Kalman Filter

Matthew Cocci

December 30, 2013

1 Basic Idea and Terminology

Here's the basic procedure associated with the Kalman Filter:

1. Start with a prior for some variable of interest in the current period, $p(x)$.
2. Observe the current measurement y_t.
3. "Filter" out the noise and compute the filtering distribution: $p_t(x|y)$.
4. Compute predictive distribution $p_{t+1}(x)$ from filtering distribution and your model.
5. Increment t and return to step 1, taking the predictive distribution as your prior.

2 Filtering Step

Suppose we want to measure some latent state variable x. We will assume a *prior* that is multivariate normal such that

$$x \sim \mathrm{N}(\hat{x}, \Sigma)$$

Next, we "measure" x by matching it to an observable in a *measurement equation*:

$$y = Gx + v \qquad v \sim \mathrm{N}(0, R)$$

where R is positive definite, while G and R are both 2×2. This forms the *likelihood*.

We then "filter" out the noise, updating our view of x in light of the data in the filtering step using Bayes' Rule. Note, this is called "filtering" because we don't use the prior and likelihood to forecast into the future. We combine the prior with the likelihood only to filter out noise and get closer to the true value of x based on the data, summarized in the posterior, or the *filtering distribution*:

$$\begin{aligned} p(x \mid y) &= \frac{p(y \mid x) \cdot p(x)}{p(y)} \propto p(y \mid x) \cdot p(x) \\ &\propto \exp\left\{-\frac{1}{2}(y - Gx)' R^{-1} (y - Gx)\right\} \exp\left\{-\frac{1}{2}(x - \hat{x})' \Sigma^{-1} (x - \hat{x})\right\} \end{aligned} \tag{1}$$

Now let's expand the term in the lefthand exponential:

$$\begin{aligned} A = (y - Gx)' R^{-1} (y - Gx) &= (y' - x'G') R^{-1} (y - Gx) = \left(y'R^{-1} - x'G'R^{-1}\right)(y - Gx) \\ &= \left(y'R^{-1}y - y'R^{-1}Gx - x'G'R^{-1}y + x'G'R^{-1}Gx\right) \end{aligned}$$

And now the same for the righthand exponential:

$$\begin{aligned} B = (x - \hat{x})'\Sigma^{-1}(x - \hat{x}) &= (x' - \hat{x}')\Sigma^{-1}(x - \hat{x}) \\ &= (x'\Sigma^{-1} - \hat{x}'\Sigma^{-1})(x - \hat{x}) \\ &= x'\Sigma^{-1}x - x'\Sigma^{-1}\hat{x} - \hat{x}'\Sigma^{-1}x + \hat{x}'\Sigma^{-1}\hat{x} \end{aligned} \tag{2}$$

Adding the two exponentials, we get:

$$\begin{aligned} C = A + B = x' \left(\Sigma^{-1} + G'R^{-1}G\right) x - x'(\Sigma^{-1}\hat{x} + G'R^{-1}y) - (\hat{x}'\Sigma^{-1} + y'R^{-1}G)x \\ + \hat{x}'\Sigma^{-1}\hat{x} + y'R^{-1}y \end{aligned}$$

Now notice that Expression 1 is the probability distribution of x *conditional* on y and pretty much anything else that isn't x. And because of the wonderful properties of the exponential function and the black-hole powers of the proportionality constant, we'll be able to simplify things nicely (and we'll worry that the distribution $p(x|y)$ integrates to one later on).

So in the expression for C, the two terms in the second row *don't* depend upon x. Therefore, letting $C(x)$ be the portion of C that depends upon x, and letting $C(\neg x)$ bet the additive terms which don't depend upon x, we can simplify

$$\begin{aligned} p(x \mid y) \propto \exp\left\{-\frac{1}{2}C\right\} &= \exp\left\{-\frac{1}{2}\left[C(x) + C(\neg x)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}C(x)\right\} + \exp\left\{-\frac{1}{2}C(\neg x)\right\} \\ &\propto \exp\left\{-\frac{1}{2}C(x)\right\} \end{aligned}$$

We just absorb the portion not relevant to $p(x|y)$ into the proportionality constant. This means our the work we did above to get C simplifies our target expression to

$$p(x \mid y) \propto \exp\left\{-\frac{1}{2}\left[x' \left(\Sigma^{-1} + G'R^{-1}G\right) x - x'(\Sigma^{-1}\hat{x} + G'R^{-1}y) - (\hat{x}'\Sigma^{-1} + y'R^{-1}G)x\right]\right\} \tag{3}$$

Goal Now this doesn't look too helpful, but with a little bit of work, we can turn this into the probability distribution for a multivariate normal random variable. In fact, the rest of the section may look complicated, but keep in mind the big picture: the likelihood and prior were both multivariate normal, so the posterior $p(x|y)$ is going to be normal. We just want to identify the mean vector and variance-covariance matrix; then we're home.

