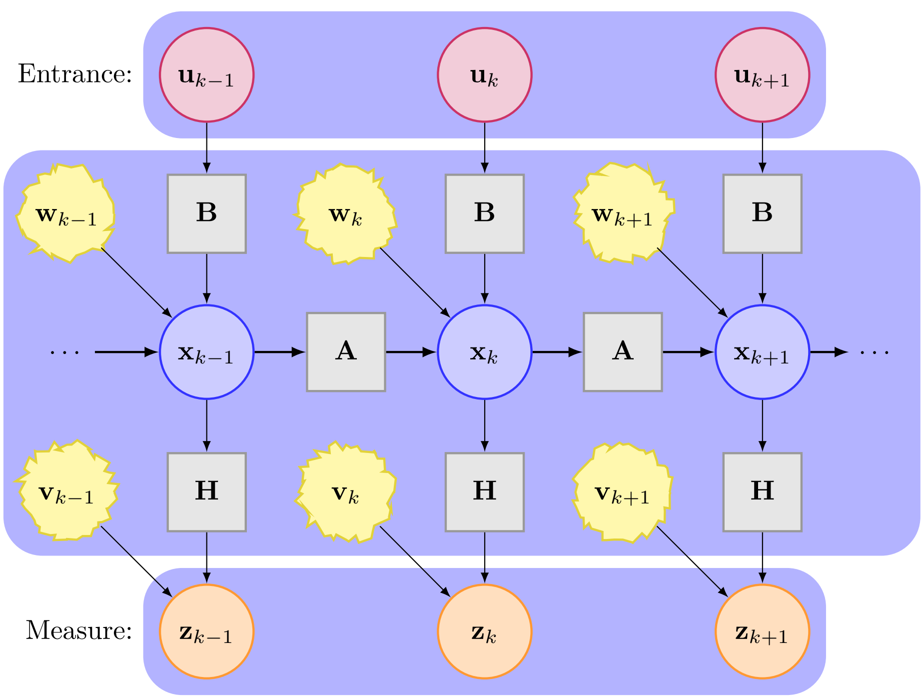
The Intuition of Kalman Filter

Weijie Chen

Department of Political and Economic Studies

University of Helsinki

August 20, 2011



Abstract

This tutorial helps you grasp the core idea of Kalman filter intuitively. This note is based on Maybeck (1979) and some graphs are borrowed from the book.

1 Introduction

Before wetting feet in daunting mathematics, we could have some casual talk about the history of Kalman filter. Last century, from 1950s to 1960s, the Air Force Office of Scientific Research sponsored some far-reaching researches on control theory applied in high-speed aircrafts (such fighters, bombers), aerospace vehicles. One of projects were leading by Dr. Rudolph Kalman, who is the main contributor of *Kalman filter*, thus entitled after his name. Later on, NASA used Kalman filter to calculate the satellite orbits, and more significant applications were on Ranger, Mariner, and Apollo missions of 1960s. Due to its outstanding performance on engineering, the Kalman filter started to be extensively made use of in missile tracking system, sonar ranging, Global Positioning System (GPS) and etc..

The essential idea of Kalman filtering is 'weighting'. Suppose you have two measurements, x_1 and x_2, you are seeking the best estimate of x based on these two measurements. So the question is how much *weight* you intend to put on either of them. This is a straightforward question, we prefer to putting more weight on the measurement with smaller variance. Thus we can easily set up a formula,

$$\hat{x} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} x_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} x_2 \tag{1}$$

If σ_2^2 is larger than σ_1^2, naturally we shall weight more on x_1 since σ_1^2 is smaller. To illustrate how this works, we can use a simple weather science example to give intuitions.

We launched an airship 1[1] into sky to collect data of weather, the sensors

[1] Picture borrowed from *The Fifth Column* bulletin.

Figure 1: Airship

Figure 2: Weather sensor

2[2] installed on the spaceship will transmit a vector of data x_1, elements such as wind speed, wind direction, humidity, temperature, coordinates, etc., to computer system, and vector x_2 is the last forecast of the state of atmosphere. In real rocket science or meteorology, the dimension of state vectors x_2 can be as large as millions. Then we naturally use covariance matrices to replace σ_1^2 and σ_2^2. So the job of Kalman filter is to find the optimal weight between the forecast and sensor data at each time period, say every 10 seconds.

1.1 Insight of Kalman Filter

The rest of this section will give more insight and intuition about Kalman filter, this is extremely important to understand the mathematical derivation later on.

