Chapter 3

Kalman Filter and its variants

Reading

1. Barfoot, Chapter 3, 4 for Kalman filter
2. Thrun, Chapter 3 for Kalman filter, 4 for particle filters
3. Russell Chapter 15.4 for Kalman filter

Hidden Markov Models (HMMs) which we discussed in the previous chapter were a very general class of models. As a consequence algorithms for filtering, smoothing and decoding that we prescribed for the HMM are also very general. In this chapter we will consider the situation when we have a little more information about our system. Instead of writing the state transition and observation matrices as arbitrary matrices, we will use the framework of linear dynamical systems to model them better. Since we know the system a bit better, algorithms that we prescribe for these models for solving filtering, smoothing and decoding will also be more efficient. We will almost exclusively focus on the filtering problem in this chapter. The other two, namely smoothing and decoding, can also be solved easily using these ideas but are less commonly used for these systems.

3.1 Background

Multi-variate random variables and linear algebra For d-dimensional random variables $X, Y \in \mathbb{R}^d$ we have

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y];$$

this is actually more surprising than it looks, it is true regardless of whether X, Y are correlated. The covariance matrix of a random variable is defined as

$$\text{cov}(X) = \text{E}[(X - \text{E}[X])\,(X - \text{E}[X])^\top];$$

we will usually denote this by $\Sigma \in \mathbb{R}^{d \times d}$. Note that the covariance matrix is, by construction, symmetric and positive semi-definite. This means it can be factorized as

$$\Sigma = U \Lambda U^\top$$

where $U \in \mathbb{R}^{d \times d}$ is an orthonormal matrix (i.e., $UU^\top = I$) and Λ is a diagonal matrix with non-negative entries. The trace of a matrix is the sum of its diagonal entries. It is also equal to the sum of its eigenvalues, i.e.,

$$\text{tr}(\Sigma) = \sum_{i=1}^{d} \Sigma_{ii} = \sum_{i=1}^{d} \lambda_i(\Sigma)$$

where $\lambda_i(S) \geq 0$ is the i^{th} eigenvalue of the covariance matrix S. The trace is a measure of the uncertainty in the multi-variate random variable X, if X is a scalar and takes values in the reals then the covariance matrix is also, of course, a scalar $\Sigma = \sigma^2$.

A few more identities about the matrix trace that we will often use in this chapter are as follows.

- For matrices A, B we have

  $$\text{tr}(AB) = \text{tr}(BA);$$

  the two matrices need not be square themselves, only their product does.
- For $A, B \in \mathbb{R}^{m \times n}$

  $$\text{tr}(A^\top B) = \text{tr}(B^\top A) = \sum_{i=1}^{m} \sum_{j=1}^{n} B_{ij} A_{ij}.$$

  This operation can be thought of as taking the inner product between two matrices.

Gaussian/Normal distribution We will spend a lot of time working with the Gaussian/Normal distribution. The multi-variate d-dimensional Normal distribution has the probability density

❓ Why is it so ubiquitous?

$$f(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \, \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

where $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$ denote the mean and covariance respectively. You should commit this formula to memory. In particular remember that

$$\int_{x \in \mathbb{R}^d} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\} \, \text{d}x = \sqrt{\det(2\pi\Sigma)}$$

which is simply expressing the fact that the probability density function integrates to 1.

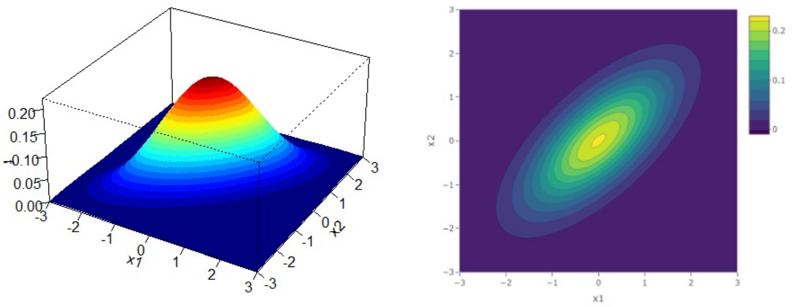


Figure 3.1: Probability density (left) and iso-probability contours (right) of a bi-variate Normal distribution. Warm colors denote regions of high probability.