Variance So first, the variance of the normal distribution corresponding to $p(x|y)$ can be derived by examining Equation 3 and likening it to Equation 2 (which gives the contents of the exponential in the prior MVN distribution of x).

Namely, the inverse of the new variance, which we'll denote as Σ^F will be sandwiched in between x' and x in Equation 3, just as it was sandwiched between x' and x in Equation 2. We use this fact, along with the the Woodbury matrix identity (stated in the appendix) to derive:

$$\begin{aligned} \Sigma^F &= \left(\Sigma^{-1} + G'R^{-1}G\right)^{-1} \\ \text{Woodbury Identity} \Rightarrow \quad &= \Sigma - \Sigma G'(R + G\Sigma G')^{-1}G\Sigma \end{aligned} \tag{4}$$

Mean Next, we want to get the mean of the distribution of $p(x|y)$, which we'll denote by $\hat{x}^F$. Again, once we take a second and compare Expression 3 to Expression 2, it's becomes clear from inspection that we must have

$$(\Sigma^{-1}\hat{x} + G'R^{-1}y) = \left(\Sigma^{-1} + G'R^{-1}G\right) Z \tag{5}$$

To see this, liken the lefthand side of Equation 5 (which itself comes from Expression 3) to the result of the matrix multiplication $\Sigma^{-1}\hat{x}$ in Equation 2. To get the righthand side, use the fact that we *know* the Equation 5 analogue to Equation 2's Σ^{-1}, which we just derived in the variance section and called Σ^F.

So all that's left to do is solve for Z in Equation 5. The result will turn out to be our mean vector for the posterior, $\hat{x}_F$:

$$\hat{x}^F = Z = \hat{x} + \left[\Sigma G'(R + G\Sigma G')^{-1}\right](y - G\hat{x}) \tag{6}$$

If you want to see the nasty linear algebra that gets you to this result, you can check out the appendix. Or you can you just take this result as given and save yourself an hour of painstaking derivation and eye-crossing complications, unlike myself.[1]

Putting together the expressions for the mean and variance (see Equations 6 and 4, respectively) of the posterior estimate for $x|y$ (i.e. the "filtering distribution"), we get that

$$x|y \sim \mathrm{N}\left(\hat{x}^F,\ \Sigma^F\right) \tag{7}$$

$$\begin{aligned} \text{where} \quad \hat{x}^F &= \hat{x} + \left[\Sigma G'(R + G\Sigma G')^{-1}\right](y - G\hat{x}) \\ \Sigma^F &= \Sigma - \Sigma G'(R + G\Sigma G')^{-1}G\Sigma \end{aligned}$$

Notice that our new "filtered" mean is simply a combination of our prior mean, $\hat{x}$, and a transformation of the "error" between our observed value and the prior guess for that observable $(y - G\hat{x})$.

[1] But if you *do* look at the appendix, you might just give my semi-wasted hour some meaning, in which case—thank you.

Recap Okay, so what did we just do?

1. We took a Multivariate Normal (MVN) prior to summarize our beliefs about a latent, imperfectly observable state variable x.
2. Knowing that we'll observe some data, y, which provides a "noisy" measure of x, we postulated a likelihood $p(y|x)$ that is also MVN.
3. Then, using Bayes' Rule, we combine the information contained in our prior $p(x)$ and the data (via the likelihhod $p(y|x)$) to get a "filtered" distribution of x, $p(x|y)$, given the data and our prior.

Why might this long, tortuous, painful process help us in economics? Well, imagine that in our model, there's some state for the "natural rate of unemployment," denoted by x. Now of course, we can't observe that value. But we'll have economic statistics, like measurements of unemployment itself along with other informative statistics such as GDP and hours, which might provide information about the natural rate of unemployment. However, those statistics are imperfect and noisy. The Kalman Filter gives us a way to combine those noisy estimates in with our beliefs in a principled, sensible manner.

3 Forecasting Step

Now let's make our model a little more dynamic and consider forecasting ahead. To do so, we specify a model of how the state, x, evolves. To make it easy on ourselves, let's assume everything's Gaussian (woohoo! that's easy):

$$x_{t+1} = Ax_t + w_{t+1} \qquad w_t \sim \mathrm{N}(0, Q) \tag{8}$$

Now, we want to come up with a *predictive distribution* given our prior and the current information encapsulated in our filtering distribution, $p(x|y)$. Since we're assuming everything is normal, we need only pin down the mean and variance of the forecast, since linear combinations of Gaussian variables are Gaussian.

