

Ecological Archives A014-023-A1 and A2

Appendix A: Kalman filtering for maximum likelihood estimation given corrupted observations.

Introduction

The Kalman filter is used to extend likelihood estimation to cases with hidden states, such as when observations are corrupted and the true population size is unobserved. The following algorithm is based on section 3.4 in Harvey (1989), which was used by Lindley (2003) for estimation for population processes. The Kalman filter is well-known and widely used in engineering and computer science applications. There are a multitude of books on the Kalman filter, including Harvey (1989). One of the more penetrable introductions of the Kalman filter alone (but not on maximum likelihood estimation) is chapter 1 of Maybeck (1979).

The state-space model

The diffusion approximation for a stochastic exponential growth model can be written as a linear state space model (written in the notation familiar in the engineering literature):

$$x_{t+1} = x_t + B + w_t, \quad \text{where } w_t \sim \text{normal}(0, Q) \quad [\text{A.1}]$$

$$y_t = x_t + v_t, \quad \text{where } v_t \sim f(0, R) \quad [\text{A.2}]$$

where $x_t = \log N_t$ is the true log population size and $y_t = \log O_t$ is the log observations of the population size. B is \mathbf{m} the mean population growth rate. Q is the \mathbf{s}^2 , otherwise known as the process error or environmental variability. R is the variability associated with sampling error or other non-process error. Only y_t is observed; the underlying parameters, B , Q , and R , and the underlying true population size, x_t , is hidden. If we make the assumption that v_t is normally distributed, then the model is a linear Gaussian state-space model.

We can calculate the probability of the observed time series, $\{y\}_1^T \equiv \{y_1, y_2, \dots, y_T\}$, as follows:

$$P(\{y\}_1^T) = \prod_{t=1}^T P(y_t | \{y\}_1^{t-1}) \quad [\text{A.3}]$$

where $P(y_t | \{y\}_1^{t-1})$ is the probability of y_t given all the observations before time t and

$P(y_1 | \{y\}_1^0) \equiv P(y_1)$. Denote the expected value of $(y_t | \{y\}_1^{t-1})$ as \tilde{y}_t^{t-1} . The conditional

probability $(y_t | \{y\}_1^{t-1})$ is distributed normal with a mean \tilde{y}_t^{t-1} and some variance, denoted F_t^{t-1} ,

which depends on the particular parameters, $\Psi = \{B, Q, R\}$, that generated the data. Thus, the

probability of the time series given a particular set of parameters, Ψ , is

$$P(\{y\}_1^T | \Psi) = \prod_{t=1}^T \exp\left\{-\frac{(y_t - \tilde{y}_t^{t-1})^2}{2F_t^{t-1}}\right\} (2\mathbf{p} | F_t^{t-1})^{-1/2} d\mathbf{J} \quad [\text{A.4}]$$

from the probability density of a normal with mean \tilde{y}_t^{t-1} and variance F_t^{t-1} . The log likelihood

of Ψ given the data, $\{y\}_1^T$, is

$$\log L(\Psi | \{y\}_1^T) = -\frac{T}{2} \log 2\mathbf{p} - \frac{1}{2} \sum_{t=1}^T \log |F_t^{t-1}| - \frac{1}{2} \sum_{t=1}^T \frac{(y_t - \tilde{y}_t^{t-1})^2}{F_t^{t-1}} + \text{a constant.} \quad [\text{A.5}]$$

For Eqn. A.5, we need estimates of $\tilde{y}_t^{t-1} \equiv E(y_t | \{y\}_1^{t-1})$ and $F_t^{t-1} = E(y_t y_t | \{y\}_1^{t-1})$. Observe from

Eqn. A.1 that

$$\begin{aligned} E(y_t | \{y\}_1^{t-1}) &= E(x_t | \{y\}_1^{t-1}) \\ E(y_t y_t | \{y\}_1^{t-1}) &= E(x_t x_t | \{y\}_1^{t-1}) + R \end{aligned} \quad [\text{A.6}]$$

The Kalman filter below gives optimal estimates of $E(x_t | \{y\}_1^{t-1})$ and $E(x_t x_t | \{y\}_1^{t-1})$ which are

then used in Eqn. A.6 to calculate the log likelihood of Ψ .

