Estimation: chapter 8

Least Squares Estimation

Lifang Feng
lffeng@ustb.edu.cn

北京科技大学
University of Science and Technology Beijing

Spring 2022

Summary

This topic will be the final topic in classical estimation theory before moving on the Bayesian estimators. So far, we have seen the following estimators:

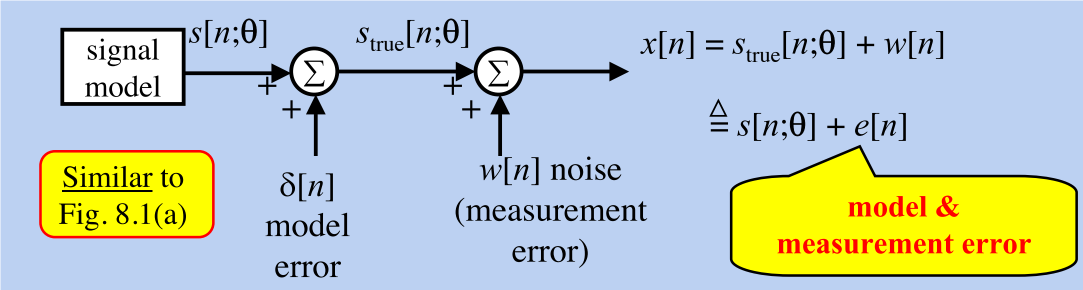
- Minimum variance unbiased estimator (MVUE): find using CRLB theorem as well as Rao-Blackwell-Lehmann-Scheffe theorem
- Efficient estimators (meet the CRLB), these are of course MVUE estimators as well and may be found using the Cramer-Rao-lower bound theorem
- Best linear unbiased estimators (BLUE): Gauss-Markov theorem
- Maximum likelihood estimator (MLE): the MLE is asymptotically efficient, and when a closed form analytic solution to the maximization problem cannot be found iterative methods (Newton-Raphson, scoring and Expectation Maximization) may be used

All need at least 2nd order statistics....

Why use LSE?

Main benefit: do NOT need a statistical model!

But we need a signal model



signal model

$s[n;\theta]$

$s_{\text{true}}[n;\theta]$

$x[n] = s_{\text{true}}[n;\theta] + w[n]$

$\triangleq s[n;\theta] + e[n]$

+ +

+ +

Similar to Fig. 8.1(a)

$\delta[n]$ model error

$w[n]$ noise (measurement error)

model & measurement error

Know what noise-free signal looks like!

No statistical performance guarantees!

What is the LSE?

Suppose we have observations/data $\mathbf{x} = [x(0), x(1), \cdots x(N-1)]^T$, and that we know what the noise-free signal should look like, i.e. we know the signal model $\mathbf{s}(\theta) = [s(0,\theta), s(1,\theta), \cdots, s(N-1,\theta)]^T$.

The least squares estimate (LSE) of θ is:

$$\hat{\theta} = \arg\min_{\theta} J(\theta) = \arg\min_{\theta} \sum_{n=0}^{N-1} (x(n) - s(n,\theta))^2 = (\mathbf{x} - \mathbf{s}(\theta))^T (\mathbf{x} - \mathbf{s}(\theta))$$

The LSE minimizes the distance or energy between the data and the signal model with the estimated value of θ. If the signal model is linear in the parameters, i.e. if

$\mathbf{s}(\theta) = \mathbf{H}\theta$ for some known matrix $\mathbf{H} \Rightarrow$ linear LSE!

In general though the signal model will be a nonlinear function of θ.

Linear LSE

Assume the signal model is $s[n] = A,\ \forall n = 0, 1, 2, \cdots, N-1$. Then the LS approach aims at minimizing

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

What happens if the noise is NOT zero mean?

Another LSE example

Assume the signal model is $s[n] = A\cos(2\pi f_0 n),\ \forall n = 0, 1, 2, \cdots, N-1$. Find the LSE of A if f_0 is known, then do the reverse.

Regression as an example of LSE

Say we are given data (x, y) : $(1, 6), (2, 5), (3, 7), (4, 10)$: and we want to find the line

$$y = \beta_1 + \beta_2 x$$

that best fits the data in the least squares sense. Find β_1 and β_2.

