Recursive Least Squares Estimation*

(Com 477/577 Notes)

Yan-Bin Jia

Dec 12, 2013

1 Estimation of a Constant

We start by with estimation of a constant based on several noisy measurements. Suppose we have a resistor but do not know its resistance. So we measure it several times using a cheap (and noisy) multimeter. How do we come up with a good estimate of the resistance based on these noisy measurements?

More formally, suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ is a constant but unknown vector, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)^T$ is a k-element noisy measurement vector. Our task is to find the "best" estimate $\tilde{\boldsymbol{x}}$ of $\boldsymbol{x}$. Here we look at perhaps the simplest case where each y_i is a linear combination of x_j, $1 \leq j \leq n$, with addition of some measurement noise ν_i. Thus, we are working with the following linear system,

$$\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{\nu},$$

where $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_k)^T$, and H is an $k \times n$ matrix; or with all terms listed,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = \begin{pmatrix} H_{11} & \cdots & H_{1n} \\ \vdots & \ddots & \vdots \\ H_{k1} & \cdots & H_{kn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \nu_1 \\ \vdots \\ \nu_k \end{pmatrix},$$

Given an estimate $\tilde{\boldsymbol{x}}$, we consider the difference between the noisy measurements and the projected values $H\tilde{\boldsymbol{x}}$:

$$\boldsymbol{\epsilon} = \boldsymbol{y} - H\tilde{\boldsymbol{x}}.$$

Under the least squares principle, we will try to find the value of $\tilde{\boldsymbol{x}}$ that minimizes the cost function

$$\begin{aligned} J(\tilde{\boldsymbol{x}}) &= \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\ &= (\boldsymbol{y} - H\tilde{\boldsymbol{x}})^T(\boldsymbol{y} - H\tilde{\boldsymbol{x}}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \tilde{\boldsymbol{x}}^T H\boldsymbol{y} - \boldsymbol{y}^T H\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{x}}^T H^T H\tilde{\boldsymbol{x}}. \end{aligned}$$

The necessary condition for the minimum is the vanishing of the partial derivative of J with respect to $\tilde{\boldsymbol{x}}$, that is,

$$\frac{\partial J}{\partial \tilde{\boldsymbol{x}}} = -2\boldsymbol{y}^T H + 2\tilde{\boldsymbol{x}}^T H^T H = 0.$$

*The material is adapted from Dan Simon's book *Optimal State Estimation* [1].

We solve the equation, obtaining

$$\tilde{\boldsymbol{x}} = (H^T H)^{-1} H^T \boldsymbol{y}. \tag{1}$$

The inverse $(H^T H)^{-1}$ exists if $k > n$ and H is non-singular. In other words, when the number of measurements is no fewer than the number of variables, and these measurements are linearly independent.

EXAMPLE 1. Suppose we are trying to estimate the resistance x of an unmarked resistor based on k noisy measurements using a multimeter. In this case,

$$\boldsymbol{y} = Hx + \boldsymbol{\nu}, \tag{2}$$

where

$$H = (1, \cdots, 1)^T. \tag{3}$$

Substitution of the above into equation (1) gives us the optimal estimate of x as

$$\begin{aligned} \tilde{x} &= (H^T H)^{-1} H^T \boldsymbol{y} \\ &= \frac{1}{k} H^T \boldsymbol{y} \\ &= \frac{y_1 + \cdots + y_k}{k}. \end{aligned}$$

2 Weighed Least Squares Estimation

So far we have placed equal confidence on all the measurements. Now we look at varying confidence in the measurements. For instance, some of our measurements of an unmarked resistor were taken with an expensive multimeter with low noise, while others were taken with a cheap multimeter by a tired student late at night. Even though the second set of measurements is less reliable, we could get some information about the resistance. We should never throw away measurements, no matter how reliable they may seem. This will be shown in the section.