Given two Gaussian rvs. $X, Y \in \mathbb{R}^d$ and $Z = X + Y$ we have

$$\mathrm{E}[Z] = \mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

with covariance

$$\mathrm{cov}(Z) = \Sigma_Z = \Sigma_X + \Sigma_Y + \Sigma_{XY} + \Sigma_{YX}$$

where

$$\mathbb{R}^{d\times d} \ni \Sigma_{XY} = \mathrm{E}\left[(X - \mathrm{E}[X])\,(Y - \mathrm{E}[Y])^\top\right];$$

the matrix S_{YX} is defined similarly. If X, Y are independent (or uncorrelated) the covariance simplifies to

$$\Sigma_Z = \Sigma_X + \Sigma_Y.$$

If we have a linear function of a Gaussian random variable X given by $Y = AX$ for some *deterministic* matrix A then Y is also Gaussian with mean

$$\mathrm{E}[Y] = \mathrm{E}[AX] = A\,\mathrm{E}[X] = A\mu_X$$

and covariance

$$\begin{aligned}\mathrm{cov}(Y) &= \mathrm{E}[(AX - A\mu_X)(AX - A\mu_X)^\top] \\ &= \mathrm{E}[A(X - \mu_X)(X - \mu_X)^\top A^\top] \\ &= A\,\mathrm{E}[(X - \mu_X)(X - \mu_X)^\top]A^\top \\ &= A\Sigma_X A^\top.\end{aligned}$$

This is an important result that you should remember.

3.2 Linear state estimation

With that background, let us now look at the basic estimation problem. Let $X \in \mathbb{R}^d$ denote the true state of a system. We would like to build an estimator for this state, this is denote by

$$\hat{X}.$$

An estimator is any quantity that indicates our belief of what X is. The estimator is created on the basis of observations and we will therefore model it as a random variable. We would like the estimator to be unbiased, i.e.,

$$\mathrm{E}[\hat{X}] = X;$$

this expresses the concept that if we were to measure the state of the system many times, say using many sensors or multiple observations from the same sensor, the resultant estimator $\hat{X}$ is correct on average. The error in our belief is

$$\tilde{X} = \hat{X} - X.$$

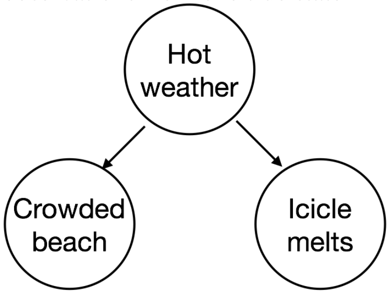
The error is zero-mean $\mathrm{E}[\tilde{X}] = 0$ and its covariance $\Sigma_{\tilde{X}}$ is called the covariance of the estimator.

Optimally combining two estimators Let us now imagine that we have two estimators $\hat{X}_1$ and $\hat{X}_2$ for the same true state X. We will assume that the two estimators were created independently (say different sensors) and therefore are conditionally independent random variables given the true state X Say both of them are unbiased but each of them have a certain covariance of the error

$$\Sigma_{\tilde{X}_1} \text{ and } \Sigma_{\tilde{X}_2}.$$

We would like to combine the two to obtain a *better* estimate of what the state could be. *Better* can mean many different quantities depending upon the problem but in general in this course we are interested in improving the error covariance. Our goal is then

ℹ Conditionally independent observations from one true state



Given two estimators $\hat{X}_1$ and $\hat{X}_2$ of the true state X combine them to obtain a new estimator

$$\hat{X} = \text{some function}(\hat{X}_1, \hat{X}_2)$$

which has the best error covariance $\mathrm{tr}(\Sigma_{\tilde{X}})$.