Linear LSE: theory

The signal model for linear LSE is

$$\mathbf{s}(\theta) = \mathbf{H}\theta$$

where θ is a $p \times 1$ vector of unknown parameters, $\mathbf{H}$ is an $N \times p$ known matrix with $N > p$ and rank p.

Linear model without any noise assumptions!

In this case, determining the linear LSE boils down to solving the following p *normal equations:*

$$\mathbf{H^T H}\theta = \mathbf{H^T x}$$

We can show the following about the LSE:

$$\hat{\theta} = \left(\mathbf{H^T H}\right)^{-1} \mathbf{H}^T\mathbf{x}$$

$$J_{min} = \mathbf{x}^T \left(\mathbf{I} - \mathbf{H}(\mathbf{H^T H})^{\mathbf{-1}}\mathbf{H^T}\right) \mathbf{x}$$

Linear LSE vs other estimators

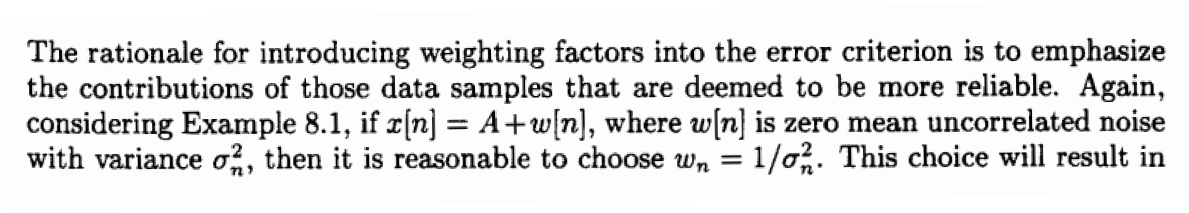
Model	Estimate
$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e}$ No Probability Model Needed	$\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$
$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ PDF Unknown, White	$\hat{\boldsymbol{\theta}}_{BLUE} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$
$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ PDF Gaussian, White	$\hat{\theta}_{ML} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$
$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ PDF Gaussian, White	$\hat{\boldsymbol{\theta}}_{MVU} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$

Weighted linear LS problem

Instead of minimizing $J(\theta) = (\mathbf{x} - \mathbf{H}\theta)^T(\mathbf{x} - \mathbf{H}\theta)$, for some $N \times N$ positive definite weighting matrix $\mathbf{W}$ we minimize

$$J(\theta) = (\mathbf{x} - \mathbf{H}\theta)^T\mathbf{W}(\mathbf{x} - \mathbf{H}\theta)$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{x} \qquad J_{\min} = \mathbf{x}^T\left(\mathbf{W} - \mathbf{W}\mathbf{H}(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\right)\mathbf{x}$$



The rationale for introducing weighting factors into the error criterion is to emphasize the contributions of those data samples that are deemed to be more reliable. Again, considering Example 8.1, if $x[n] = A + w[n]$, where $w[n]$ is zero mean uncorrelated noise with variance σ_n^2, then it is reasonable to choose $w_n = 1/\sigma_n^2$. This choice will result in

$$\hat{A} = \frac{\displaystyle\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\displaystyle\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}.$$

Geometrical interpretation of LSE

Since $\mathbf{s} = \mathbf{H}\theta$, if we call $\mathbf{h_i}$ the i-th column of $\mathbf{H}$ then it is clear that the signal $\mathbf{s} = \mathbf{H}\theta = \sum_{i=1}^{p} \theta_i \mathbf{h_i}$ lies in a p-dimensional subspace (call is S^p) of the N-dimensional space $\mathbb{R}^N$ that the signal $\mathbf{x}$ lives in. If we define

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}).$$

$$||\mathbf{y}|| = \sqrt{\sum_{i=0}^{N-1} y_i^2} = \sqrt{\mathbf{y}^T\mathbf{y}}$$

$$\begin{aligned} J(\boldsymbol{\theta}) &= ||\mathbf{x} - \mathbf{H}\boldsymbol{\theta}||^2 \\ &= ||\mathbf{x} - \sum_{i=1}^{p} \theta_i \mathbf{h}_i||^2. \end{aligned}$$