We assume that each measurement y_i, $1 \leq i \leq k$ may be taken under a different condition so that the variance ν_i of the measurement noise may be distinct too:

$$E(\nu_i^2) = \sigma_i^2, \qquad 1 \leq i \leq k.$$

Assume that the noise for each measurement has zero mean and is independent. The covariance matrix for all measurement noise is

$$\begin{aligned} R &= E(\boldsymbol{\nu}\boldsymbol{\nu}^T) \\ &= \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k^2 \end{pmatrix}. \end{aligned}$$

Write the difference $\boldsymbol{y} - H\tilde{\boldsymbol{x}}$ as $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_k)^T$. We will minimize the sum of squared differences weighted over the variations of the measurements:

$$J(\tilde{\boldsymbol{x}}) = \boldsymbol{\epsilon}^T R^{-1} \boldsymbol{\epsilon} = \frac{\epsilon_1^2}{\sigma_1^2} + \frac{\epsilon_2^2}{\sigma_2^2} + \cdots + \frac{\epsilon_k^2}{\sigma_k^2}.$$

If a measurement y_i is noisy, we care less about the discrepancy between it and the ith element of $H\tilde{\boldsymbol{x}}$ because we do not have much confidence in this measurement. The cost function J can be expanded as follows:

$$\begin{aligned} J(\tilde{\boldsymbol{x}}) &= (\boldsymbol{y} - H\tilde{\boldsymbol{x}})^T R^{-1}(\boldsymbol{y} - H\tilde{\boldsymbol{x}}) \\ &= \boldsymbol{y}^T R^{-1}\boldsymbol{y} - \tilde{\boldsymbol{x}}^T H^T R^{-1}\boldsymbol{y} - \boldsymbol{y}^T R^{-1} H\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{x}}^T H^T R^{-1} H\tilde{\boldsymbol{x}}. \end{aligned}$$

At a minimum, the partial derivative of J must vanish, yielding

$$\frac{\partial J}{\partial \tilde{\boldsymbol{x}}} = -2\boldsymbol{y}^T R^{-1} H + 2\tilde{\boldsymbol{x}}^T H^T R^{-1} H = 0.$$

Immediately, we solve the above equation for the best estimate of $\boldsymbol{x}$:

$$\tilde{\boldsymbol{x}} = (H^T R^{-1} H)^{-1} H^T R^{-1}\boldsymbol{y}. \tag{4}$$

Note that the measurement noise matrix R must be non-singular for a solution to exist. In other words, each measurement y_i must be corrupted by some noise for the estimation method to work.

EXAMPLE 2. We get back to the problem in Example 1 of resistance estimation, for which the equations are given in (2) and (3). Suppose each of the k noisy measurements has variance

$$E(\nu_i^2) = \sigma_i^2.$$

The measurement noise covariance is given as

$$R = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2).$$

Substituting $H, R, \boldsymbol{y}$ into (4), we obtain the estimate

$$\tilde{x} = \left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)^{-1} \left(\sum_{i=1}^{k} \frac{y_i}{\sigma_i^2}\right).$$

3 Recursive Least Squares Estimation

Equation (4) is adequate when we have made all the measurements. More often, we obtain measurements sequentially and want to update our estimate with each new measurement. In this case, the matrix H needs to be augmented. We would have to recompute the estimate $\tilde{\boldsymbol{x}}$ according to (4) for every new measurement. This update can become very expensive. And the overall computation can become prohibitive as the number of measurements becomes large.

This section shows how to recursively compute the weighted least squares estimate. More specifically, suppose we have an estimate $\tilde{\boldsymbol{x}}_{k-1}$ after $k-1$ measurements, and obtain a new measurement y_k. How can we update the estimate to $\tilde{\boldsymbol{x}}_k$ without solving equation (4)?

