is unobserved) then $A_t = 0$, $K_t = 0$, $X_{t|t} = X_{t|t-1}$ and $P_{t|t} = P_{t|t-1}$.

Another possibilty is to keep the q parameter constant, zeroing the terms related to unobserved components. We substitute

$$Y_{(t)} = \begin{pmatrix} Y_t^{(1)} \\ 0 \end{pmatrix}, \quad A_{(t)} = \begin{pmatrix} A_t^{(1)} \\ 0 \end{pmatrix}, \quad R_{(t)} = \begin{pmatrix} R_{11t} & 0 \\ 0 & I_{22t} \end{pmatrix}$$

for Y_t, A_t and R. Thene the innovations become:

$$\varepsilon_{(t)} = \begin{pmatrix} \varepsilon_t^{(1)} \\ 0 \end{pmatrix}, \quad \text{and} \quad \Sigma_{(t)} = \begin{pmatrix} A_t^{(1)} P_{t|t-1} (A_t^{(1)})' + R_{11t} & 0 \\ 0 & I_{22t} \end{pmatrix}$$

With these substitutions, the m.l.e via the innovations can proceed as in the usual case. Hence we can get filtered values for missing data and then proceed with the smoother using the filtered missing data.

Example. Blood data

Note that in this data set $p = q = 3$, and missing data are such that $q_{2t} = 3$ (all the three components are missing).

```
y    = cbind(WBC, PLT, HCT)
num  = nrow(y)
A    = array(0, dim=c(3,3,num))  # creates num 3x3 zero matrices
for(k in 1:num) if (y[k,1] > 0) A[,,k]= diag(1,3)
plot(WBC,type="l")

# Initial values
mu0    = matrix(0,3,1)
Sigma0 = diag(c(.1,.1,1) ,3)
Phi    = diag(1,3)
cQ     = diag(c(.1,.1,1), 3)
cR     = diag(c(.1,.1,1), 3)
(em = EM1(num, y, A, mu0, Sigma0, Phi, cQ, cR, 100, .001))
```

```
ks  = Ksmooth1(num, y, A, em$mu0, em$Sigma0, em$Phi, 0, 0,
+ chol(em$Q), chol(em$R), 0)
y1s = ks$xs[1,,]
y2s = ks$xs[2,,]
y3s = ks$xs[3,,]
p1  = 2*sqrt(ks$Ps[1,1,])
p2  = 2*sqrt(ks$Ps[2,2,])
p3  = 2*sqrt(ks$Ps[3,3,])
par(mfrow=c(3,1))
tsplot(WBC, type='p', pch=19, ylim=c(1,5), xlab='day')
lines(y1s)
lines(y1s+p1, lty=2, col=4)
lines(y1s-p1, lty=2, col=4)
tsplot(PLT, type='p', ylim=c(3,6), pch=19, xlab='day')
lines(y2s)
lines(y2s+p2, lty=2, col=4)
lines(y2s-p2, lty=2, col=4)
tsplot(HCT, type='p', pch=19, ylim=c(20,40), xlab='day')
lines(y3s)
lines(y3s+p3, lty=2, col=4)
lines(y3s-p3, lty=2, col=4)
```

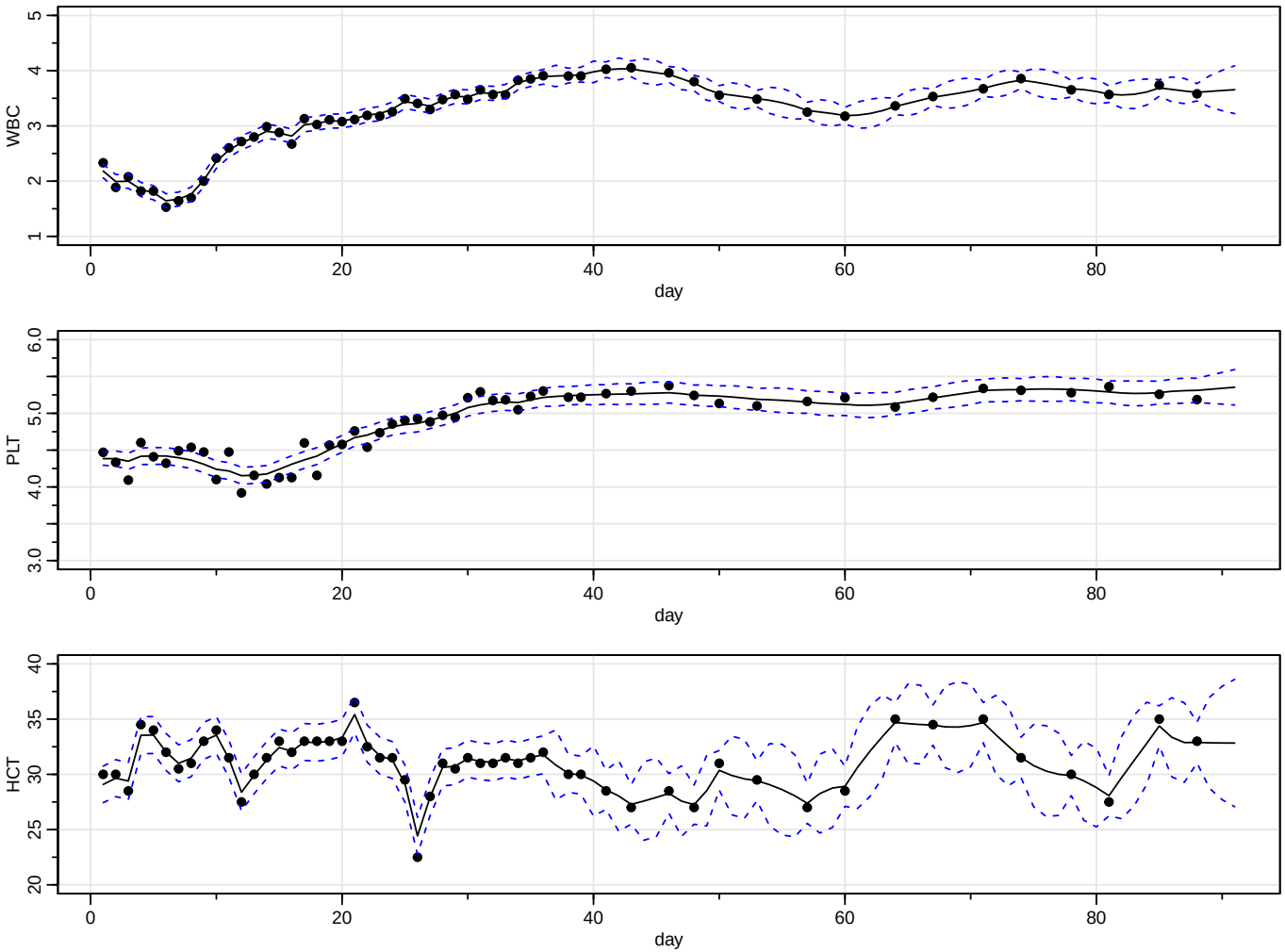

Figure 4: Blood data with missing variables