

EE6617: 05 Least Square Estimation (LSE)

Major references:

- *Fundamentals of Statistical Processing, Volume I: Estimation Theory*, 1/E, by Steven M. Kay, Prentice Hall, 1993
- *Statistical Signal Processing*, 1/E, by Louis L. Scharf, Prentice Hall, 1991

The **LS estimator** dates back to 1795 when **Gauss** used the method to study planetary motions.



The **LS estimator**

- has **no optimality properties**, e.g., minimum variance, minimum MSE, etc., associated with the classical estimators
- makes good sense for many **practical** problems of interests

Often we know nothing about $p(\mathbf{x}; \theta)$. The LS estimator assume the signal model

$$\mathbf{x} = \mathbf{s}(\theta) + \mathbf{w}, \text{ with } E[\mathbf{w}] = \mathbf{0}.$$

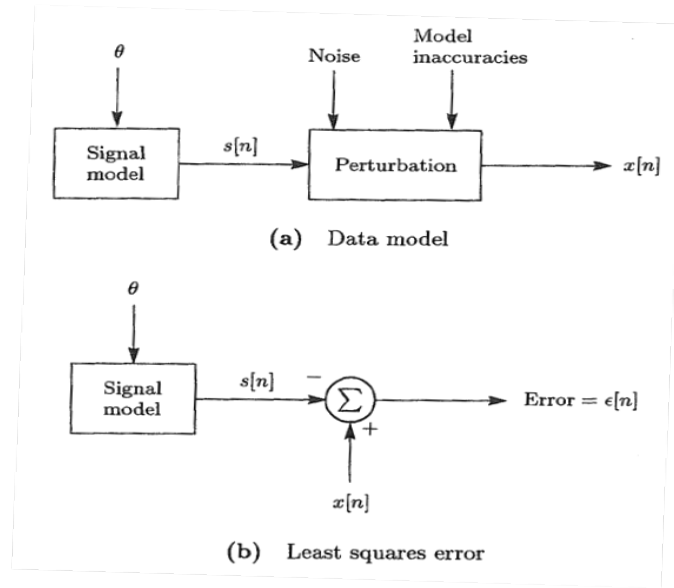
The signal $\mathbf{s}(\theta)$ is a function of θ and the **linear LSE assumes**

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

These are all necessary components to design LSE. Thus, we

- do **not know the distribution of $\mathbf{w}$** (i.e., LSE does not care Gaussian?, Correlated?) except that **$\mathbf{w}$ is zero mean**
- do **not know the likelihood function $p(\mathbf{x}; \theta)$**

1 Least Square Criterion



Least Square Principle

Minimize squared signal error to find $\hat{\theta}$ such that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} J(\theta) \quad \text{s.t.} \quad \mathbf{H}\theta = \mathbf{s}(\theta)$$

where $\mathbf{H} \in \mathbb{R}^{N \times p}$ with $N \geq p$ and the objective function (or cost function)

$$J(\theta) = \|\mathbf{x} - \mathbf{s}(\theta)\|^2 = \|\mathbf{x} - \mathbf{H}\theta\|^2.$$

Signal Model $\mathbf{s}(\theta) = \mathbf{h}\theta$ with Vector $\mathbf{h} \in \mathbb{R}^{N \times 1}$

The signal model

$$\mathbf{x} = \mathbf{h}\theta + \mathbf{w}, \quad E[\mathbf{w}] = \mathbf{0}.$$

The least square error is

$$\begin{aligned} J(\theta) &= \|\mathbf{x} - \mathbf{h}\theta\|^2 \\ &= \mathbf{x}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{h}\theta + \theta^2 \mathbf{h}^T \mathbf{h}. \end{aligned}$$

Thus, **LSE design criterion** is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{h}\theta\|^2.$$