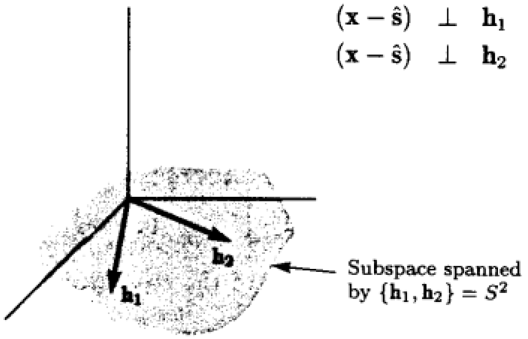
the norm in the N-dimensional Euclidean space, then finding the linear LSE can be seen as finding the vector $\mathbf{s}(\hat{\theta})$ in the subspace S^p, which is the span of the columns of $\mathbf{H}$ that lies closest to the data vector $\mathbf{x}$.

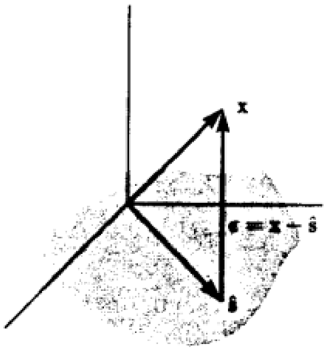
Geometrically, we want to find the projection of $\mathbf{x}$ onto the subspace S^p. The minimum distance is achieved when the error is orthogonal to the subspace S^p, which can be boiled down to the condition we already know

$$(\mathbf{x} - \mathbf{H}\theta)^T\mathbf{H} = \mathbf{0} \Rightarrow \hat{\mathbf{s}} = \mathbf{H}(\mathbf{H^T H})^{-\mathbf{1}}\mathbf{H^T x}$$

For $N = 3$ and $p = 2$ $\quad (\mathbf{x} - \hat{\mathbf{s}}) \perp S^2$

$$\begin{aligned} (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_1 \\ (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_2 \end{aligned}$$





$\boldsymbol{\epsilon} \perp S^2$

- singular

Letting $\hat{\mathbf{s}} = \theta_1\mathbf{h}_1 + \theta_2\mathbf{h}_2$, we have

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{H} = \mathbf{0}^T.$$

$$\begin{aligned} (\mathbf{x} - \theta_1\mathbf{h}_1 - \theta_2\mathbf{h}_2)^T\mathbf{h}_1 &= 0 \\ (\mathbf{x} - \theta_1\mathbf{h}_1 - \theta_2\mathbf{h}_2)^T\mathbf{h}_2 &= 0. \end{aligned}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}.$$

Referring to Figure 8.3a, if it had happened that $\mathbf{h}_1$ and $\mathbf{h}_2$ were orthogonal, then $\hat{\mathbf{s}}$ could have easily been found. This is because the component of $\hat{\mathbf{s}}$ along $\mathbf{h}_1$ or $\hat{\mathbf{s}}_1$ does not contain a component of $\hat{\mathbf{s}}$ along $\mathbf{h}_2$. If it did, then we would have the situation in Figure 8.3b. Making the orthogonality assumption and also assuming that $||\mathbf{h}_1|| = ||\mathbf{h}_2|| = 1$ (ortho*normal* vectors), we have

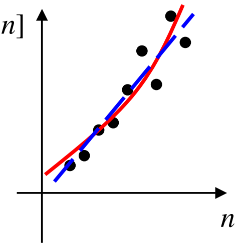
$$(\mathbf{H}^T\mathbf{H})^{-1} = (\mathbf{I})^{-1} = \mathbf{I} \qquad \hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} = \mathbf{H}^T\mathbf{x}.$$