3.2.1 One-dimensional Gaussian random variables

Consider the case when $\hat{X}_1, \hat{X}_2 \in \mathbb{R}$ are Gaussian random variables with means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively. Assume that both are unbiased estimators of $X \in \mathbb{R}$. Let us combine them linearly to obtain a new estimator

$$\hat{X} = k_1 \hat{X}_1 + k_2 \hat{X}_2.$$

How should we pick the coefficients k_1, k_2? We would of course like the new estimator to be unbiased, so

$$\begin{aligned} &\mathrm{E}[\hat{X}] = \mathrm{E}[k_1 \hat{X}_1 + k_2 \hat{X}_2] = (k_1 + k_2) X = X \\ &\Rightarrow k_1 + k_2 = 1. \end{aligned}$$

The variance of the $\hat{X}$ is

$$\mathrm{var}(\hat{X}) = k_1^2 \sigma_1^2 + k_2^2 \sigma_2^2 = k_1^2 \sigma_1^2 + (1 - k_1)^2 \sigma_2^2.$$

The optimal k_1 that leads to the smallest variance is thus given by

$$k_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

We set the derivative of var$(\hat{X})$ with respect to k_1 to zero to get this. The final estimator is

$$\hat{X} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \hat{X}_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \hat{X}_2. \tag{3.1}$$

It is unbiased of course and has variance

$$\sigma_{\hat{X}}^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Notice that since $\sigma_2^2/(\sigma_1^2 + \sigma_2^2) < 1$, the variance of the new estimator is smaller than either of the original estimators. This is an important fact to remember, combining two estimators *always* results in a better estimator.

Some comments about the optimal combination.

- It is easy to see that if $\sigma_2 \gg \sigma_1$ then the corresponding estimator, namely $\hat{X}_2$ gets less weight in the combination. This is easy to understand, if one of our estimates is very noisy, we should rely less upon it to obtain the new estimate. In the limit that $\sigma_2 \to \infty$, the second estimator is not considered at all in the combination.
- If $\sigma_1 = \sigma_2$, the two estimators are weighted equally and since $\sigma_{\hat{X}}^2 = \sigma_1^2$ there is no reduction in the variance after combination.
- The minimal variance of the combined estimator is not zero. This is easy to see because if we have two noisy estimates of the state, combining them need not lead to us knowing the true state with certainty.

3.2.2 General case

Let us now perform the same exercise for multi-variate Gaussian random variables. We will again combine the two estimators linearly to get

$$\hat{X} = K_1 \hat{X}_1 + K_2 \hat{X}_2$$

where $K_1, K_2 \in \mathbb{R}^{d\times d}$ are matrices that we would like to choose. In order for the estimator to be unbiased we again have the condition

$$\begin{aligned}\mathrm{E}[\hat{X}] &= \mathrm{E}[K_1\hat{X}_1 + K_2\hat{X}_2] = (K_1+K_2)X = X\\ \Rightarrow & K_1 + K_2 = I_{d\times d}.\end{aligned}$$

The covariance of $\hat{X}$ is

$$\begin{aligned}\Sigma_{\hat{X}} &= K_1\Sigma_1K_1^\top + K_2\Sigma_2K_2^\top\\ &= K_1\Sigma_1K_1^\top + (I-K_1)\Sigma_2(I-K_1)^\top.\end{aligned}$$

Just like the minimized the variance in the scalar case, we will minimize the trace of this covariance matrix. We know that the original covariances Σ_1 and Σ_2 are symmetric. We will use the following identity for the partial derivative of a matrix product

$$\frac{\partial}{\partial A}\,\mathrm{tr}(ABA^\top) = 2AB \tag{3.2}$$

for a symmetric matrix B. Minimizing $\mathrm{tr}(\Sigma_{\hat{X}})$ with respect to K_1 amounts to setting

$$\frac{\partial}{\partial K_1}\,\mathrm{tr}(\Sigma_{\hat{X}}) = 0$$

which yields

$$\begin{aligned}0 &= K_1\Sigma_1 - (I-K_1)\Sigma_2\\ \Rightarrow K_1 &= \Sigma_2(\Sigma_1+\Sigma_2)^{-1} \text{ and } K_2 \quad = \Sigma_1(\Sigma_1+\Sigma_2)^{-1}.\end{aligned}$$

The optimal way to combine the two estimators is thus

$$\hat{X} = \Sigma_2(\Sigma_1+\Sigma_2)^{-1}\hat{X}_1 + \Sigma_1(\Sigma_1+\Sigma_2)^{-1}\hat{X}_2. \tag{3.3}$$

You should consider the similarities of this expression with the one for the scalar case in (3.1). The same broad comments hold, i.e., if one of the estimators has a very large variance, that estimator is weighted less in the combination.