Evaluate critical point

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta} &= -2\mathbf{x}^T \mathbf{h} + 2\theta \mathbf{h}^T \mathbf{h} \\ \frac{\partial^2 J(\theta)}{\partial^2 \theta} &= 2\mathbf{h}^T \mathbf{h} = 2\|\mathbf{h}\|^2 \geq 0 \Rightarrow \text{convex} \\ &\Rightarrow -2\mathbf{x}^T \mathbf{h} + 2\theta \mathbf{h}^T \mathbf{h} = 0 \\ &\Rightarrow \hat{\theta} = \frac{\mathbf{h}^T \mathbf{x}}{\mathbf{h}^T \mathbf{h}} = \frac{\mathbf{h}^T}{\|\mathbf{h}\|^2} \mathbf{x}.\end{aligned}$$

This is

- (a) **efficient MVU when $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.**
- (b) **BLUE when $E[\mathbf{w}] = \mathbf{0}$ and $E[\mathbf{w}\mathbf{w}^T] = \sigma^2 \mathbf{I}$.**

How to Measure the LSE Performance?

The minimum cost function can be defined

$$\begin{aligned}J_{\min}(\hat{\boldsymbol{\theta}}) &= \mathbf{x}^T \mathbf{x} - 2\mathbf{h}^T \mathbf{x} \frac{\mathbf{h}^T \mathbf{x}}{\|\mathbf{h}\|^2} + \frac{(\mathbf{h}^T \mathbf{x})^2 \|\mathbf{h}\|^2}{\|\mathbf{h}\|^4} \\ &= \mathbf{x}^T \mathbf{x} - 2 \frac{(\mathbf{h}^T \mathbf{x})^2}{\|\mathbf{h}\|^2} + \frac{(\mathbf{h}^T \mathbf{x})^2}{\|\mathbf{h}\|^2} \\ &= \mathbf{x}^T \mathbf{x} - \frac{\mathbf{x}^T \mathbf{h} \mathbf{h}^T \mathbf{x}}{\|\mathbf{h}\|^2} \\ &= \mathbf{x}^T \left(\mathbf{I} - \frac{\mathbf{h} \mathbf{h}^T}{\|\mathbf{h}\|^2} \right) \mathbf{x}\end{aligned}$$

Can you find $\mathbf{x}$ that makes $J_{\min}(\hat{\boldsymbol{\theta}}) = 0$?

Signal Model $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$ with Matrix $\mathbf{H} \in \mathbb{R}^{N \times p}$

The **LSE design criterion** is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2$$

for $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, with $E[\mathbf{w}] = \mathbf{0}$ and $\mathbf{H} \in \mathbb{R}^{N \times p}$, $N \geq p$ and $\text{rank}(\mathbf{H}) = p$. Evaluate critical point

$$\begin{aligned}\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= -2\mathbf{H}^T \mathbf{x} + 2\mathbf{H}^T \mathbf{H} \boldsymbol{\theta} \\ \frac{\partial^2 J(\boldsymbol{\theta})}{\partial^2 \boldsymbol{\theta}} &= 2\mathbf{H}^T \mathbf{H} \geq 0 \Rightarrow \text{positive definite} \Rightarrow \text{convex} \\ &\Rightarrow -2\mathbf{H}^T \mathbf{x} + 2\mathbf{H}^T \mathbf{H} \boldsymbol{\theta} = \mathbf{0} \\ &\Rightarrow \hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}.\end{aligned}$$

This is

- (a) **efficient MVU when $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.**
- (b) **BLUE when $E[\mathbf{w}] = \mathbf{0}$ and $E[\mathbf{w}\mathbf{w}^T] = \sigma^2 \mathbf{I}$.**

The minimum cost function is defined

$$\begin{aligned}J_{\min}(\hat{\boldsymbol{\theta}}) &= \mathbf{x}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{H} \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}^T \mathbf{H}^T \mathbf{H} \hat{\boldsymbol{\theta}} \\ &= \mathbf{x}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} + \mathbf{x}^T \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \\ &= \mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \\ &= \mathbf{x}^T \left(\mathbf{I} - \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \right) \mathbf{x}\end{aligned}$$

Can you find $\mathbf{x}$ that makes $J_{\min}(\hat{\boldsymbol{\theta}}) = 0$?

2 Weighted Least Square Criterion

Sometimes the **errors in particular entries of $\mathbf{x} - \mathbf{s}$ are more important than others**. In order to take account for this, weight the contribution of the error

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

where $\mathbf{Y} \in \mathbb{S}^{N \times N}$ is positive definite and symmetry.