Skipping 8.6 Order-Recursive Least Squares

Motivate this idea with *Curve Fitting*

Given data: $n = 0, 1, 2, \ldots, N\text{-}1$

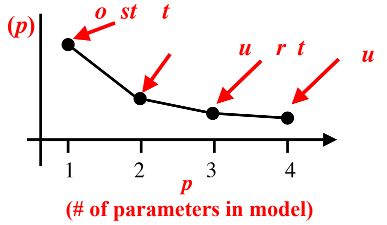
$s[0], s[1], \ldots, s[N\text{-}1]$

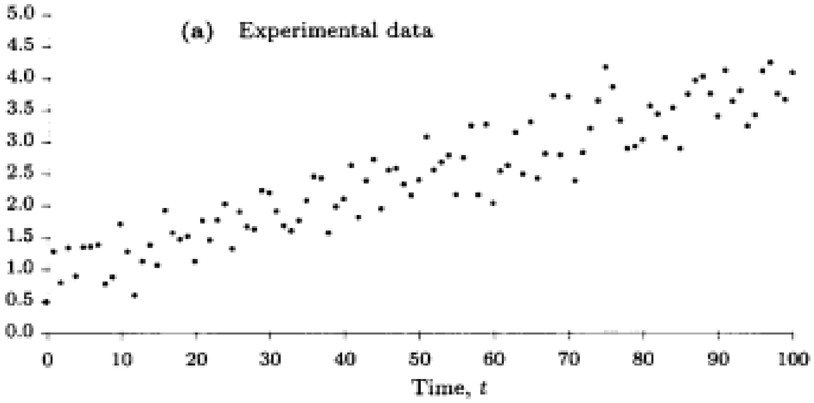


Want to fit a polynomial to data.., but which one is the right model?

- Constant
- Linear
- Quadratic
- Cubic, Etc.

Try each model, look at $J_{\min}$ … which one works "best"

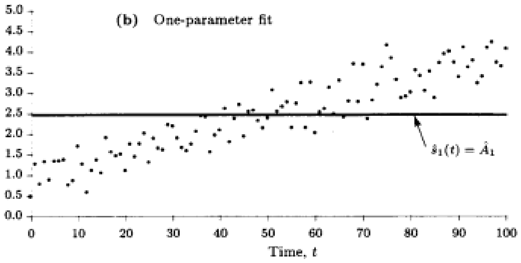




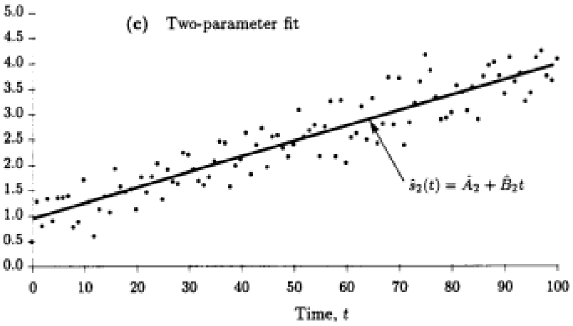
$$
\begin{aligned}
s_1[n] &= A \\
s_2[n] &= A + Bn.
\end{aligned}
$$

$$
\mathbf{H}_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \mathbf{H}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix}
$$

$$\hat{A}_1 = \bar{x}$$



$$\begin{aligned}
\hat{A}_2 &= \frac{2(2N-1)}{N(N+1)} \sum_{n=0}^{N-1} x[n] - \frac{6}{N(N+1)} \sum_{n=0}^{N-1} n x[n] \\
\hat{B}_2 &= -\frac{6}{N(N+1)} \sum_{n=0}^{N-1} x[n] + \frac{12}{N(N^2-1)} \sum_{n=0}^{N-1} n x[n].
\end{aligned}$$



A straightforward approach would compute the LSE for each model Alternatively, *order-recursive* LS approach. we *update* the LSE in order.