A linear recursive estimator can be written in the following form:

$$\begin{aligned} y_k &= H_k^T \boldsymbol{x} + \nu_k, \\ \tilde{\boldsymbol{x}}_k &= \tilde{\boldsymbol{x}}_{k-1} + K_k(y_k - H_k^T \tilde{\boldsymbol{x}}_{k-1}). \end{aligned} \tag{5}$$

Here H_k is an n-vector, and K_k is an n-vector referred to as the *estimator gain vector*. We refer to $y_k - H_k^T \tilde{\boldsymbol{x}}_{k-1}$ as the *correction term*. Namely, the new estimate $\tilde{\boldsymbol{x}}_k$ is modified from the previous estimate $\tilde{\boldsymbol{x}}_k$ with a correction via the gain vector.

The current estimation error is $\boldsymbol{\epsilon}_k = \boldsymbol{x} - \tilde{\boldsymbol{x}}_k$. We first look at the mean of this error which is computed as follows:

$$\begin{aligned} E(\boldsymbol{\epsilon}_k) &= E(\boldsymbol{x} - \tilde{\boldsymbol{x}}_k) \\ &= E\Big(\boldsymbol{x} - \tilde{\boldsymbol{x}}_{k-1} - K_k(y_k - H_k^T \tilde{\boldsymbol{x}}_{k-1})\Big) \\ &= E\Big(\boldsymbol{\epsilon}_{k-1} - K_k(H_k^T \boldsymbol{x} + \nu_k - H_k^T \tilde{\boldsymbol{x}}_{k-1})\Big) \\ &= E\Big(\boldsymbol{\epsilon}_{k-1} - K_k H_k^T(\boldsymbol{x} - \tilde{\boldsymbol{x}}_{k-1}) - K_k \nu_k\Big) \\ &= (I - K_k H_k^T) E(\boldsymbol{\epsilon}_{k-1}) - K_k E(\nu_k), \end{aligned} \tag{6}$$

where I is the $n \times n$ identity matrix. If $E(\nu_k) = 0$ and $E(\boldsymbol{\epsilon}_{k-1}) = \mathbf{0}$, then $E(\boldsymbol{\epsilon}_k) = \mathbf{0}$. So if the measurement noise ν_k has zero mean for all k, and the initial estimate of $\boldsymbol{x}$ is set equal to its expected value, then $\tilde{\boldsymbol{x}}_k = \boldsymbol{x}_k$ for all k. With this property, the estimator (5) is called *unbiased*. The property holds regardless of the value of the gain vector K_k. It says that on the average the estimate $\tilde{\boldsymbol{x}}$ will be equal to the true value $\boldsymbol{x}$.

The key is to determine the optimal value of the gain vector K_k. The optimality criterion used by us is to minimize the aggregated variance of the estimation errors at time k:

$$\begin{aligned} J_k &= E(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}_k\|^2) \\ &= E(\boldsymbol{\epsilon}_k^T \boldsymbol{\epsilon}_k) \\ &= E\left(\mathrm{tr}(\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^T)\right) \\ &= \mathrm{tr}(P_k), \end{aligned} \tag{7}$$

where tr is the trace operator[1], and $P_k = E(\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^T)$ is the estimation-error covariance, Next, we obtain P_k with a substitution of (6):

$$\begin{aligned} P_k &= E\left(\Big((I - K_k H_k^T)E(\boldsymbol{\epsilon}_k) - K_k \nu_k\Big)\Big((I - K_k H_k^T)E(\boldsymbol{\epsilon}_k) - K_k \nu_k\Big)^T\right) \\ &= (I - K_k H_k^T) E(\boldsymbol{\epsilon}_{k-1} \boldsymbol{\epsilon}_{k-1}^T)(I - K_k H_k^T)^T - K_k E(\nu_k \boldsymbol{\epsilon}_{k-1}^T)(I - K_k H_k^T)^T \\ &\quad - (I - K_k H_k^T) E(\boldsymbol{\epsilon}_{k-1} \nu_k^T) K_k^T + K_k E(\nu_k^2) K_k^T. \end{aligned}$$

The last step uses a fact that the estimation error $\boldsymbol{\epsilon}_{k-1}$ at time $k-1$ is independent of the measurement noise ν_k at time k. The latter implies that

$$E(\nu_k \boldsymbol{\epsilon}_{k-1}^T) = E(\nu_k) E(\boldsymbol{\epsilon}_{k-1}^T) = 0.$$