The maximum likelihood estimates of B , Q , and R are found by using some type of maximization routine on Eqn. A.6 to find the set of parameters $\psi = \{B, Q, R\}$ that maximize the likelihood. Matlab code for this algorithm is given at the end of this appendix.

The Kalman filter

First, some notation:

$$\begin{aligned}\{y\}_1^t &\equiv \{y_1, y_2, \dots, y_t\} \\ x_t^t &\equiv E[x_t | \{y\}_1^t] \\ V_t^t &\equiv E[x_t x_t | \{y\}_1^t]\end{aligned}$$

The Kalman recursion: Start at $t = 1$ and step forward to T . Assume an initial $x_t \equiv \mathbf{p}_1$ and initial $V_1^0 \equiv V_1$ to start the recursion. One could let these be free variables and find the maximum likelihood values when maximizing Eqn. A.6, but that is not done here and the algorithm should not be very sensitive to these starting values. At each step, compute:

$$\begin{aligned}x_t^{t-1} &= \begin{cases} \mathbf{p}_1 & \text{for } t = 1 \\ x_{t-1}^{t-1} + B & \text{for } t > 1 \end{cases} \\ V_t^{t-1} &= \begin{cases} V_1 & \text{for } t = 1 \\ V_{t-1}^{t-1} + Q & \text{for } t > 1 \end{cases} \\ K_t &= \frac{V_t^{t-1}}{(V_t^{t-1} + R)} \\ x_t^t &= x_t^{t-1} + K_t (y_t - x_t^{t-1}) \\ V_t^t &= V_t^{t-1} - K_t V_t^{t-1}\end{aligned}$$

This is the well-known Kalman filter, but it looks a little different than what you'll see in engineering texts. First generally it is assumed that y_t is a series of measurements from multiple instruments, thus the Kalman filter is always written in matrix form. Here since y_t is one measurement, it can be written in scalar form. Second, the Kalman filter is usually presented for

the model $x_{t+1} = Ax_t + Bu_t + w_t$, $y_t = Cx_t + v_t$. In this application, $A=1$, $C=1$ and $u_t=1$, so the filter is simplified quite a bit.

References

Harvey, A. C. 1991. Forecasting, structural time series models and the Kalman filter.

Cambridge University Press, Cambridge, UK.

Maybeck, P. S. 1979. Stochastic models, estimation and control. Volume 1. Academic Press,

New York, USA.

Matlab code

```
function [mu,sigma2,sigma2np]=kalman_ests(data)

y=log(data);

%Start with some reasonable initial parameter estimates
muest=mean(y(2:end)-y(1:(end-1)));
tmp1=var(y(2:end)-y(1:(end-1)));
tmp4=var(y(5:end)-y(1:(end-4)));
sigma2est=(tmp4-tmp1)/3;
sigma2npest=(var(y(2:end)-y(1:(end-1)))-max(0.0001,sigma2est))/2;
pil=max(0.0001,sigma2est)+max(0.0001,sigma2npest); %var of y(1)

%log transform the variances so the search algorithm doesn't give negative
% variances
startvals=[muest log(max(0.0001,sigma2est)) log(max(0.0001,sigma2npest))];
%fminsearch is a Nelder-Mead minimization matlab function
a=fminsearch('kalman_loglik',startvals,[],y,y(1),pil);

MLmuest=a(1);
MLsigma2est=exp(a(2));
MLsigma2npest=exp(a(3));

function negloglik = kalman_loglik(v,y,initx,V1)

T=length(y);
B = v(1); %mu
Q = exp(v(2)); %s2
R = exp(v(3)); %s2np

%initialize
xtt=zeros(1,T); Vtt=zeros(1,T); xtt1=zeros(1,T); Vtt1=zeros(1,T);
xtT=zeros(1,T); VtT=zeros(1,T); J=zeros(1,T); Vtt1T=zeros(1,T);
```

```

Ft=zeros(1,T); vt=zeros(1,T);

%forward pass gets you E[x(t) given y(1:t)]
x10=initx;
V10=V1;
for(t=1:T)
    if(t==1)
        xtt1(1) = initx; %denotes  $x_1^0$ 
        Vtt1(1) = V1; %denotes  $V_1^0$ 
    else
        xtt1(t) = xtt1(t-1) + B; %xtt1 denotes  $x_t^{(t-1)}$ ; Harvey 3.2.2a
        Vtt1(t) = Vtt1(t-1) + Q; %Harvey 3.2.2b
    end
    Kt = Vtt1(t)/(Vtt1(t)+R);
    Ft(t) = Vtt1(t)+R;
    vt(t) = y(t)-xtt1(t);
    xtt(t) = xtt1(t) + Kt*(y(t) - xtt1(t)); %Harvey 3.2.3a
    Vtt(t) = Vtt1(t)-Kt*Vtt1(t); %Harvey 3.3.3b
end

%Calculate negative log likelihood
negloglik = (1/2)*sum(vt.^2./Ft) + (1/2)*sum(log(abs(Ft))) + (T/2)*log(2*pi);

