Kalman filter

February 20, 2020

©2020 by Christopher A. Sims. This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

The Kalman Filter

Model in the form

$$\begin{aligned} \text{Plant equation :} \qquad & s_t = As_{t-1} + \varepsilon_t \\ \text{Measurement equation :} \qquad & y_t = Hs_t + \nu_t\,. \end{aligned}$$

$\mathrm{Var}(\varepsilon_t) = \Omega$, $\mathrm{Var}(\nu_t) = \Xi$. $\varepsilon_t \perp \nu_t$ and (ε_t, ν_t) i.i.d., independent of past y, s.

KF: A rule for starting with a prior $s_t \sim N(\mu_t, \Sigma_t)$, using it, plus observation of y_{t+1}, to update to a new distribution $s_{t+1} \sim N(\mu_{t+1}, \Sigma_{t+1})$.

Aside: Version with no observation error

We can relabel observation errors as elements of s, after which there are no errors in the observation equation.

$$u_t = \begin{bmatrix} s_t \\ \nu_t \end{bmatrix}$$

and then rewrite the model as

$$\begin{aligned} u_t &= \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} u_{t-1} + \begin{bmatrix} \varepsilon_t \\ \nu_t \end{bmatrix} \\ y_t &= \begin{bmatrix} H & I \end{bmatrix} u_t \,. \end{aligned}$$

The equations as an assertion about one-step-ahead conditional distribution

Assuming that information at time t, $\mathcal{I}_t$, gives us

$$s_t \sim N(\mu_t, \Sigma_t),$$

the equations imply

$$\left\{ \begin{bmatrix} s_{t+1} \\ y_{t+1} \end{bmatrix} \mid \mathcal{I}_t \right\} \sim N\left(\begin{bmatrix} A\mu_t \\ HA\mu_t \end{bmatrix}, \begin{bmatrix} A\Sigma_t A' + \Omega & A\Sigma_t A'H' + \Omega H' \\ HA\Sigma_t A' + H\Omega & HA\Sigma_t A'H' + H\Omega H' + \Xi \end{bmatrix} \right).$$

Linear Regression

Suppose we have a jointly normal random vector split into two pieces, X_1, X_2:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) .$$

Then familiar results about linear regression tell us that

$$\big[X_1 \mid X_2 \big] \sim N\big(\beta(X_2 - \mu_2) + \mu_1, \Omega \big)$$

$$\beta = \Sigma_{12}\Sigma_{22}^{-1}, \qquad \Omega = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} .$$

KF formulas

Applying the formula for the conditional distribution of one Gaussian random variable given another, we get

$$\{s_{t+1} \mid \mathcal{I}_{t+1}\} \sim N(\mu_{t+1}, \Sigma_{t+1})$$

$$\mu_{t+1} = A\mu_t + (A\Sigma_t A'H' + \Omega H')(HA\Sigma_t A'H' + H\Omega H' + \Xi)^{-1}(y_{t+1} - HA\mu_t)$$

$$\begin{aligned}\Sigma_{t+1} = {} & A\Sigma_t A' + \Omega \\ & - (A\Sigma_t A'H' + \Omega H')(HA\Sigma_t A'H' + H\Omega H' + \Xi)^{-1}(HA\Sigma_t A' + H\Omega)\end{aligned}$$

Note that, though this looks like messy algebra, if y_t is a scalar, there is no matrix inversion involved. There is a lot of experience in using this algorithm, so it worthwhile consulting numerical analysis literature or using an optimized program if you want to do this with large matrices or many times.

Likelihood

At each date, the Kalman filter involves forming a normal distribution for $y_{t+1} \mid \mathcal{I}_t$. Calling the pdf of this distribution $p(y_{t+1} \mid \mathcal{I}_t)$, the pdf of the entire observed sample of y's is then

$$\prod_{t=1}^{T} p(y_t \mid \mathcal{I}_{t-1})$$

This formula applies because we assume the information available at t consists of the time zero information $\mathcal{I}_0$ plus the sequence of y_s values for $s \leq t$. The Kalman filter only tells us how to derive $p(\cdot \mid \mathcal{I}_{t+1})$ from $p(\cdot \mid \mathcal{I}_t)$ and y_t. The initial distribution $p(\cdot \mid \mathcal{I}_0)$ is determined by an initial Gaussian prior on the initial state s_0.

The log posterior density (often imprecisely called the log likelihood, despite the fact that it involves a prior density) is then just the sum of the $\log(p(y_t \mid \mathcal{I}_t))$ terms. A single one of those terms is

$$-\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}(y_t - \hat{y}_t)'\Phi_t^{-1}(y_t - \hat{y}_t) - \tfrac{1}{2}\log|\Phi_t| \ ,$$

where n is the dimension of y and $\hat{y}_t = HA\mu_{t-1}$ and $\Phi_t = HA\Sigma_{t-1}A'H' + H\Omega H' + \Xi$ are the mean and variance matrix of the one-step-ahead distribution for y_t. Since these quantities are computed as part of the KF, the log likelihood element, or the two pieces of it separately, are usually provided, along with the filtered μ_t, Σ_t, as part of the results of the filter.