Given the definition of $R_k = E(\nu_k^2)$ as covariance of ν_k, the expression of P_k becomes

$$P_k = (I - K_k H_k^T) P_{k-1} (I - K_k H_k^T)^T + K_k R_k K_k^T. \tag{8}$$

Equation (8) is the recurrence for the covariance of the least squares estimation error. It is consistent with the intuition that as the measurement noise (R_k) increases, the uncertainty (P_k) increases. Note that P_k as a covariance matrix is positive definite.

[1]The trace of a matrix is the sum of its diagonal elements.

What remains is to find the value of the gain vector K_k that minimizes the cost function given by (6). The mean of the estimation error is zero independent of the value of K_k already. Thus the minimizing value of K_k will make the cost function consistently close to zero. We need to differentiate J_k with respect to K_k. Note that the derivative of a function f with respect to a matrix $A = (a_{ij})$ is a matrix $\frac{\partial f}{\partial A} = (\frac{\partial f}{\partial a_{ij}})$. Also, we make use of a fact that $\frac{\partial}{\partial A}\mathrm{tr}(ABA^T) = 2AB$ when B is symmetric. With this in mind, we first substitute (8) into (7) and differentiate the resulting expression with respect to K_k:

$$\left(\frac{\partial J_k}{\partial K_k}\right)^T = 2(I - K_k H_k^T)P_{k-1}(-H_k) + 2K_k R_k.$$

Setting the partial derivative to zero, we solve for K_k:

$$K_k = P_{k-1}H_k(H_k^T P_{k-1}H_k + R_k)^{-1}. \tag{9}$$

Write $S_k = H_k^T P_{k-1}H_k + R_k$, so

$$K_k = P_{k-1}H_k S_k^{-1}. \tag{10}$$

Substitute the above for K_k into equation (8) for P_k. The operation followed by an expansion leads to a few steps of manipulation as follows:

$$\begin{aligned}
P_k &= (I - P_{k-1}H_k S_k^{-1}H_k^T)P_{k-1}(I - P_{k-1}H_k S_k^{-1}H_k^T)^T + P_{k-1}H_k S_k^{-1}R_k S_k^{-1}H_k^T P_{k-1} \\
&= P_{k-1} - P_{k-1}H_k S_k^{-1}H_k^T P_{k-1} - P_{k-1}H_k S_k^{-1}H_k^T P_{k-1} + \\
&\qquad P_{k-1}H_k S_k^{-1}\underline{H_k^T P_{k-1}H_k}S_k^{-1}H_k^T P_{k-1} + P_{k-1}H_k S_k^{-1}\underline{R_k}S_k^{-1}H_k^T P_{k-1} \\
&= P_{k-1} - P_{k-1}H_k S_k^{-1}H_k^T P_{k-1} - P_{k-1}H_k S_k^{-1}H_k^T P_{k-1} + P_{k-1}H_k S_k^{-1}S_k S_k^{-1}H_k^T P_{k-1} \\
&\qquad\qquad \text{after merging the underlined terms into } S_k \\
&= P_{k-1} - 2P_{k-1}H_k S_k^{-1}H_k^T P_{k-1} + P_{k-1}H_k S_k^{-1}H_k^T P_{k-1} \\
&= P_{k-1} - P_{k-1}H_k S_k^{-1}H_k^T P_{k-1} \\
&= P_{k-1} - K_k H_k^T P_{k-1} \qquad\qquad \text{by (10)} \\
&= (I - K_k H_k^T)P_{k-1}. \qquad (11)
\end{aligned}$$

Note that in the above P_k is symmetric as a covariance matrix, and so is S_k.

4 The Estimation Algorithm

The algorithm for recursive least squares estimation is summarized as follows.