$$\begin{aligned}
s_1[n] &= A_1 \\
s_2[n] &= A_2 + B_2 n
\end{aligned}$$

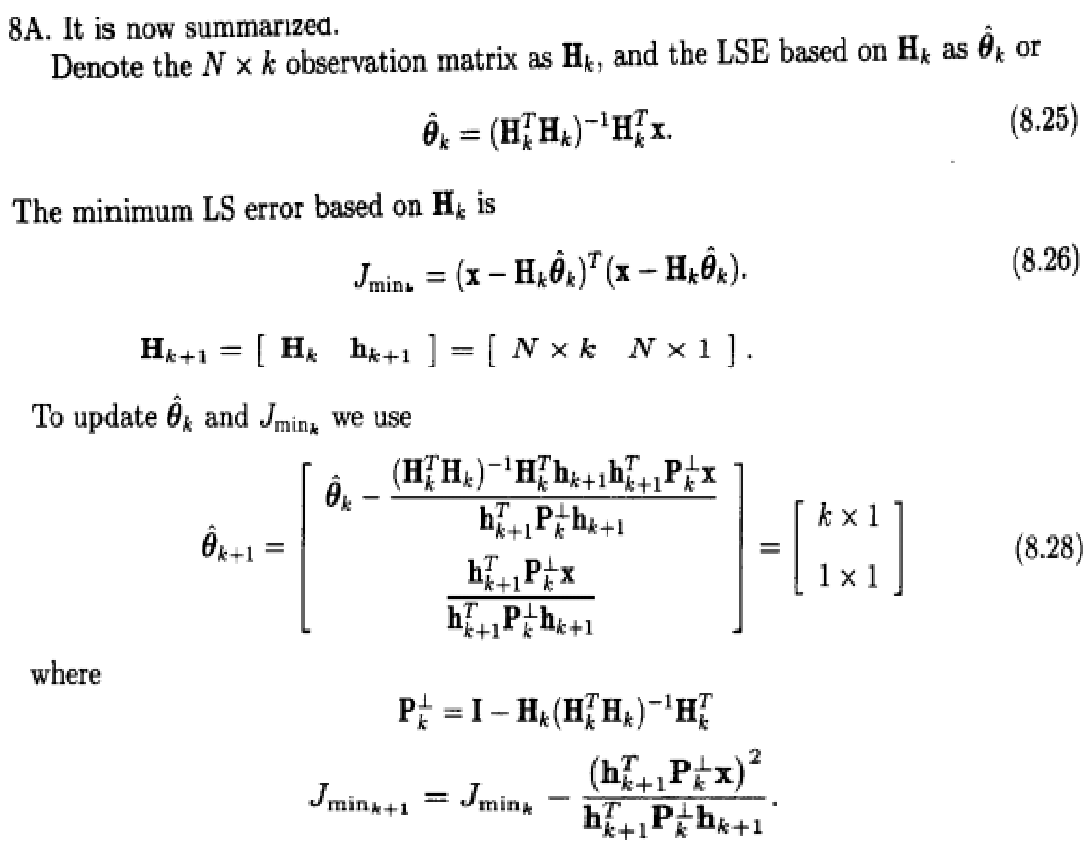
$$\mathbf{H}_2 = \begin{bmatrix} 1 & -M \\ 1 & -(M-1) \\ \vdots & \vdots \\ 1 & M \end{bmatrix}.$$

$$\hat{A}_1 = \frac{1}{2M+1} \sum_{n=-M}^{M} x[n]$$

for $-M \leq n \leq M$. We have now altered the observation interval to be the symmetric interval $[-M, M]$ as opposed to the original $[0, N-1]$ interval. The effect of this assumption is to orthogonalize the columns of $\mathbf{H}_2$ since now

$$\mathbf{H}_2^T \mathbf{H}_2 = \begin{bmatrix} 2M+1 & 0 \\ 0 & \sum_{n=-M}^{M} n^2 \end{bmatrix}$$

$$\begin{aligned}
\hat{A}_2 &= \frac{1}{2M+1} \sum_{n=-M}^{M} x[n] \\
\hat{B}_2 &= \frac{\sum_{n=-M}^{M} n x[n]}{\sum_{n=-M}^{M} n^2}.
\end{aligned}$$



8A. It is now summarized.

Denote the $N \times k$ observation matrix as $\mathbf{H}_k$, and the LSE based on $\mathbf{H}_k$ as $\hat{\boldsymbol{\theta}}_k$ or

$$\hat{\boldsymbol{\theta}}_k = (\mathbf{H}_k^T\mathbf{H}_k)^{-1}\mathbf{H}_k^T\mathbf{x}. \tag{8.25}$$

The minimum LS error based on $\mathbf{H}_k$ is

$$J_{\min_k} = (\mathbf{x} - \mathbf{H}_k\hat{\boldsymbol{\theta}}_k)^T(\mathbf{x} - \mathbf{H}_k\hat{\boldsymbol{\theta}}_k). \tag{8.26}$$

$$\mathbf{H}_{k+1} = [\ \mathbf{H}_k \quad \mathbf{h}_{k+1}\ ] = [\ N \times k \quad N \times 1\ ].$$