By the positive definite and symmetry, we have by default

$$(\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T \mathbf{Y} (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) \geq 0$$

and if $\mathbf{Y} = \text{diag}(y_1, y_2, \dots, y_N)$,

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

The LSE is designed by solving

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T \mathbf{Y} (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})),$$

where $\mathbf{s}(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$. For instance, if the error of the first parameter, i.e., $x_1 - s_1$ has massive impact than the other parameters, one might prefer to allocate larger weights to $y_{11}, y_{12}, \dots, y_{1N}$, and $y_{21}, \dots, y_{N1}$, meaning to further mitigate the error terms stemmed by $x_1 - s_1$.

Let $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}$ and

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \quad E[\mathbf{w}] = \mathbf{0}.$$

LSE design criterion is

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} J(\boldsymbol{\theta}) \text{ s.t. } \mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta} \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{Y} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}). \end{aligned}$$

Now,

$$\begin{aligned} J(\boldsymbol{\theta}) &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{Y} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \\ &= \mathbf{x}^T \mathbf{Y} \mathbf{x} - 2\mathbf{x}^T \mathbf{Y} \mathbf{H} \boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{Y} \mathbf{H} \boldsymbol{\theta}, \end{aligned}$$

evaluate critical point

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= -2\mathbf{H}^T \mathbf{Y} \mathbf{x} + 2\mathbf{H}^T \mathbf{Y} \mathbf{H} \boldsymbol{\theta} \\ \frac{\partial^2 J(\boldsymbol{\theta})}{\partial^2 \boldsymbol{\theta}} &= 2\mathbf{H}^T \mathbf{Y} \mathbf{H} \geq 0 \Rightarrow \text{positive definite} \Rightarrow \text{convex} \\ &\Rightarrow -2\mathbf{H}^T \mathbf{Y} \mathbf{x} + 2\mathbf{H}^T \mathbf{Y} \mathbf{H} \boldsymbol{\theta} = \mathbf{0} \\ &\Rightarrow \hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{Y} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{Y} \mathbf{x}. \end{aligned}$$

Then, the minimum cost function is

$$\begin{aligned} J_{\min}(\hat{\boldsymbol{\theta}}) &= \mathbf{x}^T \mathbf{Y} \mathbf{x} - 2\mathbf{x}^T \mathbf{Y} \mathbf{H} (\mathbf{H}^T \mathbf{Y} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{Y} \mathbf{x} + \mathbf{x}^T \mathbf{Y} \mathbf{H} (\mathbf{H}^T \mathbf{Y} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{Y} \mathbf{x} \\ &= \mathbf{x}^T \mathbf{Y} \mathbf{x} - \mathbf{x}^T \mathbf{Y} \mathbf{H} (\mathbf{H}^T \mathbf{Y} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{Y} \mathbf{x} \\ &= \mathbf{x}^T \left(\mathbf{Y} - \mathbf{Y} \mathbf{H} (\mathbf{H}^T \mathbf{Y} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{Y} \right) \mathbf{x}. \end{aligned}$$

Can you find $\mathbf{x}$ that makes $J_{\min}(\hat{\boldsymbol{\theta}}) = 0$?

3 Least Square Geometrical Interpretation

To minimize $J(\boldsymbol{\theta}) = \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2$, we find $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$. The error vector is

$$\begin{aligned} \mathbf{e} &= \mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}} \\ &= \mathbf{x} - \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \\ &= \left(\mathbf{I} - \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \right) \mathbf{x}. \end{aligned}$$

Consider to compute

$$\begin{aligned}\mathbf{H}^T \mathbf{e} &= \mathbf{H}^T \mathbf{x} - \mathbf{H}^T \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \\ &= \mathbf{H}^T \mathbf{x} - \mathbf{H}^T \mathbf{x} = \mathbf{0}.\end{aligned}\tag{1}$$

What Does This Mean?