Of course, these kinds of things are well known for MVN random variables, which has the nice properties

$$\begin{aligned} E[AX] &= AE[X] = A\mu \\ \mathrm{Var}(AX) &= A\mathrm{Var}(X)A' = A\Sigma A' \\ \text{where} \quad X &\sim \mathrm{N}(\mu, \Sigma) \end{aligned}$$

Now let's use these facts, along with the assumption that we're predicting x_{t+1} by starting with a *filtered* x_t (denoted x_t^F which has the distribution in Equation 7), and assuming x_t^F is uncorrelated with w_{t+1}:

$$\begin{aligned} E[x_{t+1}] &= E\left[Ax_t^F + w_{t+1}\right] = AE\left[x_t^F\right] + E\left[w_{t+1}\right] \\ &= A\hat{x}_t^F + 0 = A\hat{x}_t^F \qquad (9) \\ \mathrm{Var}(x_{t+1}) &= \mathrm{Var}(Ax_t^F + w_{t+1}) = A\mathrm{Var}\left(x_t^F\right)A' + \mathrm{Var}(w_{t+1}) \\ &= A\Sigma_t^F A' + Q \qquad (10) \end{aligned}$$

where $\hat{x}_t^F$ and Σ_t^F are as above in Equation 7. This characterizes the distribution for the one step ahead forecasting distribution.

Now, we can simplify what we have a bit more by defining the *Kalman Gain*:

$$K_\Sigma = A\Sigma G'(R + G\Sigma G')^{-1} \tag{11}$$

We see the practical use of this by subbing in the full expressions for $\hat{x}_t^F$ and Σ_t^F into Equations 9 and 10 above, and then simplifying our expressions for the mean and variance as a function of the Kalman Gain:

$$\begin{aligned}
\hat{x}_{t+1} &= E[x_{t+1}] = A\left\{\hat{x}_t + \left[\Sigma_t G'(R + G\Sigma_t G')^{-1}\right](y - G\hat{x}_t)\right\} \\
&= A\hat{x}_t + K_{\Sigma_t}(y - G\hat{x}_t) \qquad (12) \\
\Sigma_{t+1} &= \mathrm{Var}(x_{t+1}) = A\left\{\Sigma_t - \Sigma_t G'(R + G\Sigma_t G')^{-1} G\Sigma_t\right\} A' + Q \\
&= A\Sigma_t A' - K_{\Sigma_t} G\Sigma_t A' + Q \qquad (13)
\end{aligned}$$

4 Full Recursive Procedure

Now we need to build up the recursive algorithm to forecast further into the future. Let t be the current time period, and we proceed as follows:

1. Start with a prior in the current period at time t, as given above
$$p_t(x) \sim \mathrm{N}\left(\hat{x}_t,\ \Sigma_t\right)$$
2. Observe the current measurement, y_t.
3. Update and filter out the noise as we did in the Filtering Section to get, you guessed it, the filtering distribution: $p_t(x|y) = \mathrm{N}(\hat{x}_t^F, \Sigma_t^F)$.
4. Compute the predictive distribution $p_{t+1}(x) = \mathrm{N}(\hat{x}_{t+1}, \Sigma_{t+1})$ from the filtering distribution, $p_t(x|y)$, and the law of motion in Equation 8.
5. Increment t. Return to step 1, taking the predictive distribution, $p_{t+1}(x)$, as the prior.

5 Convergence

Now since x_t is random from the perspective of time $t-1$ (i.e. there is some irreducable uncertainty resulting from shocks at time t), we know that Σ_t will never be zero, unless w_t in Equation 8 is degenerate. However, we might ask whether Σ_t, our measure of uncertainty for our prediction $\hat{x}_t$ of x_t, will ever converge to a *constant* matrix over time.

Recall how Σ_t evolves, as specified in Equation 13 (substituting in for the Kalman Gain, K_Σ), which gives the non-linear difference equation:

$$\Sigma_{t+1} = A\Sigma_t A' - K_{\Sigma_t} G\Sigma_t A' + Q \tag{14}$$

If it were the case that Σ_t converges to some fixed matrix, then there would be a fixed point, Σ^*, satisfying Equation 14 as follows:

$$\Sigma^* = A\Sigma^* A' - A\Sigma^* G'(R + G\Sigma^* G')^{-1} G\Sigma^* A' + Q \tag{15}$$

This is known as the *Discrete Time Algebraic Riccati Equation*. A sufficient condition for convergence is that all the eigenvalues of A, λ_i, satisfy $|\lambda_i| < 1$.