To update $\hat{\boldsymbol{\theta}}_k$ and $J_{\min_k}$ we use

$$\hat{\boldsymbol{\theta}}_{k+1} = \begin{bmatrix} \hat{\boldsymbol{\theta}}_k - \dfrac{(\mathbf{H}_k^T\mathbf{H}_k)^{-1}\mathbf{H}_k^T\mathbf{h}_{k+1}\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{x}}{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{h}_{k+1}} \\ \dfrac{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{x}}{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{h}_{k+1}} \end{bmatrix} = \begin{bmatrix} k \times 1 \\ 1 \times 1 \end{bmatrix} \tag{8.28}$$

where

$$\mathbf{P}_k^{\perp} = \mathbf{I} - \mathbf{H}_k(\mathbf{H}_k^T\mathbf{H}_k)^{-1}\mathbf{H}_k^T$$

$$J_{\min_{k+1}} = J_{\min_k} - \frac{(\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{x})^2}{\mathbf{h}_{k+1}^T\mathbf{P}_k^{\perp}\mathbf{h}_{k+1}}.$$

Skipping 8.7 Sequential Least Squares

In Last Section:

- Data Stays Fixed
- Model Order Increases

In This Section:

- Data Length Increases
- Model Order Stays Fixed

You have received new data sample!

Say we have $\hat{A}[N-1]$ based on $\{x[0], \ldots, x[N\text{-}1]\}$

If we get $x[N]$… can we compute $\hat{A}[N]$ based on $\hat{A}[N-1]$ and $x[N]$?
(w/o solving using full data set!)

We want… $\hat{A}[N] = f(\hat{A}[N-1], x[N])$

Consider Example 8.1 in which the DC signal level is to be estimated. We saw that the LSE is

$$\hat{A}[N-1] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

$$\hat{A}[N] = \frac{1}{N+1}\sum_{n=0}^{N} x[n].$$

$$\begin{aligned}\hat{A}[N] &= \frac{1}{N+1}\left(\sum_{n=0}^{N-1} x[n] + x[N]\right)\\ &= \frac{N}{N+1}\hat{A}[N-1] + \frac{1}{N+1}x[N].\end{aligned}$$

$$\hat{A}[N] = \hat{A}[N-1] + \frac{1}{N+1}\left(x[N] - \hat{A}[N-1]\right).$$

The new estimate is equal to the old one plus a *correction* term. The correction term decreases with N, reflecting the fact that the estimate $\hat{A}[N-1]$ is based on many more data samples and therefore should be given more weight. Also, $x[N] - \hat{A}[N-1]$ can be thought of as the error in predicting $x[N]$ by the previous samples, which are summarized by $\hat{A}[N-1]$. If this error is zero, then no correction takes place for that update. Otherwise, the new estimate differs from the old one.

The minimum LS error may also be computed recursively. Based on data samples up to time $N-1$, the error is

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

and thus using (8.36)

$$J_{\min}[N] = \sum_{n=0}^{N}(x[n] - \hat{A}[N])^2$$

$$J_{\min}[N] = J_{\min}[N-1] + \frac{N}{N+1}(x[N] - \hat{A}[N-1])^2.$$

Skipping 8.8 Constrained Least Squares

Why Constrain? Because sometimes we know (or believe!) certain values are not allowed for θ

For example: In emitter location you may know that the emitter's range can't exceed the "radio horizon"

You may also know that the emitter is on the left side of the aircraft (because you got a strong signal from the left-side antennas and a weak one from the right-side antennas)

Thus, when finding $\hat{\theta}_{LS}$ you want to constrain it to satisfy these conditions

Say that S_c is the set of allowable θ values (due to constraints).

Then we seek $\hat{\theta}_{CLS} \in S_c$ such that

$$\left\|\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}_{CLS}\right\|^2 = \min_{\boldsymbol{\theta}\in S_c}\left\|\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\right\|^2$$

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b} \qquad J_c = (\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}) + \boldsymbol{\lambda}^T(\mathbf{A}\boldsymbol{\theta}-\mathbf{b})$$

$$\frac{\partial J_c}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} + \mathbf{A}^T\boldsymbol{\lambda}$$

$$\begin{aligned}\hat{\boldsymbol{\theta}}_c &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} - \frac{1}{2}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\boldsymbol{\lambda}\\ &= \hat{\boldsymbol{\theta}} - (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\frac{\boldsymbol{\lambda}}{2}\end{aligned}$$

$$\mathbf{A}\boldsymbol{\theta}_c = \mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\frac{\boldsymbol{\lambda}}{2} = \mathbf{b}$$

$$\frac{\boldsymbol{\lambda}}{2} = \left[\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\right]^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}-\mathbf{b}).$$

$$\hat{\boldsymbol{\theta}}_c = \hat{\boldsymbol{\theta}} - (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\left[\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\right]^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}-\mathbf{b})$$

