Linear Least-Squares Estimation

Report By:
Prashant Tripathi (Y8104047)
Waquar Ahmad (Y8104077)

Instructor: Dr. Rajesh M. Hegde

Indian Institute of Technology Kanpur

Abstract

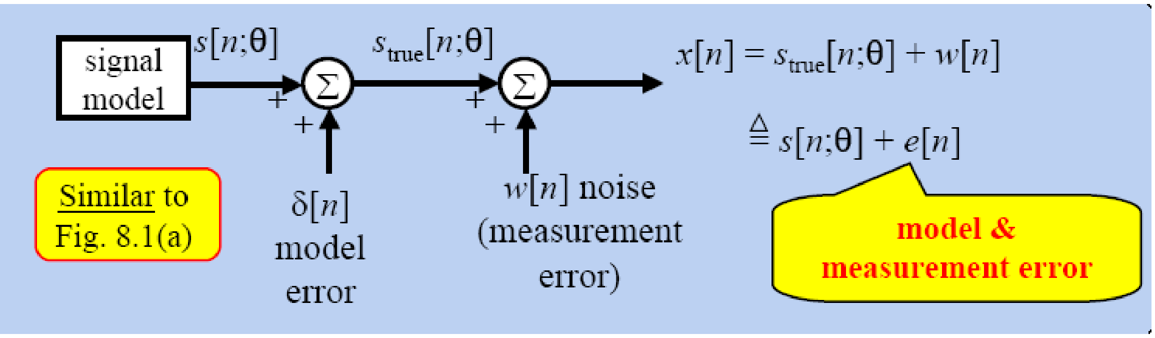
This report is about the Linear Least square approach for estimating the parameter. This method is different from other method as it involves no probabilistic model instead it involve signal model or data model and then minimizing the error function. This also tells about the geometric approach to linear least square approach. The geometric interpretation of least squares leads to the important orthogonality principle.

Contents

1. Least Square Approach

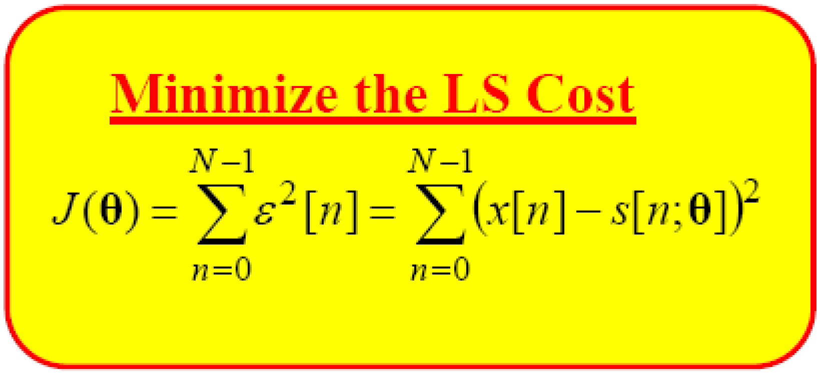
All the methods to find an optimal or nearly optimal estimator by considering the class of unbiased estimator and determining the one exhibiting minimum variance required a probabilistic model for the data i.e. they need a pdf p(x;θ). The least square approach is not statistically based. It does not need a pdf model but it need a deterministic signal model. The disadvantage of this method is that no claim can be made about optimality; the statistical performance cannot be assessed without some specific assumption about the probability structure of the data. Still, the least square estimator is widely used in practice due to its ease of implementation, amounting to the minimization of a least square error criterion.

The Block diagram below give the idea of least square approach.


signal model
$s[n;\theta]$
$s_{\text{true}}[n;\theta]$
$x[n] = s_{\text{true}}[n;\theta] + w[n]$
$\triangleq s[n;\theta] + e[n]$
+
+
+
+
Σ
Σ
Similar to Fig. 8.1(a)
$\delta[n]$ model error
$w[n]$ noise (measurement error)
model & measurement error

In Least square approach we attempt to minimize the squared difference between the given data x[n] and the assumed signal or noiseless data as shown in above figure. The signal is generated by some model which in turn depends on unknown parameter θ. The signal s[n] is purely deterministic. Due to observation noise or model inaccuracies we observe a perturbed version of s[n], which we denote by x[n].

The least square estimator of θ chooses the value that makes s[n] closest to the observed data x[n]. This closeness is measured by the LS error function or cost function given by



Minimize the LS Cost

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \varepsilon^2[n] = \sum_{n=0}^{N-1} \left(x[n] - s[n;\boldsymbol{\theta}]\right)^2$$

Thus to summarize Least Square approach we can say

1. The Value Of ‘θ’ that minimizes the J(θ) is the LSE.
2. Method is equally valid for a Gaussian as well as non- Gaussian noise.
3. It is applied when statistical characterization of data is unknown.

Now we see one example which involves the estimate of DC level in a signal.