As we can see from the meteorology example above, Kalman filter is an algorithm to decide weight on sensor data and last forecast at each time instance. In general, Kalman filter will absorb all information, such as data from sensors and forecast results, to generate the an overall best estimates. To give another concrete example, to measure the velocity of an vehicle, we have several ways: using Doppler radar, inertial navigation system, pitot-static tube and wind flow in air data system, all these data will be made use of, regardless of their precision. The Kalman filter will assess all information you have in hand and make the overall best estimate.

But one thing to notice, since Kalman filter is a *recursive* algorithm,

[2] Picture borrowed from http://www.zeework.co.uk/pb100.htm, a website of supplying hydrogenomatics equipment.

which does not process all previous data together when a new measurement comes. In the long run, Kalman filter stands out among all available filters, we could say Kalman filter is an optimal filter.

2 Three Assumptions

To formulate Kalman filter, there are three assumptions we need: linearity, whiteness, Gaussian.

Linearity

Linear models, compare with nonlinear counterparts, are usually the priorities we would like to work on. Even if we don't have a linear model, we can always linearise the nonlinear one around some fixed point, such as linearising a nonlinear differential equations around its steady-state value. For reasons of doing this, first, we have a full set of tools to handle linear dynamic systems, such as differential and difference equations; second, they are easier to be handled by computer than nonlinear ones.

Whiteness

'Whiteness' means error term follows a white noise process, which is independent of time. The Fourier transform of white noise process shows that the noise has equal power at all frequencies, which does not exist in real life. The reason we need to use white noise is mainly from a physical point of view, see figure 3. 'Bandpass' is a frequency range which a certain physical system can respond, for instance, human ears cannot hear the sound with

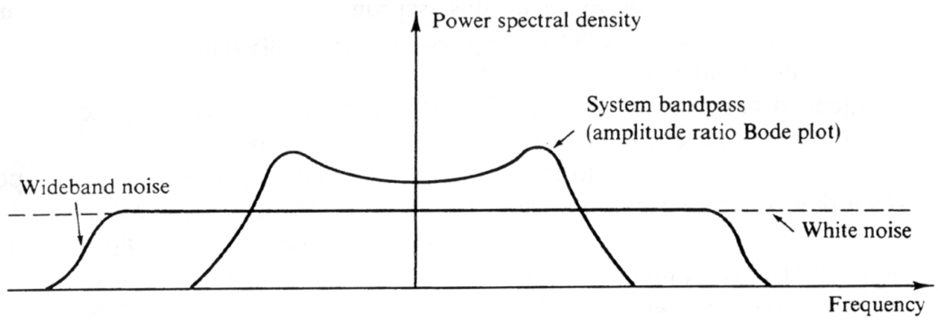


Figure 3: Power spectral density

too high or too low frequency, the bandpass of human ear is between 20 Hz to 20,000 Hz. The wideband noise has power over the frequency beyond and below the system of bandpass, but within the system of bandpass frequency, the spectral density is a constant (horizontal line). The white noise density is represented by dashed line, which has the same spectral density across the whole frequency range. The part within the bandpass, white noise 'coincides' with real wideband noise. There is no coincidence, we merely choose a white noise to be identical to wideband noise within system bandpass. The reason is that the white noise is considerably tractable, then we could replace real bandpass with white noise.

Gaussian

The last assumption states that the measurement noise should also be Gaussian. The first reason is because of Central Limit Theorem, the measurement noise is typically caused by various sources, the addition of these noise sources will produce a joint Gaussian probability density. The second, we only need first and second moment to characterise a Gaussian distribution.