Example 8.8 - Constrained Signal

If the signal model is

$$s[n] = \begin{cases} \theta_1 & n=0 \\ \theta_2 & n=1 \\ 0 & n=2 \end{cases}$$

and we observe $\{x[0], x[1], x[2]\}$, then the observation matrix is

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Observe that the signal vector

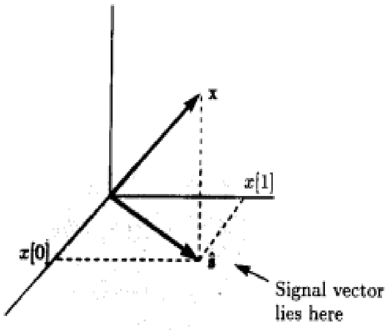
$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ 0 \end{bmatrix}$$

must lie in the plane shown in Figure 8.11a. The unconstrained LSE is

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \end{bmatrix}$$

and the signal estimate is

$$\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}} = \begin{bmatrix} x[0] \\ x[1] \\ 0 \end{bmatrix}.$$



(a) Unconstrained least squares

As shown in Figure 8.11a, this is intuitively reasonable. Now assume that we know a priori that $\theta_1 = \theta_2$. In terms of (8.50) we have

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \boldsymbol{\theta} = 0$$

so that $\mathbf{A} = [1 \ -1]$ and $\mathbf{b} = 0$. Noting that $\mathbf{H}^T\mathbf{H} = \mathbf{I}$, we have from (8.52)

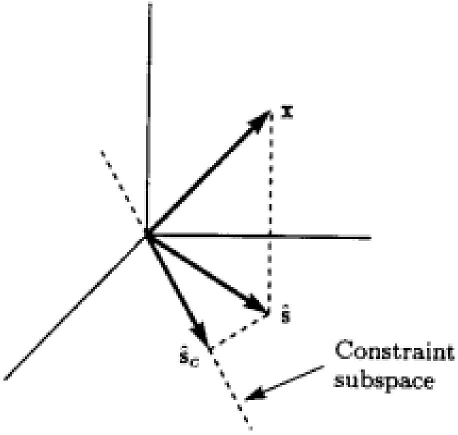
$$\begin{aligned} \hat{\boldsymbol{\theta}}_c &= \hat{\boldsymbol{\theta}} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\hat{\boldsymbol{\theta}} \\ &= \left[\mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\right]\hat{\boldsymbol{\theta}}. \end{aligned}$$

After some simple algebra this becomes

$$\hat{\boldsymbol{\theta}}_c = \begin{bmatrix} \frac{1}{2}(x[0]+x[1]) \\ \frac{1}{2}(x[0]+x[1]) \end{bmatrix}$$

and the constrained signal estimate becomes

$$\hat{\mathbf{s}}_c = \mathbf{H}\hat{\boldsymbol{\theta}}_c = \begin{bmatrix} \frac{1}{2}(x[0]+x[1]) \\ \frac{1}{2}(x[0]+x[1]) \\ 0 \end{bmatrix}$$



(b) Constrained least squares

Non-linear LSE

In general though, the signal model will not be linear in the unknown parameters. In this case, we wish to minimize

$$J(\theta) = (\mathbf{x} - \mathbf{s}(\theta))^T (\mathbf{x} - \mathbf{s}(\theta))$$

Differentiating (or taking the gradient) with respect to θ, we obtain the following nonlinear equations to be solved:

$$\frac{\partial \mathbf{s}(\theta)^T}{\partial \theta} (\mathbf{x} - \mathbf{s}(\theta)) = \mathbf{0}$$

We can sometimes use the tricks 1) transform the parameters so that the signal model is linear in the new parameters via a 1-1 mapping 2) separate the parameters into linear and non-linear parts. But in general we need to solve that equation, which may be done iteratively if it is impossible in closed form, via say the Newtown-Raphson method.

