CHAPTER 20

DATA ASSIMILATION

In this homework set you will apply the Kalman Filter to assimilate observations using a vector autoregressive model. A vector autoregressive model allows probabilities of forecasted variables to be solved exactly, so it illustrates the Kalman Filter in an "ideal" situation. However, this ideal situation is very unrealistic, so there is potential for being mislead.

A vector autoregressive model is of the form

$$\begin{array}{cccccc} \mathbf{y}_t & = & \mathbf{A} & \mathbf{y}_{t-1} & + & \mathbf{w}_t \\ M \times 1 & & M \times M & M \times 1 & & M \times 1\text{'} \end{array} \tag{20.1}$$

where $\mathbf{A}$ is a matrix called the *dynamical operator* and $\mathbf{w}_t$ is Gaussian white noise with zero mean and covariance matrix $\mathbf{Q}$. After specifying the matrices $\mathbf{A}$ and $\mathbf{Q}$ and an initial condition $\mathbf{y}_0$, the vectors $\mathbf{y}_1, \mathbf{y}_2, \dots$ can be solved for a particular realization of the noise $\mathbf{w}_t$.

In this homework set, you will consider the following experimental situation. First, you will create a vector time series $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{100}$ from (20.1) for a particular choice of $\mathbf{A}$, $\mathbf{Q}$, and $\mathbf{y}_0$ and for a particular realization of $\mathbf{w}_t$. The resulting vector time series will be the "truth." To mimic reality, we will pretend that we do not know the truth. Instead, you will create "observations" using the model

$$\begin{array}{cccccc} \mathbf{o}_t & = & \mathbf{H} & \mathbf{y}_t & + & \mathbf{r}_t \\ K \times 1 & & K \times M & M \times 1 & & K \times 1\text{'} \end{array} \tag{20.2}$$

for a particular choice of $\mathbf{H}$ and covariance matrix $\mathbf{R}$ of $\mathbf{r}_t$. The observations $\mathbf{o}_t$ are known and our goal is to estimate the truth $\mathbf{y}_t$ from these observations (pretending that we don't know the truth). You will use the Kalman Filter to assimilate these observations to estimate the true state. Then you will compare this estimate with the true state to see how well the Kalman Filter works.

Download the R program `kf_student.R` from the class website. This program generates a time series from (20.1) for the choice

$$\mathbf{A} = \begin{pmatrix} 0.9 & 0.5 & 0.3 \\ 0.0 & 0.5 & 2.0 \\ 0.0 & 0.0 & 0.4 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{y}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \tag{20.3}$$

and generates "observations" using the choice

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{R} = 49. \tag{20.4}$$

Note that this choice implies that only one variable (i.e., the first element of $\mathbf{y}_t$) is observed with noise. The program assumes that $\mathbf{Q}$ is diagonal, so that the elements of the noise vector $\mathbf{w}_t$ are independent and have variances equal to the diagonal elements of $\mathbf{Q}$. This special form allows the noise vector to be specified very efficiently numerically as

```
w = rnorm(M,sd=sqrt(diag(Q)))
```

Similarly, $\mathbf{R}$ is assumed to be diagonal and hence the observations generated by (20.2) can be specified efficiently in a similar way. This is done in the R code that you will download.

After the R code runs (hopefully without error messages!), you should have the following quantities available:

```
# NDIM:  THE DIMENSION OF THE VECTOR AUTOREGRESSIVE MODEL (VAR); DEFAULT = 3
# NTOT:  THE TOTAL LENGTH OF THE TIME SERIES; DEFAULT = 100
# NOBS:  NUMBER OF OBSERVATIONS PER TIME STEP; DEFAULT = 1
# X:     [NDIM, NTOT] MATRIX SPECIFYING THE "TRUTH" AT EACH TIME STEP
# OBS:   [NOBS, NTOT] MATRIX SPECIFYING THE "OBSERVATIONS" AT EACH TIME STEP
# DYNOP: [NDIM, NDIM] MATRIX SPECIFYING THE DYNAMICAL OPERATOR OF THE VAR
# HOP:   [NOBS, NDIM] MATRIX SPECIFYING THE INTERPOLATION OPERATOR
# Q.COV: [NDIM, NDIM] MATRIX SPECIFYING THE NOISE COVARIANCE MATRIX OF THE VAR
# R.COV: [NOBS, NOBS] MATRIX SPECIFYING THE ERROR COVARIANCE MATRIX OF THE OBS
```

Exercise 20.1. Plot the first element of the true time series. Also plot the corresponding observations. The result should look something like the curve and dots in fig. 20.1. □

Exercise 20.2. Write a function to evaluate the Kalman Filter equations

$$\boldsymbol{\mu}_t^A = \boldsymbol{\mu}_t^B + \boldsymbol{\Sigma}_t^B \mathbf{H}_t^T \left(\mathbf{H}_t \boldsymbol{\Sigma}_t^B \mathbf{H}_t^T + \mathbf{R}_t \right)^{-1} \left(\mathbf{o}_t - \mathbf{H}_t \boldsymbol{\mu}_t^B \right) \tag{20.5}$$

$$\boldsymbol{\Sigma}_t^A = \boldsymbol{\Sigma}_t^B - \boldsymbol{\Sigma}_t^B \mathbf{H}_t^T \left(\mathbf{H}_t \boldsymbol{\Sigma}_t^B \mathbf{H}_t^T + \mathbf{R}_t \right)^{-1} \mathbf{H}_t \boldsymbol{\Sigma}_t^B. \tag{20.6}$$