1. Initialize the estimator:

$$\begin{aligned}
\tilde{\boldsymbol{x}}_0 &= E(\boldsymbol{x}), \\
P_0 &= E\left((\boldsymbol{x} - \tilde{\boldsymbol{x}}_0)(\boldsymbol{x} - \tilde{\boldsymbol{x}}_0)^T\right).
\end{aligned}$$

   In the case of no prior knowledge about $\boldsymbol{x}$, simply let $P_0 = \infty I$. In the case of perfect prior knowledge, let $P_0 = 0$.

2. Iterate the follow two steps.

   (a) Obtain a new measurement y_k, assuming that it is given by the equation

   $$y_k = H_k^T \boldsymbol{x} + \nu_k,$$

   where the noise ν_k has zero mean and covariance R_k. The measurement noise at each time step k is independent. So,

   $$E(\nu^2) = \begin{cases} 0, & \text{if } i \neq j, \\ R_j, & \text{if } i = j. \end{cases}$$

   Essentially, we assume white measurement noise.

   (b) Update the estimate $\tilde{\boldsymbol{x}}$ and the covariance P of the estimation error sequentially as follows:

$$\begin{aligned} K_k &= P_{k-1} H_k (H_k^T P_{k-1} H_k + R_k)^{-1}, && (12) \\ \tilde{\boldsymbol{x}}_k &= \tilde{\boldsymbol{x}}_{k-1} + K_k (y_k - H_k^T \tilde{\boldsymbol{x}}_{k-1}), && (13) \\ P_k &= (I - K_k H_k^T) P_{k-1}. && (14) \end{aligned}$$

EXAMPLE 3. We revisit the resistance estimation problem presented in Examples 1 and 2. Now, we want to iteratively improve our estimate of the resistance x. At the kth sampling, our measurement is

$$\begin{aligned} y_k &= H_k^T x + \nu_k = x + \nu_k, \\ R_k &= E(\nu_k^2). \end{aligned}$$

Here, the measurement vector H_k is a scalar 1. Furthermore, we suppose that each measurement has the same covariance so R_k is a constant written as R.

Before the first measurement, we have some idea about the resistance x. This becomes our initial estimate. Also, we have some uncertainty about this initial estimate, which becomes our initial covariance. Together we have

$$\begin{aligned} \tilde{x}_0 &= E(x), \\ P_0 &= E((x - \tilde{x}_0)^2). \end{aligned}$$

If we have no idea about the resistance, set $P(0) = \infty$. If we are certain about the resistance value, set $P(0) = 0$. (Of course, then there would be no need to take measurements.)

After the first measurement (k=1), we update the estimate and the error covariance according to equations (12)–(14) as follows:

$$\begin{aligned} K_1 &= \frac{P_0}{P_0 + R}, \\ \tilde{x}_1 &= \tilde{x}_0 + \frac{P_0}{P_0 + R}(y_1 - \tilde{x}_0), \\ P_1 &= \frac{P_0 R}{P_0 + R}. \end{aligned}$$

After the second measurement, the estimates become

$$K_2 = \frac{P_1}{P_1 + R} = \frac{P_0}{2P_0 + R},$$

$$
\begin{aligned}
P_2 &= \frac{P_1 R}{P_1 + R} = \frac{P_0 R}{2P_0 + R}, \\
\tilde{x}_2 &= \tilde{x}_1 + \frac{P_1}{P_1 + R}(y_2 - \tilde{x}_1) \\
&= \frac{P_0 + R}{2P_0 + R}\tilde{x}_1 + \frac{P_0}{2P_0 + R} y_2.
\end{aligned}
$$

By induction, we can show that

$$
\begin{aligned}
K_k &= \frac{P_0}{kP_0 + R}, \\
\tilde{x}_k &= \frac{(k-1)P_0 + R}{kP_0 + R}\tilde{x}_{k-1} + \frac{P_0}{kP_0 + R} y_k, \\
P_k &= \frac{P_0 R}{kP_0 + R}.
\end{aligned}
$$