```

Appendix B: Correction for inputs

Consider a population A whose stochastic dynamics due to reproduction and survival can be described by a Leslie matrix, \mathbf{A}_t , but that experiences external inputs of individuals into the population. Such a situation can arise, if the population is supplemented such as for management purposes or if it is adjacent to a source population that continually provides immigrants. If the input individuals reproduce and their offspring cannot be distinguished, their presence (particularly the presence of their offspring) masks the underlying dynamics of population A. Here I present a method for estimating population A's underlying growth rate, λ_A , if the inputs were halted.

Throughout this discussion, I refer to resident and input individuals at year t . Residents at year t refers all individuals minus the inputs during year t *only*; residents may include individuals that were externally input at an earlier year or that were born from parents which were externally input in previous years. Inputs at year t refer only to those individuals that are externally input into the population at year t . The method assumes that the following data or estimates thereof are available: 1) a census count, O_t ; this could be a total population count or an age- or stage-specific count (such as an egg survey or breeder index). Since residents and inputs are indistinguishable, this count is assumed to include a mixture of both types of individuals, 2) an estimate of the fraction of individuals in O_t that are year t inputs, and 3) an estimate of the ratio of residents to (residents + inputs) for all ages or stages in the population. Number 3 is not sufficient to give number 2 since the census count may be a combination of ages, such as a total population count, without age information.

With external inputs, \mathbf{E}_t , at year t into a resident population, \mathbf{N}_t , the population process is:

$$\mathbf{N}_{t+1} = \mathbf{A}_t (\mathbf{N}_t + \mathbf{E}_t) \quad [\text{B.1}]$$

where \mathbf{A}_t is the stochastic projection matrix for year t . The objective is to estimate the mean and variance of $\lambda_{A,t}$, the dominant eigenvalue of \mathbf{A}_t . \mathbf{N}_t is the age-specific vector of resident individuals in the population at year t and \mathbf{E}_t is the age-specific vector of individuals input into the population at year t :

$$\mathbf{N}_t = \begin{bmatrix} N_{1,t} \\ N_{2,t} \\ N_{3,t} \\ \dots \\ N_{m,t} \end{bmatrix}, \quad \mathbf{E}_t = \begin{bmatrix} E_{1,t} \\ E_{2,t} \\ E_{3,t} \\ \dots \\ E_{m,t} \end{bmatrix}, \quad [\text{B.2}]$$

The maximum age of individuals is m .

The vector \mathbf{E}_t can be defined in terms of \mathbf{N}_t by using the age-specific fraction of residents relative to inputs at year t :

$$r_{i,t} = N_{i,t} / (N_{i,t} + E_{i,t}). \quad [\text{B.3}]$$

We can then solve for \mathbf{E}_t in terms of \mathbf{N}_t :

$$\mathbf{E}_t = \begin{bmatrix} 1/r_{1,t} - 1 & 0 & 0 & \dots & 0 \\ 0 & 1/r_{2,t} - 1 & 0 & \dots & 0 \\ 0 & 0 & 1/r_{3,t} - 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1/r_{m,t} - 1 \end{bmatrix} \mathbf{N}_t = (\mathbf{B}_t - \mathbf{I})\mathbf{N}_t. \quad [\text{B.4}]$$