The function call should be

```
kalman.population = function(mub,sigmab,hop,r.cov,obs) {
##########
## EVALUATES THE KALMAN FILTER EQUATIONS FOR THE MEAN AND COVARIANCE MATRIX OF THE ANALYSIS
## BASED ON *POPULATION* COVARIANCE MATRICES AND MEANS
## INPUT:
#    MUB:    [NDIM] MEAN OF THE BACKGROUND DISTRIBUTION
#    SIGMAB: [NDIM,NDIM] BACKGROUND COVARIANCE MATRIX
#    HOP:    [NOBS,NDIM] INTERPOLATION OPERATOR
#    R.COV:  [NOBS,NOBS] COVARIANCE MATRIX OF THE OBSERVATIONAL ERROR
#    OBS:    [NOBS] THE OBSERVATIONS
## OUTPUT
#    MUA:    [NDIM] VECTOR OF THE ANALYSIS DISTRIBUTION
#    SIGMAA: [NDIM,NDIM] COVARIANCE MATRIX OF THE ANALYSIS DISTRIBUTION
```

As you can see from the equations, you will need to compute the inverse of a matrix. You should invert the matrix using the commands `chol2inv(chol(MATRIX))`, which takes advantage of the fact that the matrix being inverted is symmetric. □

Exercise 20.3. The background distribution is obtained from the *previous* analysis. In particular, the mean and covariance matrix of the background distribution are

$$\boldsymbol{\mu}_t^B = \mathbf{A}\boldsymbol{\mu}_{t-1}^A \quad \text{and} \quad \boldsymbol{\Sigma}_t^B = \mathbf{A}\boldsymbol{\Sigma}_{t-1}^A\mathbf{A}^T + \mathbf{Q}. \tag{20.7}$$

Assume the initial analysis $\boldsymbol{\mu}_0^A = \mathbf{0}$ and

$$\boldsymbol{\Sigma}_0^A = \begin{pmatrix} 400 & 0 & 0 \\ 0 & 400 & 0 \\ 0 & 0 & 400 \end{pmatrix}. \tag{20.8}$$

Combine these equations together with your Kalman Filter function to construct an analysis for one hundred time steps. Plot the mean analysis and its uncertainty for the first element of $\mathbf{y}_t$. The result should look something like the grey shading in fig. 20.1, which shows $(\boldsymbol{\mu}^A)_1 \pm \sqrt{(\boldsymbol{\Sigma}^A)_{11}}$. □

Exercise 20.4. Plot the standard error of the analysis for each element of $\mathbf{y}_t$ as a function of time (i.e., plot the square roots of the diagonal elements of $\boldsymbol{\Sigma}_t^A$). What happens to the errors after a long time? □

Exercise 20.5. In this problem only, set $\mathbf{R}_t = 1$ and show the analysis, observations (if available), and the truth for the first and second elements of $\mathbf{y}_t$. How do these plots differ from the previous case? Explain why this difference makes sense. □

Exercise 20.6. Suppose that, instead of observing $(\mathbf{y}_t)_1$, we can observe only the sum $z(t) = (\mathbf{y}_t)_1 + (\mathbf{y}_t)_2$. Furthermore, suppose the error of this observation is normally distributed with zero mean and variance 25. Explain how you would assimilate these observations (e.g., explain how you would change your R code to handle this case). Actually do this and show the standard error of the analysis and the truth for $(\mathbf{y}_t)_1$. □

Exercise 20.7 (Least Squares Derivation of the Kalman Filter Equations)**.** The Kalman Filter equations can be derived by the method of least squares. The essential problem is that we are given a set of observations $\mathbf{o}$, and based on this want to estimate the state $\mathbf{x}$. To do this, we assume a linear prediction model