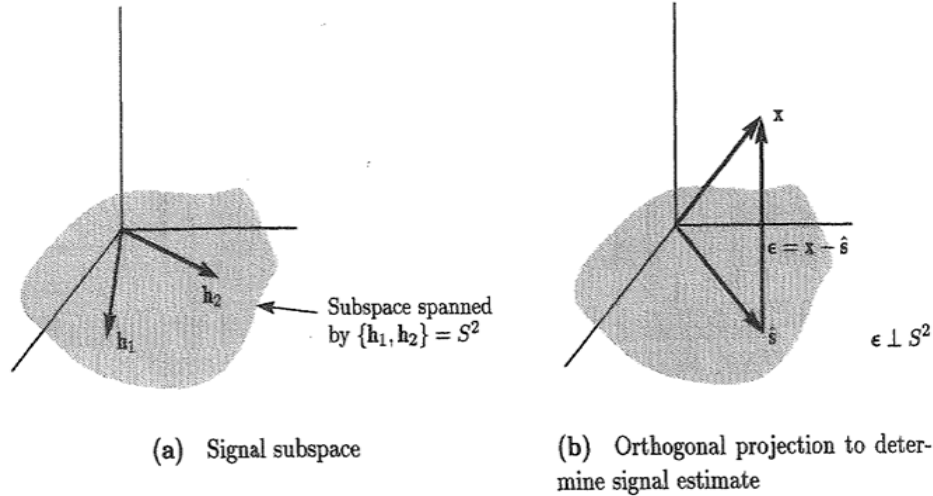
This means that the **error vector $\mathbf{e}$ is orthogonal to the columns of $\mathbf{H}$.**

Why?

Estimating $\boldsymbol{\theta}$ is same as estimating $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$, because the estimator knows $\mathbf{H} \in \mathbb{R}^{N \times p}$, $N \geq p$. We write $\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}}$. Then the error vector $\mathbf{e} = \mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{x} - \hat{\mathbf{s}}$. For this construction, $\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}}$ must lie in the column space of $\mathbf{H}$, i.e., $\hat{\mathbf{s}} \in \mathcal{R}(\mathbf{H})$.

Hence, the (1) means that **the error $\mathbf{e}$ is orthogonal to columns of $\mathbf{H}$ and is the one that minimizes $\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 = \|\mathbf{x} - \mathbf{s}\|^2 = \|\mathbf{e}\|^2$, i.e., minimizes its norm.**

For instance, assume $p = 2$ and $N = 3$. We have $\mathbf{H} = [\mathbf{h}_1 \ \mathbf{h}_2]$. In order that $\mathbf{x} - \hat{\mathbf{s}}$ is orthogonal to $\mathcal{R}(\mathbf{H})$, $\hat{\mathbf{s}}$ must be the projection of $\mathbf{x}$ onto the subspace created by $\mathbf{H}$, i.e., $\mathcal{R}(\mathbf{H})$ and this is the minimum norm solution. This has an intuitive geographical interpretation as below.



Optimal Projection

Want to optimally estimate the projection of $\mathbf{x}$ onto $\mathcal{R}(\mathbf{H})$. Assume $\text{rank}(\mathbf{H}) = p$. Let the singular value decomposition (SVD) of $\mathbf{H} \in \mathbb{R}^{N \times p}$ as

$$\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

where $\mathbf{U} \in \mathbb{R}^{N \times N}$ satisfies $\mathbf{U}^T \mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_N$, $\mathbf{V} \in \mathbb{R}^{p \times p}$ satisfies $\mathbf{V}^T \mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$, and $\boldsymbol{\Sigma}$ is the singular value matrix. The $[\mathbf{U}]_{1,p} = \bar{\mathbf{U}}$ is called left singular matrix (orthonormal basis of

$\mathcal{R}(\mathbf{H})$ of $\mathbf{H}$, the $\mathbf{V}$ is right singular matrix (orthonormal basis of $\mathcal{R}(\mathbf{H}^T)$), and we denote the numbers σ_i as the singular values. The singular value decomposition can be rewritten

$$\mathbf{H} = [\bar{\mathbf{U}} | [\mathbf{U}]_{p+1:N}] \begin{bmatrix} \bar{\Sigma} \\ \mathbf{0}_{(N-p) \times p} \end{bmatrix} \mathbf{V}^T = \sum_{i=1}^p \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

where $\bar{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_p)$, $\mathbf{u}_i$ is the i th column of $\mathbf{U}$, and $\mathbf{v}_i$ is the i th column of $\mathbf{V}$.