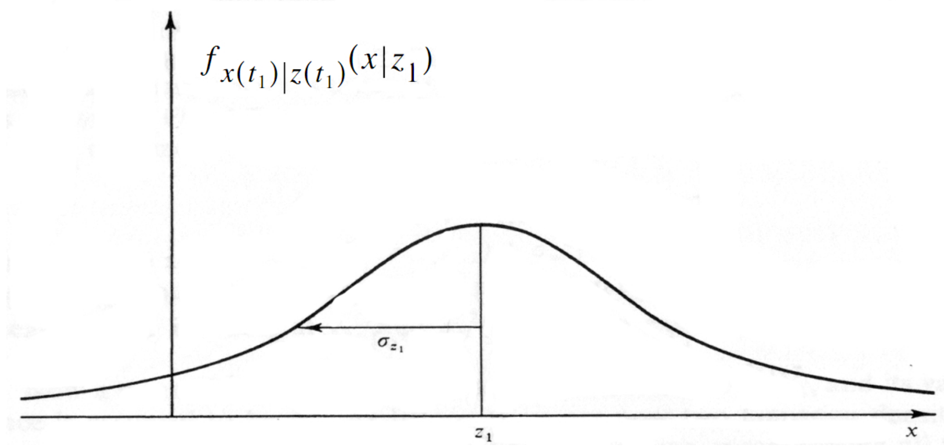


Figure 4: Conditional density of position

3 The Simplest Example

This example is the simplest introductory example of Kalman filter modified from Maybeck (1979) and pictures are borrowed from his book, it gives you the most essential intuitions. Work through this example carefully.

You and your friend are sailing in the Mediterranean sea, and now lost direction in the night. You can infer the location by watching stars, and your friend is trained in navigating on the sea. Say you are moving on a one-dimension route, see figure 4. The x-axis denote the one-dimension route, and you are inferring your location to be at z_1 at time t_1. Certainly it is not precise, your inference has a standard deviation of σ_{z_1}. The density in the figure is conditional on z_1, and best estimate is $\hat{x}(t_1) = z_1$, which is mean, mode and median due to the Guassian assumption.

Now your navigator friend uses his professional skills to decide at time $t_2 \approx t_1$ that the location is at z_2 with a standard deviation σ_{z_2}. Because of

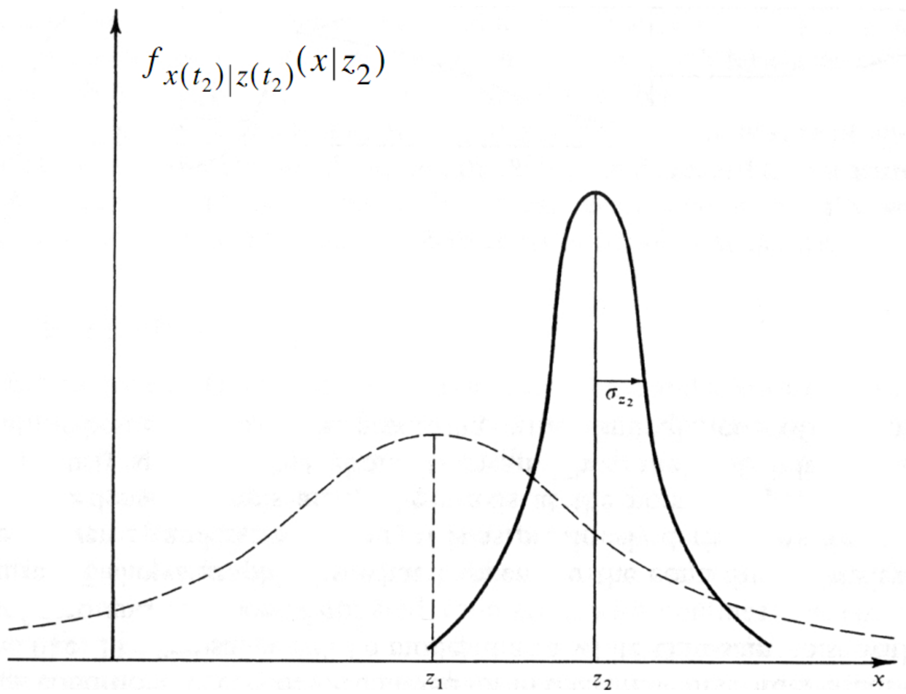


Figure 5: Conditional density on z_2

his professional skill, $\sigma_{z_2} < \sigma_{z_1}$. See figure 5.

The only two sources of information is from you and your friend, thus the problem is how to combine these information to make the best estimate. Refer to equation (1), we have

$$\mu = \frac{\sigma_{z_2}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_1 + \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_2 \tag{2}$$

And define

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_{z_1}^2} + \frac{1}{\sigma_{z_2}^2} \tag{3}$$

which means the addition of precisionness of z_1 and z_2 equals a larger precisionness $1/\sigma^2$. σ^2 is less than either of $\sigma_{z_1}^2$ and $\sigma_{z_2}^2$. See figure 6. We combine

