CHAPTER 3

Least squares estimation

> The most probable value of the unknown quantities will be that in which the sum of the squares of the differences between the actually observed and the computed values multiplied by numbers that measure the degree of precision is a minimum.
>
> —Karl Friedrich Gauss [Gau04]

In this chapter, we will discuss least squares estimation, which is the basic idea of Karl Gauss's quote above.[1] The material in this chapter relies on the theory of the previous two chapters, and will enable us to derive optimal state estimators later in this book.

Section 3.1 discusses the estimation of a constant vector on the basis of several linear but noisy measurements of that vector. Section 3.2 extends the results of Section 3.1 to the case in which some measurements are more noisy than others; that is, we have less confidence in some measurements than in others. Sections 3.1 and 3.2 use matrices and vectors whose dimensions grow larger as more measurements are obtained. This makes the problem cumbersome if many measurements are available. This leads us to Section 3.3, which presents a recursive way of estimating a constant on the basis of noisy measurements. Recursive estimation in this chapter is a method of estimating a constant without increasing the computa-

[1] Gauss published his book in 1809, although he claimed to have worked out his theory as early as 1795 (when he was 18 years old).

Optimal State Estimation, First Edition. By Dan J. Simon
ISBN 0471708585

tional effort of the algorithm, regardless of how many measurements are available. Finally, Section 3.4 presents the Wiener filter, which is a method of estimating a time-varying signal that is corrupted by noise, on the basis of noisy measurements. Until 1960, Wiener filtering was the state of the art in signal estimation. The paradigm of signal estimation was shattered with the publication of Rudolph Kalman's work and related papers in the early 1960s, but it is still worthwhile understanding Wiener filtering because of its historical place in the history of signal estimation. Furthermore, Wiener filtering is still very useful in signal processing and communication theory.

3.1 ESTIMATION OF A CONSTANT

In this section, we will determine how to estimate a constant on the basis of several noisy measurements of that constant. For example, suppose we have a resistor but we do not know its resistance. We take several measurements of its resistance using a multimeter, but the measurements are noisy because we have a cheap multimeter. We want to estimate the resistance on the basis of our noisy measurements. In this case, we want to estimate a constant scalar but, in general, we may want to estimate a constant vector.

To put the problem in mathematical terms, suppose x is a constant but unknown n-element vector, and y is a k-element noisy measurement vector. How can we find the "best" estimate $\hat{x}$ of x? Let us assume that each element of the measurement vector y is a linear combination of the elements of x, with the addition of some measurement noise:

$$\begin{aligned} y_1 &= H_{11}x_1 + \cdots + H_{1n}x_n + v_1 \\ &\vdots \\ y_k &= H_{k1}x_1 + \cdots + H_{kn}x_n + v_k \end{aligned} \tag{3.1}$$

This set of equations can be put into matrix form as

$$y = Hx + v \tag{3.2}$$

Now define ϵ_y as the difference between the noisy measurements and the vector $H\hat{x}$:

$$\epsilon_y = y - H\hat{x} \tag{3.3}$$

ϵ_y is called the measurement residual. As Karl Gauss wrote [Gau04], the most probable value of the vector x is the vector $\hat{x}$ that minimizes the sum of squares between the observed values y and the vector $H\hat{x}$. So we will try to compute the $\hat{x}$ that minimizes the cost function J, where J is given as

$$\begin{aligned} J &= \epsilon_{y1}^2 + \cdots + \epsilon_{yk}^2 \\ &= \epsilon_y^T \epsilon_y \end{aligned} \tag{3.4}$$

J is often referred to in control and estimation books and papers as a cost function, objective function, or return function. We can substitute for ϵ_y in the above equation to rewrite J as

$$\begin{aligned} J &= (y - H\hat{x})^T (y - H\hat{x}) \\ &= y^T y - \hat{x}^T H^T y - y^T H\hat{x} + \hat{x}^T H^T H\hat{x} \end{aligned} \tag{3.5}$$

In order to minimize J with respect to $\hat{x}$, we compute its partial derivative and set it equal to zero:

$$\begin{aligned} \frac{\partial J}{\partial \hat{x}} &= -y^T H - y^T H + 2\hat{x}^T H^T H \\ &= 0 \end{aligned} \tag{3.6}$$

Solving this equation for $\hat{x}$ results in

$$\begin{aligned} H^T y &= H^T H \hat{x} \\ \hat{x} &= (H^T H)^{-1} H^T y \\ &= H^L y \end{aligned} \tag{3.7}$$

where H^L, the left pseudo inverse of H, exists if $k \geq n$ and H is full rank. This means that the number of measurements k is greater than the number of variables n that we are trying to estimate, and the measurements are linearly independent. In order to prove that we have found a minimum rather than some other type of stationary point[2] of J, we need to prove that the second derivative of J is positive semidefinite (see Problem 3.1).