where \mathbf{I} is the identity matrix (all ones along the diagonal; zeros elsewhere) and \mathbf{B}_t is the matrix with $1/r_{i,t}$ on the diagonal. Combining Eqs. B.1 and B.4, we have

$$\mathbf{N}_{t+1} = \mathbf{A}_t (\mathbf{N}_t + (\mathbf{B}_t - \mathbf{I})\mathbf{N}_t) = \mathbf{A}_t \mathbf{B}_t \mathbf{N}_t = \mathbf{C}_t \mathbf{N}_t. \quad [\text{B.5}]$$

where \mathbf{B}_t is defined in Eq. B.4. Our censuses, combined with information on the fraction of new inputs in the census, give us an estimate $N_{i,t}$ or of $\sum_{i \in j} N_{i,t}$ if the census is a combination of ages.

We can then estimate $I_{C,t}$, the dominant eigenvalue of \mathbf{C}_t using the ratio of $N_{i,t+1}$ to $N_{i,t}$, or the sums thereof (estimation of $I_{C,t}$ from the census data is discussed below).

However, our goal is to estimate $I_{A,t}$, the dominate eigenvalue of \mathbf{A}_t , thus we need to understand the relationship between $I_{C,t}$, which we can estimate from our censuses and $I_{A,t}$, which we want to estimate. One strategy is to use net reproductive rates, R_0 , defined as the mean number of offspring produced by a female over her lifetime. There is an approximate relationship between the net reproductive rate and I : (Caswell 2000)

$$I^T \approx R_0 \text{ or } T \ln(I) \approx \ln(R_0) \quad [\text{B.6}]$$

where T is the mean generation time. Using this relationship, we can solve for the relationship between the observed $I_{C,t}$ and the unobserved $I_{A,t}$:

$$\begin{aligned} \ln(I_{C,t}) &= \frac{1}{T} \ln(\tilde{R}_{0,t}) \text{ and } \ln(I_{A,t}) = \frac{1}{T} \ln(R_{0,t}) \\ \ln(I_{A,t}) &= \frac{1}{T} \ln\left(\frac{R_{0,t}}{\tilde{R}_{0,t}}\right) + \ln(I_{C,t}) \end{aligned} \quad [\text{B.7}]$$

The net reproductive rate of matrix \mathbf{A}_t is $R_{0,t}$. This is the true net reproductive rate if inputs were not occurring. The apparent net reproductive rate from the observed time series is $\tilde{R}_{0,t}$. This is the net reproductive rate of matrix \mathbf{C}_t . The key is to estimate the ratio $R_{0,t} / \tilde{R}_{0,t}$ without knowing much about \mathbf{A}_t , in other words not knowing the survivorships, s_i 's, or fecundities, f_i 's.

Depending on the life history and ages at which inputs occur, there may be a simple $R_{0,t} / \tilde{R}_{0,t}$ relationship that is independent of \mathbf{A}_t ; at the minimum, one may be able to make a reasonable guess at $R_{0,t} / \tilde{R}_{0,t}$.

As an example, suppose that the form of \mathbf{A}_t is

$$\mathbf{A}_t = \begin{bmatrix} f_{1,t} & f_{2,t} & f_{3,t} & \dots & f_{m,t} \\ s_{2,t} & 0 & 0 & \dots & 0 \\ 0 & s_{3,t} & 0 & \dots & 0 \\ 0 & 0 & s_{4,t} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \quad [\text{B.8}]$$

The net reproductive rate of \mathbf{A}_t in Eq. B.8 is (with t subscripts dropped for simplification)

$$R_{0,t} = f_1 + s_2 f_2 + s_2 s_3 f_3 + \dots + s_2 s_3 \dots s_m f_m. \quad [\text{B.9}]$$

The net reproductive rate of the matrix $\mathbf{C}_t (= \mathbf{A}_t \mathbf{B}_t)$ as defined in Eq. B.4) would then be

$$\tilde{R}_{0,t} = \frac{f_1}{r_1} + \frac{s_2 f_2}{r_1 r_2} + \frac{s_2 s_3 f_3}{r_1 r_2 r_3} + \dots + \frac{s_2 s_3 \dots s_m f_m}{r_1 r_2 r_3 \dots r_m}. \quad [\text{B.10}]$$