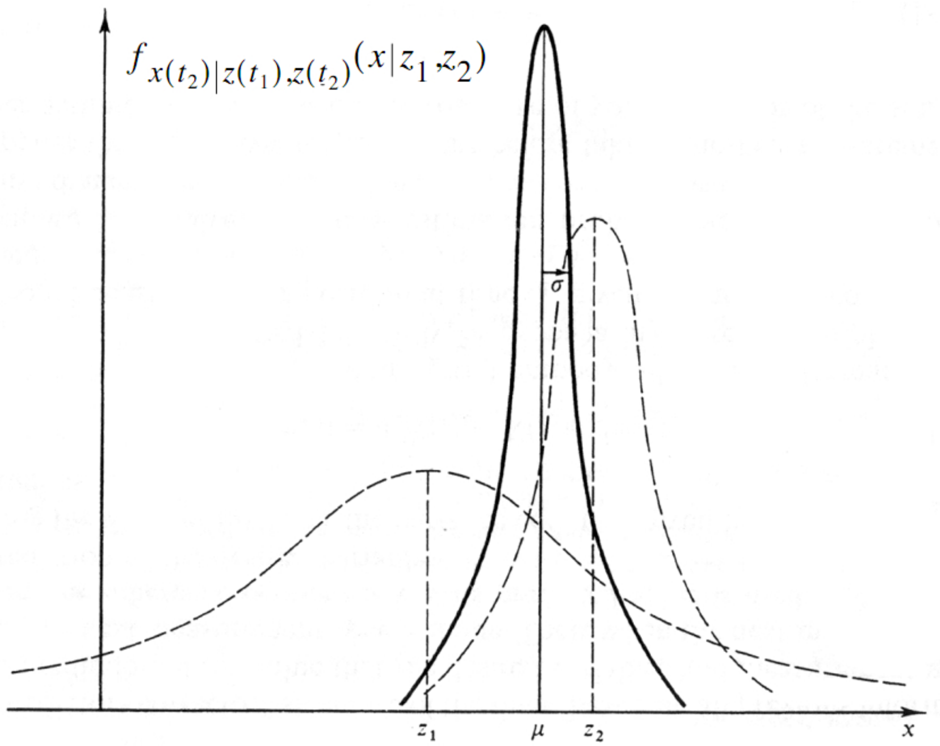


Figure 6: Conditional density based on z_1 and z_2

information from z_1 and z_2 to make the optimal estimate $\hat{x}(t_2) = \mu$. Note that we use more information to update $\hat{x}(t_1)$ to $\hat{x}(t_2)$, this is basically how Kalman filter works.

What if $\sigma_{z_2}^2 = \sigma_{z_1}^2$, then (2) becomes

$$\mu = \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_1}^2} z_1 + \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_1}^2} z_2 = \frac{1}{2} z_1 + \frac{1}{2} z_2 = \frac{1}{2}(z_1 + z_2) \tag{4}$$

which shows that once you put equal weight on both information sources, you simple get an arithmetic average.

To modify (2),

$$\begin{aligned}
\hat{x}(t_2) &= \frac{\sigma_{z_1}^2 + \sigma_{z_2}^2 - \sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_1}^2} z_1 + \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_2 \\
&= \left(1 - \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2}\right) z_1 + \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_2 \\
&= z_1 - \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2}(z_2 - z_1)
\end{aligned}$$

or in the final form,

$$\hat{x}(t_2) = \hat{x}(t_1) + K(t_2)[z_2 - \hat{x}(t_1)] \tag{5}$$

where

$$K(t_2) = \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} \tag{6}$$

From (5), we can see that the subsequent best estimate is based on last one, here $\hat{x}(t_1)$, plus a correction term $K(t_2)[z_2 - \hat{x}(t_1)]$. If $z_2 = \hat{x}(t_1)$, then $\hat{x}(t_2) = \hat{x}(t_1)$.

Use (6), rewrite (3) as

$$\sigma_x^2(t_2) = \sigma_x^2(t_1) - K(t_2)\sigma_x^2(t_1)$$

Once we have $\hat{x}(t_2)$ and $\sigma_x^2(t_2)$, the conditional density is completely specified. This is the advantage of three assumptions, we could always feature the distribution, and update it with most recent information.

References

[1] Maybeck, P.S. (1979): *Stochastic models, estimation and control*, Vol.1, Academic Press.

[2] Kalman, R. (1960): 'A New Approach to Linear Filtering and Prediction Problem,' *Journal of Basic Engineering*,pp. 35-45 (March 1960)