Note that if x is known perfectly *a priori*, then $P_0 = 0$, which implies that $K_k = 0$ and $\tilde{x}_k = \tilde{x}_0$, for all k. The optimal estimate of x is independent of any measurements that are obtained. At the opposite end of the spectrum, if x is completely unknown *a priori*, then $P_0 = \infty$. The above equation for $\tilde{x}_k$ becomes,

$$
\begin{aligned}
\tilde{x}_k &= \lim_{P_0 \to \infty} \frac{P_0 + R}{kP_0 + R}\tilde{x}_{k-1} + \frac{P_0}{kP_0 + R} y_k \\
&= \frac{k-1}{k}\tilde{x}_{k-1} + \frac{1}{k} y_k \\
&= \frac{1}{k}\Big((k-1)\tilde{x}_{k-1} + y_k\Big).
\end{aligned}
$$

The right hand side of the last equation above is just the running average $\bar{y}_k = \frac{1}{k}\sum_{j=1}^{k} y_j$ of the measurements, for we have

$$
\begin{aligned}
\sum_{j=1}^{k} y_j &= \sum_{j=1}^{k-1} y_j + y_k \\
&= (k-1)\left(\frac{1}{k-1}\sum_{j=1}^{k-1} y_j\right) + y_k \\
&= (k-1)\bar{y}_{k-1} + y_k.
\end{aligned}
$$

Since $\tilde{x}_1 = \bar{y}_1$, the recurrences for $\tilde{x}_k$ and $\bar{y}_k$ are the same. Hence $\tilde{x}_k = \bar{y}_k$ for all k.

EXAMPLE 4. Suppose that a tank contains a concentration x_1 of chemical 1, and a concentration x_2 of chemical 2. We have an instrument to detect the combined concentration $x_1 + x_2$ of the two chemicals but not able to tell the values of x_1 and x_2. Chemical 2 leaks from the tank so that its concentration decreases by 1% from one measurement to the next. The measurement equation is given as

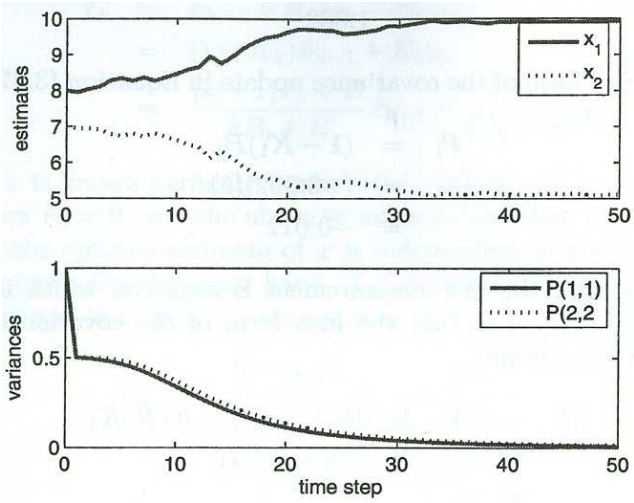
$$
y_k = x_1 + 0.99^{k-1} x_2 + \nu_k,
$$

where $H_k = (1, 0.99^{k-1})^T$, and ν_k is a random variable with zero mean and a variance $R = 0.01$.

Let the real values be $\tilde{x} = (x_1, x_2)^T = (10, 5)^T$. Suppose the initial estimates are $\tilde{x}_1 = 8$ and $\tilde{x}_2 = 7$ with P_0 equal to the identity matrix. We apply the recursive least squares algorithm. The next figure[2] shows the evolutions of the estimates $\tilde{x}_1$ and $\tilde{x}_2$, along with those of the variance of the estimation errors. It can be

[2]Figure 3.1, p. 92 of [1].

seen that after a couple dozen measurements, the estimates are getting very close to the true values 10 and 5. The variances of the estimation errors asymptotically approach zero. This means that we have increasingly more confidence in the estimates with more measurements obtained.



References

[1] D. Simon. *Optimal State Estimations.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.