The KF assumes that A, Ω, H, and Ξ are known quantities, while in applications they usually are not known. In applications in econometrics usually these parameters of the KF are specified as functions of some other underlying parameters, and the KF is executed to evaluate the posterior density at many values of the underlying parameters, either as part of a maximization routine or as part of a scheme for exploring the shape of the posterior density.

Since the KF assumes A, Ω, H, and Ξ are known, and since the KF operates one date at a time, it can handle time subscripts on all these parameters. Of course if we have to estimate them, we don't want underlying parameters to be changing freely at every date, so the fact that the KF allows this is not of much help. However it is quite common in applications for H_t to be an observable matrix of exogenous variables that changes with t.

The Kalman smoother

- The Kalman filter: $s_t \mid \mathcal{I}_t$ for each t.
- Sometimes (e.g. guiding a spacecraft) this is what we need.
- In economics often we have a fixed data set for $t = 1, \ldots, T$ and would like to know: $s_t \mid \mathcal{I}_T$ for each t.
- With the results of the KF in hand, we can find these distributions recursively, by an algorithm much like the KF, that works backwards from the end of the sample.

The smoother: details

- The smoother at each t uses the distribution of $s_{t+1} \mid \mathcal{I}_T$ and that of $s_t \mid \mathcal{I}_t$, to deliver that of $s_t \mid \mathcal{I}_T$.

- Initialize using the fact that the KF itself gives us $s_T \mid \mathcal{I}_T$ at the end of the sample.

- Then apply the filter recursively backward from the sample's end.

- For $t = 0$, we will have a distribution $s_0 \mid \mathcal{I}_{\mathcal{T}}$ that is generally very different from the prior $s_0 \mid \mathcal{I}_0$. Other observations for small t also may have smoothed distributions very different from the KF results, because for these observations the smoother uses much more information than the KF.

The smoother: algebra

First note the joint distribution, which we have used before,

$$\begin{bmatrix} s_{t+1} \\ s_t \end{bmatrix} \mid \mathcal{I}_t \sim N\left(\begin{bmatrix} A\mu_t \\ \mu_t \end{bmatrix}, \begin{bmatrix} A\Sigma_t A' + \Omega & A\Sigma_t \\ \Sigma_t A' & \Sigma_t \end{bmatrix} \right) . \tag{†}$$

From this we can see, by applying the formulas for normal conditional distributions, that

$$s_t = \mu_t + \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1}(s_{t+1} - A\mu_t) + \zeta_t \,, \tag{$*$}$$

where $\zeta_t \sim N(0, \Sigma_t - \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1} A\Sigma_t)$ and ζ_t is uncorrelated with the past observations on y that generate $\mathcal{I}_t$ and also with s_{t+1}.

Why ζ_t is uncorrelated with the entire future

- The fact that it is uncorrelated with s_{t+1} means that it is necessarily also uncorrelated with s_{t+v+1} for all $v > 0$, because…

- The plant equation can be solved recursively to tell us that $s_{t+v+1} = A^v s_{t+1} + \eta_{t+v+1}$, where η_{t+v+1} is a linear combination of the plant error terms ε_{t+1+u} for $u = 1, \ldots, v$.

- Since ζ_t is uncorrelated with s_{t+1}, and since ε_{t+u+1} is uncorrelated with any s_r or y_r for $r \leq t+1$, ζ_t, a function of s_t, s_{t+1} and y_v, $v \leq t$, is uncorrelated with s_{t+1+v} for $v \geq 1$.

The smoother: the formulas

- Everything on the right-hand side of $(*)$ is in $\mathcal{I}_t$ except s_{t+1} and ζ_t.

- Therefore to find the $s_t \mid \mathcal{I}_T$ distribution, we just use $(*)$, which defines s_t as a linear transformation of s_{t+1}, ζ_t. We use the notation $\{s_t \mid \mathcal{I}_T\} \sim N(m_t, S_t)$. The mean of ζ_t given $\mathcal{I}_T$ is zero, as we have noted, and the mean of s_{t+1} given $\mathcal{I}_T$ is m_{t+1}, so plugging in to $(*)$ gives us

$$m_t = \mu_t + \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1}(m_{t+1} - A\mu_t)\,.$$

For the conditional variance, we get two components, one from the variance S_{t+1} of $s_{t+1} \mid \mathcal{I}_T$, the other from the variance of ζ_t. Note that we have observed that these two components of s_t are uncorrelated. The variance is therefore

$$S_t = \Sigma_t - \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1} A\Sigma_t \\ + \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1} S_{t+1} (A\Sigma_t A' + \Omega)^{-1} A\Sigma_t \, .$$

Here as with the formulas for the filter, you don't need to commit the formulas to memory. You should understand how they are derived.

Interpreting filtered and smoothed estimates

- Even if the underlying state is not changing at all, the filtered estimate of it will generally change a lot toward the beginning of the sample.

- Even for the smoothed states, changes reflect both estimation error and actual movement in the state.