Assuming the given signal model as s[n] = A and observed signal x[n] for n = 0,1,…,N-1.

Now using LS approach we have

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2$$

$$set \frac{\partial J(A)}{\partial A} = 0 \Rightarrow \hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] = \bar{x}$$

Thus we have seen that the estimator is the mean of the observed data. Let us take another example.

Consider the signal model s[n] = $A\cos 2\pi f_0 n$ in which the frequency to be estimated.

Assume ‘A’ is known parameter.

LSE is found by minimizing

$$J(f_0) = \sum_{n=0}^{N-1} (x[n] - A\cos 2\pi f_0 n)^2$$

In this case LS error is highly non-linear in frequency. Hence minimization cannot be done in closed form. Since the error criterion is the quadratic function of the signal, a signal that is linear in the unknown parameter yields a quadratic function for J, as in the previous example. A signal model that is a linear in the unknown parameter is said to generate linear least squares problem. Otherwise, the problem is nonlinear least squares problem.

Suppose in the previous example if ‘A’ is unknown parameter and is known then it can become a problem of linear least squares estimation. It may also be the possibility that both ‘A’ and fo are unknown and it become a case of vector parameter estimation.

Weighted Least Square Criterion:

Sometimes not all data samples are equally good:

x[0], x[1],….,x[N-1]
Say we know x[10] was poor in quality compared to other data. Thus we want to de-emphasize the x[10] in the sum of squares.

$$J(\theta)=\sum_{n=0}^{N-1} w_n (x[n]-S[n,\theta])^2$$

Depending on ‘n’ we can choose particular w_n

2. Linear Least Squares

In applying the Linear LS approach for a scalar parameter we must assume that

s[n] =θ.h[n]

where h[n] is the known sequence. The LS criterion may becomes

$$J(\theta)=\sum_{n=0}^{N-1}(x[n]-\theta h[n])^2$$

Minimization of this LS criterion produces the LSE as

$$\overset{\wedge}{\theta}=\frac{\sum_{n=0}^{N-1}x[n]h[n]}{\sum_{n=0}^{N-1}h^2[n]}$$

The minimum LS error is obtained by substituting the above equation in LS criterion equation. Thus we can directly write the equation for minimum LS error as

$$J_{\min}=\sum_{n=0}^{N-1}x^2[n]-\frac{\left(\sum_{n=0}^{N-1}x[n]h[n]\right)^2}{\sum_{n=0}^{N-1}h^2[n]}$$

The extension of these results to a vector parameter θ of dimensions pX1 is straight forward and of great utility.

For the signal **s**= {s [0], s[1],...s[N-1]}' to be linear in the unknown parameters, we assume, using matrix notation

S=Hθ

Where H is a known Nxp matrix (N>p) of full rank p. The matrix H is referred to the observation matrix. This is, of course, the linear model, albeit without the usual noise PDF assumption. The LSE is found by minimizing

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n])^2$$

$$= (X - H\theta)^T (X - H\theta)$$

$$J(\theta) = X^T X - X^T H\theta - \theta^T H^T X + \theta^T H^T H\theta$$

$$= X^T X - 2X^T H\theta + \theta^T H^T H\theta$$

Now differentiating J(θ) w.r.t θ

$$\frac{\partial J(\theta)}{\partial \theta} = -2H^T X + 2H^T H\theta$$

$$\mathit{Setting\ the\ gradient\ equal\ to\ zero\ yields\ the\ LSE}$$

$$\overset{\wedge}{\theta} = (H^T H)^{-1} H^T X$$

The Equation $H^T H\theta = H^T x$ to be solved for $\overset{\wedge}{\theta}$ are termed the normal equations. The assumed full rank of H guarantees the invertibility of $H^T H$. The minimum LS error is found as

$$J_{\min} = J(\overset{\wedge}{\theta})$$

$$= (X - H\overset{\wedge}{\theta})^T (X - H\overset{\wedge}{\theta})$$

$$= X^T (I - H(H^T H)^{-1} H^T) X$$

The last step results from the fact that I-H$(H^T H)^{-1} H^T$ is an idempotent matrix.

An extension of Linear LS problem is weighted LS

J(θ) $= (X - H\overset{\wedge}{\theta})^T W (X - H\overset{\wedge}{\theta})$

The General form of weighted LSE is readily shown to be

$$\overset{\wedge}{\theta} = (H^T W H)^{-1} H^T W X$$

And its minimum LS error is

J$_{\text{min}}$ $= X^T (W - WH(H^T WH)^{-1} H^T W) X$.

3. Geometrical interpretations

We will see now Linear LS approach using Geometrical Perspective.
Let S=Hθ be the general signal model.
If we denote columns of H by hi then we can write

$$S = [h1 \quad h2 \quad . \quad . \quad hp]\begin{bmatrix} \theta 1 \\ \theta 2 \\ . \\ . \\ \theta p \end{bmatrix}$$

$$= \sum_{i=1}^{p} \theta_i h_i$$

Signal model can be seen to be linear combination of the “signal vectors” {h1, h2 ...hp}.