If individuals enter the population at a single age k , which is at or before the age of first reproduction, then $R_{0,t} / \tilde{R}_{0,t} = r_{k,t}$ where $r_{k,t}$ is the fraction of resident individuals at age k when the inputs occur. If this is not the case, going through the exercise of specifying the form of \mathbf{A} , R_0 , and \tilde{R}_0 , even if the parameters are unknown, may still allow one to make an educated estimate of $R_{0,t} / \tilde{R}_{0,t}$. The reader is cautioned to go through this exercise rather than making a “seat of the pants” guess at $R_{0,t} / \tilde{R}_{0,t}$. The ratio $R_{0,t} / \tilde{R}_{0,t}$ is often larger than one would guess since the product of r_i rather than simply r_i appears in Eq. B.10.

The last step is to estimate $\mathbf{I}_{C,t}$ from the census data. At equilibrium, $\mathbf{I}_{C,t}$ equals the ratio of any element (or sum of elements) of \mathbf{N}_{t+1} by the corresponding element (or sum) of \mathbf{N}_t . Thus, $\mathbf{I}_{C,t}$ could be estimated a number of ways:

If the census, O_b , is of a single age class, k :

$$\hat{\mathbf{I}}_{C,t} = \frac{N_{i,t+1}}{N_{i,t}} = \frac{(1 - io_{t+1})O_{t+1}}{(1 - io_t)O_t} \text{ where } io_t \text{ is the fraction of year } t \text{ inputs in } O_t.$$

If the census, O_t , is comprised of multiple age classes, j :

$$\hat{I}_{C,t} = \frac{\sum_{i \in j} N_{i,t+1}}{\sum_{i \in j} N_{i,t}} = \frac{(1 - io_{t+1})O_{t+1}}{(1 - io_t)O_t} \text{ for } j = \text{some set of ages}$$

Another option for both these cases would be to use

$$\hat{I}_{Ct} = \frac{O_{t+1}}{O_t} \text{ rather than } \frac{(1 - io_{t+1})O_{t+1}}{(1 - io_t)O_t}$$

and assume that $\frac{io_{t+1}}{io_t} \approx 1$. If the fraction of inputs is changing over time, $\frac{io_{t+1}}{io_t} \neq 1$, and using

$\frac{O_{t+1}}{O_t}$ means that you will correspondingly under- or overestimate $I_{C,t}$. Still this may be a

justified approximation if your information on inputs, io_t , is rather poor.

With estimates of $R_{0,t} / \tilde{R}_{0,t}$ and $I_{C,t}$, the estimated mean and variance of $I_{A,t}$ are:

$$\begin{aligned} \hat{m}_A &= \text{mean of } \ln(\hat{I}_{A,t}) = \frac{1}{T} \ln \left(\frac{R_{0,A,t}}{R_{0,C,t}} \right) + \ln(\hat{I}_{C,t}) \text{ for all } t \\ \hat{s}_A^2 &= \text{variance of } \ln(\hat{I}_{A,t}) = \frac{1}{T} \ln \left(\frac{R_{0,A,t}}{R_{0,C,t}} \right) + \ln(\hat{I}_{C,t}) \text{ for all } t \end{aligned} \quad [\text{B.11}]$$

These are the maximum likelihood estimates. The example with Pacific salmon in the main text uses running sum and slope estimates instead to deal with extraneous variability in the census counts.

To summarize, the algorithm for input correction involves the following steps:

- 1) Estimating the fraction of individuals in each year's census that came from outside the population that year (this io_t).

- 2) Estimating for year t , what fraction of age i individuals were external inputs from the current year ($r_{i,t}$)
- 3) Estimating the form of \mathbf{A}_t . Estimating at what ages the inputs occur (\mathbf{E}_t). These two can be used to estimate the form of \mathbf{C}_t .
- 4) Using 1-3 above, to estimate or infer $R_{0,t} / \tilde{R}_{0,t}$.
- 5) Estimating $I_{C,t}$ from the census data.
- 6) Estimating the mean and variance of $I_{A,t}$ using $R_{0,t} / \tilde{R}_{0,t}$ and $I_{C,t}$.