3.2.3 Incorporating Gaussian observations of a state

Let us now imagine that we have a sensor that can give us observations of the state. The development in this section is analogous to our calculations in Chapter 2 with the recursive application of Bayes rule or the observation matrix of the HMM. We will consider a special type of sensor that gives observations

$$\mathbb{R}^p \ni Y = CX + \epsilon \tag{3.4}$$

which is a linear function of the true state $X \in \mathbb{R}^d$ with the matrix $C \in \mathbb{R}^{p\times d}$ being something that is unique to the particular sensor. This observation is not precise and we will model the sensor as having zero-mean Gaussian noise

$$\epsilon \sim N(0, Q)$$

of covariance $Q \in \mathbb{R}^{p \times p}$. Notice something important here, the dimensionality of the observations need not be the same as the dimensionality of the state. This should not be surprising, after all the the number of observations in the HMM need not be the same as the number of the states in the Markov chain.

We will solve the following problem. Given an existing estimator $\hat{X}'$ we want to combine it with the observation Y to update the estimator to $\hat{X}$, in the best way, i.e., in a way that gives the minimal variance. We will again use a linear combination

$$\hat{X} = K'\hat{X}' + KY.$$

Again we want the estimator to be unbiased, so we set

$$\begin{aligned} \mathrm{E}[\hat{X}] &= \mathrm{E}[K'\hat{X}' + KY] \\ &= K'X + K\,\mathrm{E}[Y] \\ &= K'X + K\,\mathrm{E}[CX + \epsilon] \\ &= K'X + KCX \\ &= X. \end{aligned}$$

to get that

$$\begin{aligned} I &= K' + KC. \\ \Rightarrow \hat{X} &= (I - KC)\hat{X}' + KY \qquad (3.5) \\ &= \hat{X}' + K(Y - C\hat{X}'). \end{aligned}$$

This is special form which you will do well to remember. The old estimator $\hat{X}'$ gets an additive term $K(Y - C\hat{X}')$. For reasons that will soon become clear, we call this term

$$\text{innovation} = Y - C\hat{X}'.$$

Let us now optimize K as before to compute the estimator with minimal variance. We will make the following important assumption in this case.

> We will assume that the observation Y is independent of the estimator $\hat{X}'$ given X. This is a natural assumption because presumably our original estimator $\hat{X}'$ was created using past observations and the present observation Y is therefore independent of it given the state X.

The covariance of $\hat{X}$ is

$$\Sigma_{\tilde{X}} = (I - KC)\Sigma_{\tilde{X}'}(I - KC)^\top + KQK^\top.$$

We optimize the trace of $\Sigma_{\tilde{X}}$ with respect to K to get

$$\begin{aligned} 0 &= \frac{\partial}{\partial K}\,\mathrm{tr}(\Sigma_{\tilde{X}}) \\ 0 &= -2(I - KC)\Sigma_{\tilde{X}'} + 2KQ \\ \Rightarrow \Sigma_{\tilde{X}'}C^\top &= K(C\Sigma_{\tilde{X}'}C^\top + Q) \\ \Rightarrow K &= \Sigma_{\tilde{X}'}C^\top(C\Sigma_{\tilde{X}'}C^\top + Q)^{-1}. \end{aligned}$$

The matrix $K \in \mathbb{R}^{d \times p}$ is called the “Kalman gain” after Rudoph Kalman who developed this method in the 1960s.

Kalman gain This is an important formula and it helps to have a mnemonic and a slightly simpler notation to remember it by. If Σ' is the covariance of the previous estimator, Q is the covariance of the zero-mean observation and C is the matrix that gives the observation from the state, then the Kalman gain is

$$K = \Sigma_{\tilde{X}'} C^\top (C \Sigma_{\tilde{X}'} C^\top + Q)^{-1}. \tag{3.6}$$

and the new estimator for the state is

$$\hat{X} = \hat{X}' + K(Y - C\hat{X}').$$

Some comments

- If $C = I$, the Kalman gain is the same expression as the optimal coefficient in (3.3). This should not be surprising because the observation is an estimator for the state.
-

3.3 Background on linear dynamical systems

3.4 Kalman Filter (KF)

3.5 Extended-Kalman Filter (EKF)

3.6 Unscented Kalman Filter (UKF)

3.7 Particle Filters (PFs)

Bibliography