■ EXAMPLE 3.1

Let us go back to our original problem of trying to estimate the resistance x of an unmarked resistor on the basis of k noisy measurements from a multimeter. In this case, x is a scalar so our k noisy measurements are given as

$$\begin{aligned} y_1 &= x + v_1 \\ &\vdots \\ y_k &= x + v_k \end{aligned} \tag{3.8}$$

These k equations can be combined into a single matrix equation as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} x + \begin{bmatrix} v_1 \\ \vdots \\ v_k \end{bmatrix} \tag{3.9}$$

Equation (3.7) shows that the optimal estimate of the resistance x is given as

$$\begin{aligned} \hat{x} &= (H^T H)^{-1} H^T y \\ &= \left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \\ &= \frac{1}{k}(y_1 + \cdots + y_k) \end{aligned} \tag{3.10}$$

In this simple example, we see that least squares estimation agrees with our intuition to simply compute the average of the measurements.

▽▽▽

[2] A stationary point of a function is any point at which its derivative is equal to zero. A stationary point of a scalar function could be a maximum, a minimum, or an inflection point. A stationary point of a vector function could be a maximum, a minimum, or a saddle point.

3.2 WEIGHTED LEAST SQUARES ESTIMATION

In the previous section, we assumed that we had an equal amount of confidence in all of our measurements. Now suppose we have more confidence in some measurements than others. In this case, we need to generalize the results of the previous section to obtain weighted least squares estimation. For example, suppose we have several measurements of the resistance of an unmarked resistor. Some of the measurements were taken with an expensive multimeter with low noise, but other measurements were taken with a cheap multimeter by a tired student late at night. We have more confidence in the first set of measurements, so we should somehow place more emphasis on those measurements than on the others. However, even though the second set of measurements is less reliable, it seems that we could get at least *some* information from them. This section shows that we can indeed get some information from less reliable measurements. We should never throw away measurements, no matter how unreliable they may be.

To put the problem in mathematical terms, suppose x is a constant but unknown n-element vector, and y is a k-element noisy measurement vector. We assume that each element of y is a linear combination of the elements of x, with the addition of some measurement noise, and the variance of the measurement noise may be different for each element of y:

$$\begin{aligned} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} &= \begin{bmatrix} H_{11} & \cdots & H_{1n} \\ \vdots & \ddots & \vdots \\ H_{k1} & \cdots & H_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_k \end{bmatrix} \\ E(v_i^2) &= \sigma_i^2 \quad (i = 1, \ldots, k) \end{aligned} \tag{3.11}$$

We assume that the noise for each measurement is zero-mean and independent. The measurement covariance matrix is

$$\begin{aligned} R &= E(vv^T) \\ &= \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_k^2 \end{bmatrix} \end{aligned} \tag{3.12}$$

Now we will minimize the following quantity with respect to $\hat{x}$.

$$J = \epsilon_{y1}^2/\sigma_1^2 + \cdots + \epsilon_{yk}^2/\sigma_k^2 \tag{3.13}$$

Note that instead of minimizing the sum of squares of the ϵ_y elements as we did in Equation (3.4), we will minimize the *weighted* sum of squares. If y_1 is a relatively noisy measurement, for example, then we do not care as much about minimizing the difference between y_1 and the first element of $H\hat{x}$ because we do not have much confidence in y_1 in the first place. The cost function J can be written as

$$\begin{aligned} J &= \epsilon_y^T R^{-1} \epsilon_y \\ &= (y - H\hat{x})^T R^{-1} (y - H\hat{x}) \\ &= y^T R^{-1} y - \hat{x}^T H^T R^{-1} y - y^T R^{-1} H\hat{x} + \hat{x}^T H^T R^{-1} H\hat{x} \end{aligned} \tag{3.14}$$

Now we take the partial derivative of J with respect to $\hat{x}$ and set it equal to zero to compute the best estimate $\hat{x}$:

$$\begin{aligned} \frac{\partial J}{\partial \hat{x}} &= -y^T R^{-1} H + \hat{x}^T H^T R^{-1} H \\ &= 0 \\ H^T R^{-1} y &= H^T R^{-1} H \hat{x} \\ \hat{x} &= (H^T R^{-1} H)^{-1} H^T R^{-1} y \end{aligned} \tag{3.15}$$

Note that this method requires that the measurement noise matrix R be nonsingular. In other words, each of the measurements y_i must be corrupted by at least *some* noise for this method to work.