Now the following points give the main reasons behind geometric interpretation.

1. Linear LS approach attempts to minimize the square of the distance from the data vector x to a signal vector.

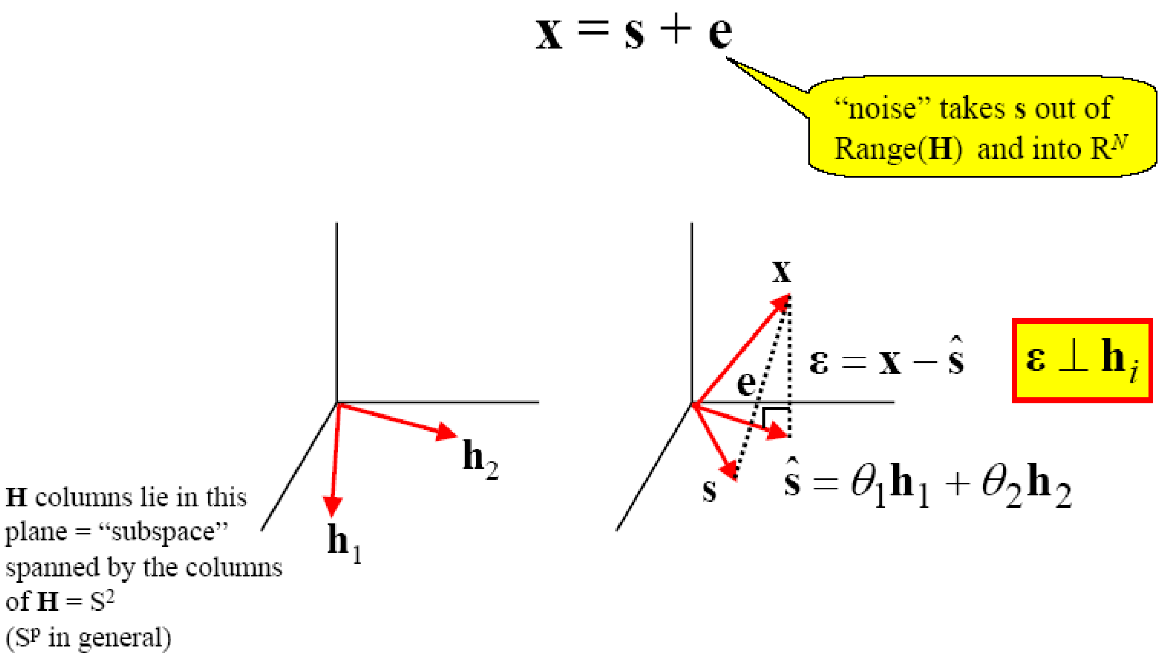
2. Data vector can lie anywhere in an N dimensional space.

3. Signal vector must lie in p-dimensional subspace of N-dimensional space.

4. Full rank of H assures that columns are linearly independent

Let us take an example with N=3 and p=2

The geometrical sketch in 3 dimensional plane can be drawn as shown below



$\mathbf{x} = \mathbf{s} + \mathbf{e}$
"noise" takes s out of Range(**H**) and into R^N
H columns lie in this plane = "subspace" spanned by the columns of **H** = S^2 (S^p in general)
$\mathbf{h}_1$
$\mathbf{h}_2$
x
e
s
$\varepsilon = \mathbf{x} - \hat{\mathbf{s}}$
$\varepsilon \perp \mathbf{h}_i$
$\hat{\mathbf{s}} = \theta_1 \mathbf{h}_1 + \theta_2 \mathbf{h}_2$

Now we know the following points

1. Error vector is orthogonal to signal subspace, then we have following relationship

$$\left(X - \hat{S}\right) \perp S^2$$

$$\Rightarrow \left(X - \hat{S}\right) \perp h1$$

$$\left(X - \hat{S}\right) \perp h2$$

2. We can use the definition of orthogonality

$$\left(X - \hat{S}\right)^T h1 = 0$$

$$\left(X - \hat{S}\right)^T h2 = 0$$

3. Using , $S = \theta_1 h_1 + \theta_2 h_2$ we can write

$$(X - H\theta)^T [h1\ h2] = 0^T$$
$$(X - H\theta)^T H = 0^T$$
$$\Rightarrow \theta = (H^T H)^{-1} H^T X$$

Finally we have arrived at the same estimate for θ hence proving that geometrical interpretation gives the same results.

Bibliography:

1. S.M. Kay, Fundamentals of Statistical Signal Processing, Vol. 1
2. http://www.aiaccess.net/english/glossaries/glosMode/e_gm_least_squares_estimation.htm.