A Woodbury Matrix Identity

For matrices A, U, C, and V:

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1} \tag{16}$$

Now consider the special case we have above with the Kalman filter:

$$(A+V'CV)^{-1} = A^{-1} - A^{-1}V'(C^{-1}+VA^{-1}V')^{-1}VA^{-1} \tag{17}$$

B Derivation of the Mean

Recall what we want to show. For $\hat{x}_F \equiv Z$, we want to show that

$$\begin{aligned}(\Sigma^{-1}\hat{x} + G'R^{-1}y) &= \left(\Sigma^{-1} + G'R^{-1}G\right)Z \\ \Rightarrow \hat{x}_F = Z &= \hat{x} + \left[\Sigma G'(R+G\Sigma G')^{-1}\right](y - G\hat{x})\end{aligned}$$

And so we solve this equation by using the Woodbury matrix identity representation from above:

$$\begin{aligned}(\Sigma^{-1}\hat{x} + G'R^{-1}y) &= \left(\Sigma^{-1} + G'R^{-1}G\right)Z \\ \Rightarrow \quad Z &= \left(\Sigma^{-1} + G'R^{-1}G\right)^{-1}(\Sigma^{-1}\hat{x} + G'R^{-1}y) \\ Z &= \left(\Sigma - \Sigma G'(R+G\Sigma G')^{-1}G\Sigma\right)\left(\Sigma^{-1}\hat{x} + G'R^{-1}y\right)\end{aligned}$$

Now let's simplify $\hat{x}_F = Z$ a bit, expanding out the multiplication:

$$\begin{aligned}\hat{x}_F = Z &= \left(\Sigma - \Sigma G'(R+G\Sigma G')^{-1}G\Sigma\right)\left(\Sigma^{-1}\hat{x} + G'R^{-1}y\right) \\ \text{FOIL} \quad &= \hat{x} + \Sigma G'R^{-1}y - \left[\Sigma G'(R+G\Sigma G')^{-1}G\Sigma\right]\left[\Sigma^{-1}\hat{x}\right] \\ &\quad - \left[\Sigma G'(R+G\Sigma G')^{-1}G\Sigma\right]\left[G'R^{-1}y\right] \\ \text{Simpify} \quad &= \hat{x} + \Sigma G'R^{-1}y - \left[\Sigma G'(R+G\Sigma G')^{-1}\right](G\hat{x}) \\ &\quad - \Sigma G'(R+G\Sigma G')^{-1}G\Sigma G'R^{-1}y \\ \text{Change Order} \quad &= \hat{x} - \left[\Sigma G'(R+G\Sigma G')^{-1}\right](G\hat{x}) \\ &\quad + \Sigma G'R^{-1}y - \Sigma G'(R+G\Sigma G')^{-1}G\Sigma G'R^{-1}y \\ \text{Regroup} \quad &= \hat{x} - \left[\Sigma G'(R+G\Sigma G')^{-1}\right](G\hat{x}) \\ &\quad + \Sigma G'\left\{R^{-1} - (R+G\Sigma G')^{-1}G\Sigma G'R^{-1}\right\}y\end{aligned}$$

Okay, now let's take a breather. We'll make this a bit easier on ourselves, and just consider simplifying the guy in the brackets, {} by using $A^{-1}A = I$ with a very special choice of A:

$$\begin{aligned}\left\{R^{-1} - (R+G\Sigma G')^{-1}G\Sigma G'R^{-1}\right\} &= (R+G\Sigma G')^{-1}(R+G\Sigma G')R^{-1} \\ &\quad - (R+G\Sigma G')^{-1}G\Sigma G'R^{-1} \\ \text{Group} \quad &= (R+G\Sigma G')^{-1}\left[(R+G\Sigma G')R^{-1} - G\Sigma G'R^{-1}\right] \\ \text{Distribute} \quad &= (R+G\Sigma G')^{-1}\left[RR^{-1} + G\Sigma G'R^{-1} - G\Sigma G'R^{-1}\right] \\ \text{Simplify} \quad &= (R+G\Sigma G')^{-1}\left[I + 0\right] \\ &= (R+G\Sigma G')^{-1}\end{aligned}$$

Substituting back in above for the term in braces, {}, we get the following expression for $Z = \hat{x}_F$:

$$\begin{aligned} \hat{x}_F = Z &= \hat{x} - \left[\Sigma G'(R + G\Sigma G')^{-1}\right](G\hat{x}) + \Sigma G'(R + G\Sigma G')^{-1} y \\ &= \hat{x} + \left[\Sigma G'(R + G\Sigma G')^{-1}\right](y - G\hat{x}) \end{aligned} \tag{18}$$