Transformation of parameters

Let $\alpha = \mathbf{g}(\theta)$ where $\mathbf{g}(\cdot)$ is a p-dimensional function whose inverse exists. Ideally, want to find $\mathbf{g}(\cdot)$ such that

$$\mathbf{s}(\theta(\alpha)) = \mathbf{s}(\mathbf{g}^{-1}(\alpha)) = \mathbf{H}\alpha \text{ is linear!!! Then what?}$$

$$\hat{\boldsymbol{\alpha}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}.$$

$$\hat{\boldsymbol{\theta}} = \mathbf{g}^{-1}(\hat{\boldsymbol{\alpha}})$$

Example: estimate the phase ϕ and amplitude $A > 0$ in the least squares sense if the signal model is $s[n] = A\cos(2\pi f_0 n + \phi)$.

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

$$\begin{aligned} \alpha_1 &= A\cos\phi \\ \alpha_2 &= -A\sin\phi, \end{aligned}$$

$$\begin{aligned} A &= \sqrt{\alpha_1^2 + \alpha_2^2} \\ \phi &= \arctan\left(\frac{-\alpha_2}{\alpha_1}\right) \end{aligned}$$

$$A\cos(2\pi f_0 n + \phi) = A\cos\phi\cos 2\pi f_0 n - A\sin\phi\sin 2\pi f_0 n$$

$$s[n] = \alpha_1 \cos 2\pi f_0 n + \alpha_2 \sin 2\pi f_0 n. \qquad \mathbf{s} = \mathbf{H}\boldsymbol{\alpha}$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ \cos 2\pi f_0 & \sin 2\pi f_0 \\ \vdots & \vdots \\ \cos 2\pi f_0(N-1) & \sin 2\pi f_0(N-1) \end{bmatrix}$$

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{A} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} \sqrt{\hat{\alpha}_1^2 + \hat{\alpha}_2^2} \\ \arctan\left(\frac{-\hat{\alpha}_2}{\hat{\alpha}_1}\right) \end{bmatrix}$$

Separability of parameters

The signal model may be linear in some of the unknown parameters - which helps. The signal model is called separable if

$$\mathbf{s} = \mathbf{H}(\alpha)\beta, \quad \theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

This model is linear in $\boldsymbol{\beta}$ but nonlinear in $\boldsymbol{\alpha}$

Then what? $J(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{x} - \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta})^T (\mathbf{x} - \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta})$ $\quad \hat{\boldsymbol{\beta}} = \left(\mathbf{H}^T(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})\right)^{-1} \mathbf{H}^T(\boldsymbol{\alpha})\mathbf{x}$

$J(\boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}) = \mathbf{x}^T \left[\mathbf{I} - \mathbf{H}(\boldsymbol{\alpha}) \left(\mathbf{H}^T(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})\right)^{-1} \mathbf{H}^T(\boldsymbol{\alpha})\right] \mathbf{x}$. The problem now reduces to a *maximization* of

$$\mathbf{x}^T \mathbf{H}(\boldsymbol{\alpha}) \left(\mathbf{H}^T(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})\right)^{-1} \mathbf{H}^T(\boldsymbol{\alpha})\mathbf{x}$$

Example: Find the LSE of $\{A_1, A_2, A_3, r\}$ if $0 < r < 1$ in the signal model

$$s[n] = A_1 r^n + A_2 r^{2n} + A_3 r^{3n}$$

the model is linear in the amplitudes $\boldsymbol{\beta} = [A_1\, A_2\, A_3]^T$, and nonlinear in the damping factor $\boldsymbol{\alpha} = r$. Using (8.54), the nonlinear LSE is obtained by maximizing

$$\mathbf{x}^T \mathbf{H}(r) \left(\mathbf{H}^T(r)\mathbf{H}(r)\right)^{-1} \mathbf{H}^T(r)\mathbf{x}$$

over $0 < r < 1$, where

$$\mathbf{H}(r) = \begin{bmatrix} 1 & 1 & 1 \\ r & r^2 & r^3 \\ \vdots & \vdots & \vdots \\ r^{N-1} & r^{2(N-1)} & r^{3(N-1)} \end{bmatrix}.$$

Once $\hat{r}$ is found we have the LSE for the amplitudes

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{H}^T(\hat{r})\mathbf{H}(\hat{r})\right)^{-1} \mathbf{H}^T(\hat{r})\mathbf{x}.$$

This maximization is easily carried out on a digital computer.