KALMAN FILTER

1. THE KALMAN FILTER

Model in the form

Plant equation: $s_t = As_{t-1} + \varepsilon_t$

Measurement equation: $y_t = Hs_t + \nu_t$.

$\text{Var}(\varepsilon_t) = \Omega$, $\text{Var}(\nu_t) = \Xi$. $\varepsilon_t \perp \nu_t$ 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 $s_{t+1} \sim N(\mu_{t+1}, \Sigma_{t+1})$.

2. 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.

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

and then rewrite the model as

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

3. 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

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

Date: September 29, 2008.

4. 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

$$\begin{aligned} \begin{bmatrix} X_1 \mid X_2 \end{bmatrix} &\sim N\big(\beta(X_2 - \mu_2) + \mu_1, \Omega\big) \\ \beta = \Sigma_{12}\Sigma_{22}^{-1}, &\qquad \Omega = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} . \end{aligned}$$

5. KF formulas

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

$$\begin{aligned} \{s_{t+1} \mid \mathcal{I}_{t+1}\} &\sim N(\mu_{t+1}, \Sigma_{t+1}) \\ \mu_{t+1} &= A\mu_t + (A\Sigma_t A'H' + \Omega H')(HA\Sigma_t A'H' + H\Omega H' + \Xi)^{-1}(y_{t+1} - HA\mu_t) \\ \Sigma_{t+1} &= A\Sigma_t A' + \Omega \\ &\quad - (A\Sigma_t A'H' + \Omega H')(HA\Sigma_t A'H' + H\Omega H' + \Xi)^{-1}(HA\Sigma_t A' + H\Omega) \end{aligned}$$

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.

6. 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

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

where $\hat{y}_t = HA\mu_{t-1}$ 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.

7. THE KALMAN SMOOTHER

The Kalman filter gives us the distribution for $s_t \mid \mathcal{I}_t$ for each t. If we are making a decision in real time at each t, this is exactly what we need. But in economic applications we are often looking at historical data to gain insight about parameter values and about past events. For these purposes we would like to use the full sample information, i.e. to find the distribution of $s_t \mid \mathcal{I}_T$, where T is the end of the sample. With the Kalman filter results in hand, this can be done recursively. Note that the distribution of $s_T \mid \mathcal{I}_T$ is delivered by the Kalman filter. 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 $s_t \mid \mathcal{I}_T$. Obviously this can be applied recursively, starting at the end of the sample, to deliver the whole sequence of desired distributions. Note that this includes generating a posterior distribution for $s_0 \mid \mathcal{I}_T$ that will be different from the prior on s_0.

To derive the smoother, note first the joint distribution

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

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

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

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}. But the fact that it is uncorrelated with s_{t+1} means that it is necessarily also uncorrelated with s_{t+v+1} for all $v > 0$. This follows from the fact that 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 $v \geq 1$.

With this reasoning in hand, we can see that all we need do now is replace s_{t+1} in $(*)$ by its conditional distribution given $\mathcal{I}_T$, since everything else on the right-hand side is in $\mathcal{I}_t$. If we use the notation $\{s_t \mid \mathcal{I}_T\} \sim N(m_t, S_t)$, we therefore have

$$\begin{aligned} m_t &= \mu_t + \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1}(m_{t+1} - A\mu_t) \\ S_t &= \Sigma_t - \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1} A\Sigma_t \\ &\quad + \Sigma_t A'(A\Sigma_t A' + \Omega)^{-1} S_{t+1} (A\Sigma_t A' + \Omega)^{-1} A\Sigma_t . \end{aligned}$$

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.

8. Filtering vs. smoothing in practice

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. Usually the prior is not very informative, so the distribution for the state changes strongly in response to the arrival of the initial data, then settles down a the sample lengthens. If this is happening, the smoothed estimates of the state will not show the rapid changes at the start of the sample, since they are using the full information set to estimate the state at every date.

If changes in s_t are the focus of interest, therefore, it is very important to use smoothed estimates of the state in interpreting changes in the estimated state. Also, even for the smoothed states, changes reflect both estimation error and actual movement in the state. Plots of smoothed estimates of the state should show standard error bands around the estimates, to give an idea of how much of the plotted movement might be due to estimation error.

9. 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

$$\begin{aligned} p_{it} &= \alpha_i + \beta_i \bar{p}_t + \nu_{it} \\ \bar{p}_t &= \gamma_0 + \theta_0 \bar{p}_{t-1} + \varepsilon_{0t} \\ \nu_{it} &= \gamma_i + \theta_i p_{i,t-1} + \nu_{i,t-1} \end{aligned}$$

ε parameters are i.i.d. across time. They are independent across equations and independent of all lagged variables. Their variances, σ_i^2 , may differ across equations.

To set this up as a KF problem, take the state to be $\bar{p}_t$ together with ν_{it}, $i = 0, \ldots, N$ and a constant. We can either use a single constant that is always 1, or else treat the α and γ coefficients as part of the state. The former keeps the KF simpler. The latter 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 the first listed. The plant equation consists of the next two, plus either a single equation stating $u_t = u_{t-1}$, with the prior on u_0 degenerate at $u_0 = 1$, or else 2N+1 equations of the form $\alpha_{it} - \alpha_{i,t-1}$ and $\gamma_{it} = \gamma_{i,t-1}$.

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

10. 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.

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.]

11. 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?