$$\hat{\mathbf{y}} = \mathbf{A}\mathbf{o} + \mathbf{b}. \tag{20.9}$$

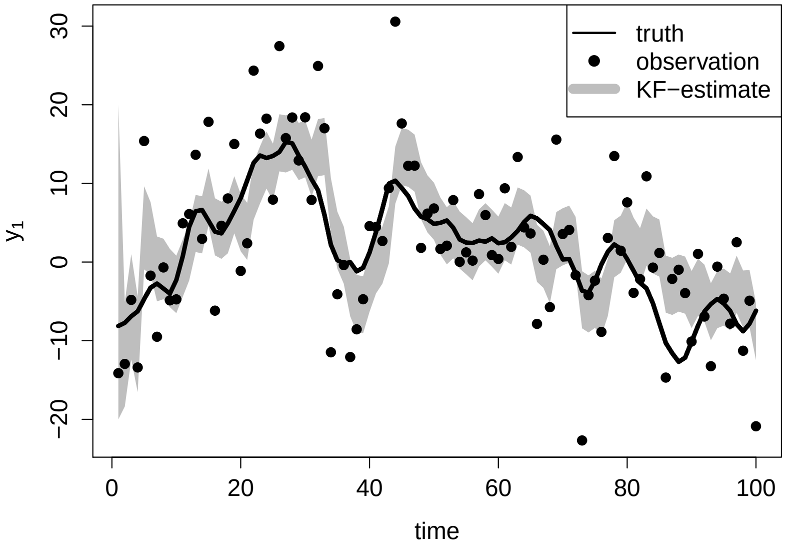

Figure 20.1 A particular realization of the first element $y_1(t)$ generated by the vector autoregressive model (20.1) (solid curve), corresponding observations (dots), and the analysis mean plus a standard deviation, as estimated state by the Kalman Filter (grey shading).

Show that the method of least squares gives

$$\mathbf{A} = \mathbf{\Sigma}_{YO}\mathbf{\Sigma}_O^{-1} \quad \text{and} \quad \mathbf{b} = \mathbb{E}[\mathbf{y}] - \mathbf{A}\mathbb{E}[\mathbf{o}], \tag{20.10}$$

where the expectation operator $E[\cdot]$ is an average over all possible realizations of the observations and background quantities. Under this definition, $E[\mathbf{y}] = \boldsymbol{\mu}_B$ and $\text{cov}[\mathbf{y}] = \mathbf{\Sigma}_B$. The observations are assumed to be related to the state $\mathbf{y}$ through the model

$$\mathbf{o} = \mathbf{H}\mathbf{y} + \mathbf{r}, \tag{20.11}$$

where $\mathbf{H}$ is the interpolation operator and $\mathbf{r}$ is independent and normally distributed with zero mean and covariance matrix $\mathbf{\Sigma}_R$. Show that this equation implies

$$\mathbf{\Sigma}_{OY} = \mathbf{H}\mathbf{\Sigma}_Y = \mathbf{H}\mathbf{\Sigma}_B \tag{20.12}$$

and

$$\mathbf{\Sigma}_O = \mathbf{H}\mathbf{\Sigma}_Y\mathbf{H}^T + \mathbf{\Sigma}_R = \mathbf{H}\mathbf{\Sigma}_B\mathbf{H}^T + \mathbf{\Sigma}_R. \tag{20.13}$$

and

$$\mathbb{E}[\mathbf{o}] = \mathbf{H}\boldsymbol{\mu}_B \tag{20.14}$$

Substituting these equations into (20.10) gives

$$\mathbf{A} = \mathbf{\Sigma}_B\mathbf{H}^T\left(\mathbf{H}\mathbf{\Sigma}_B\mathbf{H}^T + \mathbf{\Sigma}_R\right)^{-1} \tag{20.15}$$

and

$$\mathbf{b} = \boldsymbol{\mu}_B - \boldsymbol{\Sigma}_B \mathbf{H}^T \left(\mathbf{H} \boldsymbol{\Sigma}_B \mathbf{H}^T + \boldsymbol{\Sigma}_R \right)^{-1} \mathbf{H} \mu_B. \tag{20.16}$$

It follows that the linear regression model (20.9) yields

$$\mathbf{y}_a = \boldsymbol{\mu}_B + \boldsymbol{\Sigma}_B \mathbf{H}^T \left(\mathbf{H} \boldsymbol{\Sigma}_B \mathbf{H}^T + \boldsymbol{\Sigma}_R \right)^{-1} \left(\mathbf{o} - \mathbf{H} \boldsymbol{\mu}_B \right), \tag{20.17}$$

which is precisely the Kalman Filter equation for the mean analysis.

Show that the error covariance of the regression model (20.9) is

$$\begin{aligned}
\boldsymbol{\Sigma}_a &= \mathbb{E} \left[\left(\mathbf{y} - \mathbf{y}_a \right) \left(\mathbf{y} - \mathbf{y}_a \right)^T \right] \\
&= \mathbb{E} \left[\left(\mathbf{y} - \boldsymbol{\mu}_B - \mathbf{A}(\mathbf{o} - \mathbf{H}\boldsymbol{\mu}_B) \right) \left(\mathbf{y} - \boldsymbol{\mu}_B - \mathbf{A}(\mathbf{o} - \mathbf{H}\boldsymbol{\mu}_B) \right)^T \right] \\
&= \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{YO} \mathbf{A}^T - \mathbf{A} \boldsymbol{\Sigma}_{OY} + \mathbf{A} \boldsymbol{\Sigma}_O \mathbf{A}^T \\
&= \boldsymbol{\Sigma}_B - \boldsymbol{\Sigma}_B \mathbf{H}^T \left(\mathbf{H} \boldsymbol{\Sigma}_B \mathbf{H}^T + \boldsymbol{\Sigma}_R \right)^{-1} \mathbf{H} \boldsymbol{\Sigma}_B, \qquad (20.18)
\end{aligned}$$

which is precisely the Kalman Filter equation for the analysis error covariance.

□