- Therefore, the fact that the filtered or smoothed *estimates* vary, even if the variation seems economically significant, should not be naively interpreted as implying s_t varies substantially over time.

- If you are trying to determine whether there is substantial time variation, plot smoothed estimates, with their associated error bands.

- Or, using methods we will discuss later, find posterior odds on the model with time variation vs. the model without time variation.

Sampling time paths of states

In MCMC sampling from the posterior density, it often is useful to sample from the time path of the unobservable states as part of the Markov Chain. This can be done via a backward recursion that is much like that for smoothing.

I. Start by drawing a value of s_T from the distribution of $s_T \mid \mathcal{I}_T$ that is available from the Kalman filter run.

II. For each $t < T$, with a draw s_{t+1} from $s_{t+1} \mid \mathcal{I}_T$ in hand, use the joint normal distribution for $(s_t, s_{t+1}) \mid \mathcal{I}_t$ given in (†) to make a draw from $s_t \mid \{\mathcal{I}_t, s_{t+1}\}$. Since all the influence of s_t on later y's and s's is through s_{t+1} this is in fact a draw from $s_t \mid \mathcal{I}_T$

This procedure generates a complete sequence of values for s_t, including for the initial conditions. Often the initial conditions have a diffuse prior distribution, but conditional on the data, they may nonetheless be sharply determined.

Application: index numbers

Suppose we have a collection of N price time series p_{it}, measured in logs, that we think are each made up of an unobservable general "price level" component and an idiosyncratic component that is independent of the general price level and of other idiosyncratic components. We would like to use them to estimate the general price level. The equations are

$$p_{it} = \alpha_i + \beta_i \bar{p}_t + \nu_{it} \tag{1}$$

$$\bar{p}_t = \gamma_0 + \theta_0 \bar{p}_{t-1} + \varepsilon_{0t} \tag{2}$$

$$\nu_{it} = \gamma_i + \theta_i \nu_{i,t-1} + \varepsilon_{it}\,, i = 1, \dots, N\,. \tag{3}$$

$$\varepsilon_{it} \sim N(0, \sigma_i^2), \text{independent across } t, i\,. \tag{4}$$

Index numbers: setting it up in KF form

$$s_t : \qquad \begin{cases} (\bar{p}_t, v_{it},\, i = 1, \dots, N,\, \mu_t) \text{ or} \\ \big(\bar{p}_t, (v_{it}, \alpha_{it}, \gamma_{it}, i = 1, \dots, N)\big) \end{cases}$$

The first version will have the α's and γ's coefficients of the constant $\mu_t \equiv 1$. The second version instead exploits the KF's ability to deliver posterior means for these constant term parameters, conditional on the other parameters, and thereby makes the dimension of an iterative posterior density maximization problem smaller.

The measurement equation is (1). The plant equation consists of (2)-(3), plus either a single equation stating $\mu_t = \mu_{t-1}$, with the prior on μ_0 degenerate at $\mu_0 = 1$, or else 2N+1 equations of the form $\alpha_{it} = \alpha_{i,t-1}$ and $\gamma_{it} = \gamma_{i,t-1}$.

Index number practice exercise:

Figure out what the A, H, Ω and Ξ matrices are in this problem, for both ways of treating the constant terms.

Application: Time-varying parameter regression

The model is

$$y_t = X_t\beta_t + \varepsilon_t$$
$$\beta_t = A\beta_{t-1} + \nu_t\,.$$

The error terms ε_t and ν_t are uncorrelated across equations and across time and are uncorrelated with X_t and with any lagged variables. Here the state is β_t, the first equation is the observation equation, and the second equation is the plant equation. We assume y_t and X_t are observable, β_t and the error terms are not.

Time varying parameters practice exercise:

Figure out what the A, H, Ω and Ξ matrices are in this problem. For the validity of the Kalman Filter, does it matter whether the X's are strictly exogenous or instead predetermined? [Reminder: For strictly exogenous X's, X_t and ε_s are uncorrelated for all t, s combinations, while for predetermined X's X_t and ε_s are uncorrelated for $s \geq t$, but not necessarily for other t, s pairs. Predetermined X's can be lagged y's.]

Application: A finite order MA model

Suppose $y_t = a_0\varepsilon_t + a_1\varepsilon_{t-1}$ with ε_t i.i.d. $N(0,1)$. If we let $[\varepsilon_t, \varepsilon_{t-1}]$ be the state vector, Then this equation becomes the observation equation and the plant equation is

$$\begin{bmatrix}\varepsilon_t\\ \varepsilon_{t-1}\end{bmatrix} = \begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}\begin{bmatrix}\varepsilon_{t-1}\\ \varepsilon_{t-2}\end{bmatrix} + \begin{bmatrix}\nu_t\\ 0\end{bmatrix}.$$

Here of course ν_t is the same thing as ε_t; we only distinguish them to make the notation line up with that of the KF.

Practice exercise: Define the state and set up the KF plant and observation equations for an ARMA(1,1) model (i.e. a model of the form $B(L)y_t = A(L)\varepsilon_t$, with B and A both first-order polynomials). Can we treat any of the parameters in this model as part of the state?