■ EXAMPLE 3.2

We return to our original problem of trying to estimate the resistance x of an unmarked resistor on the basis of k noisy measurements from a multimeter. In this case, x is a scalar so our k noisy measurements are given as

$$\begin{aligned} y_i &= x + v_i \\ E(v_i^2) &= \sigma_i^2 \quad (i = 1, \ldots, k) \end{aligned} \tag{3.16}$$

The k measurement equation can be combined into a single matrix equation as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} x + \begin{bmatrix} v_1 \\ \vdots \\ v_k \end{bmatrix} \tag{3.17}$$

and the measurement noise covariance is given as

$$R = \text{diag}(\sigma_1^2, \ldots, \sigma_k^2) \tag{3.18}$$

Equation (3.15) shows that the optimal estimate of the resistance x is given as

$$\begin{aligned} \hat{x} &= (H^T R^{-1} H)^{-1} H^T R^{-1} y \\ &= \left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \times \\ &\quad \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k^2 \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \\ &= \left(\sum 1/\sigma_i^2 \right)^{-1} (y_1/\sigma_1^2 + \cdots + y_k/\sigma_k^2) \end{aligned} \tag{3.19}$$

We see that the optimal estimate $\hat{x}$ is a weighted sum of the measurements, where each measurement is weighted by the inverse of its uncertainty. In other words, we put more emphasis on certain measurements, in agreement

with our intuition. Note that if all of the σ_i constants are equal, this estimate reduces to the simpler form given in Equation (3.10).

$\nabla\nabla\nabla$

3.3 RECURSIVE LEAST SQUARES ESTIMATION

Equation (3.15) gives us a way to compute the optimal estimate of a constant, but there is a problem. Note that the H matrix in (3.15) is a $k \times n$ matrix. If we obtain measurements sequentially and want to update our estimate of x with each new measurement, we need to augment the H matrix and completely recompute the estimate $\hat{x}$. If the number of measurements becomes large, then the computational effort could become prohibitive. For example, suppose we obtain a measurement of a satellite's altitude once per second. After one hour has passed, the number of measurements is 3600 and growing. The computational effort of least squares estimation can rapidly outgrow our resources.

In this section, we show how to *recursively* compute the weighted least squares estimate of a constant. That is, suppose we have $\hat{x}$ after $(k-1)$ measurements, and we obtain a new measurement y_k. How can we update our estimate without completely reworking Equation (3.15)?

A linear recursive estimator can be written in the form

$$\begin{aligned} y_k &= H_k x + v_k \\ \hat{x}_k &= \hat{x}_{k-1} + K_k(y_k - H_k\hat{x}_{k-1}) \end{aligned} \tag{3.20}$$

That is, we compute $\hat{x}_k$ on the basis of the previous estimate $\hat{x}_{k-1}$ and the new measurement y_k. K_k is a matrix to be determined called the estimator gain matrix. The quantity $(y_k - H_k\hat{x}_{k-1})$ is called the correction term. Note that if the correction term is zero, or if the gain matrix is zero, then the estimate does not change from time step $(k-1)$ to k.

Before we compute the optimal gain matrix K_k, let us think about the mean of the estimation error of the linear recursive estimator. The estimation error mean can be computed as

$$\begin{aligned} E(\epsilon_{x,k}) &= E(x - \hat{x}_k) \\ &= E[x - \hat{x}_{k-1} - K_k(y_k - H_k\hat{x}_{k-1})] \\ &= E[\epsilon_{x,k-1} - K_k(H_k x + v_k - H_k\hat{x}_{k-1})] \\ &= E[\epsilon_{x,k-1} - K_k H_k(x - \hat{x}_{k-1}) - K_k v_k] \\ &= (I - K_k H_k)E(\epsilon_{x,k-1}) - K_k E(v_k) \end{aligned} \tag{3.21}$$

So if $E(v_k) = 0$ and $E(\epsilon_{x,k-1}) = 0$, then $E(\epsilon_{x,k}) = 0$. In other words, if the measurement noise v_k is zero-mean for all k, and the initial estimate of x is set equal to the expected value of x [i.e., $\hat{x}_0 = E(x)$], then the expected value of $\hat{x}_k$ will be equal to x_k for all k. Because of this, the estimator of Equation (3.20) is called an unbiased estimator. Note that this property holds regardless of the value of the gain matrix K_k. This is a desirable property of an estimator because it says that, *on average*, the estimate $\hat{x}$ will be equal to the true value x.

Next we turn our attention to the determination of the optimal value of K_k. Since the estimator is unbiased regardless of what value of K_k we use, we must