It can be shown that the LSE $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$ satisfies

$$\hat{\mathbf{s}} = \mathbf{H} \hat{\boldsymbol{\theta}} = \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} = \underbrace{\bar{\mathbf{U}} \bar{\mathbf{U}}^T}_{\text{projection onto } \mathcal{R}(\mathbf{H})} \mathbf{x}.$$

This can be interpreted that

- (a) $\bar{\mathbf{U}}^T \mathbf{x}$ projects $\mathbf{x}$ to obtain weights in each orthonormal direction
- (b) $\bar{\mathbf{U}} \bar{\mathbf{U}}^T \mathbf{x}$ use weights to linearly combine columns of $\bar{\mathbf{U}}$ to reconstruct $\hat{\mathbf{s}}$.

4 Constrained Least Square Estimation

With LS what we did was to find $\boldsymbol{\theta}$ that minimizes

$$\begin{aligned} J(\boldsymbol{\theta}) &= (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) \\ &= (\mathbf{x} - \underbrace{\mathbf{H}\boldsymbol{\theta}}_{\text{linear}})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}). \end{aligned}$$

We solved this assuming $\boldsymbol{\theta}$ could be any vector in $\mathbb{R}^p$. However, in practice, you might need to satisfy additional constraints on $\boldsymbol{\theta}$. Talk about one case of $\boldsymbol{\theta}$ being **constrained linearly**,

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$$

where $\mathbf{A} \in \mathbb{R}^{r \times p}$ and $\mathbf{b} \in \mathbb{R}^{r \times 1}$.

Linearly constrained least square problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \quad \text{s.t.} \quad \mathbf{A}\boldsymbol{\theta} = \mathbf{b}. \quad (2)$$

We can set-up the **Lagrangian objective**

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

where $\lambda \in \mathbb{R}^{r \times 1}$ is the Lagrange vector. The first and second order derivatives are

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} &= -2\mathbf{H}^T \mathbf{x} + 2\mathbf{H}_T^T \mathbf{H}\boldsymbol{\theta} + \mathbf{A}^T \lambda \\ \frac{\partial^2 J(\boldsymbol{\theta}, \lambda)}{\partial^2 \boldsymbol{\theta}} &= 2\mathbf{H}_T^T \mathbf{H} \geq 0 \Rightarrow \text{convex} \\ &\Rightarrow \text{the minimizer solves } -2\mathbf{H}^T \mathbf{x} + 2\mathbf{H}_T^T \mathbf{H}\boldsymbol{\theta} + \mathbf{A}^T \lambda = \mathbf{0} \\ &\Rightarrow \hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H} \mathbf{x} - \frac{1}{2} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T \lambda. \end{aligned} \quad (3)$$

Since $\hat{\boldsymbol{\theta}}$ must satisfies the linear constraint

$$\begin{aligned}\frac{\partial J(\boldsymbol{\theta}, \lambda)}{\partial \lambda} &= \mathbf{A}\boldsymbol{\theta} = \mathbf{b} \\ \Rightarrow \mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}\mathbf{x} - \frac{1}{2}\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\lambda &= \mathbf{b} \\ \Rightarrow \lambda &= 2\left(\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\right)^{-1}\left(\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}\mathbf{x} - \mathbf{b}\right).\end{aligned}$$

Then, plugging λ in (3),

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

If we set $(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}\mathbf{x} = \hat{\boldsymbol{\theta}}_{uc}$, then

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{uc} - (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\left(\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\right)^{-1}\left(\mathbf{A}\hat{\boldsymbol{\theta}}_{uc} - \mathbf{b}\right).$$

5 Sequential Least Squares

Usually, data is taken from sampling a continuous signal $x(t)$ and we normally observe this **casually**. Thus, we can only use samples from $x(t)$ with $t \leq t_0$.

The **sequential LS** concerns **how to update the estimator at time N given the computed LS estimation at time $N - 1$** . Assume sample n is

$$x[n] = \mathbf{h}^T[n]\boldsymbol{\theta} + w[n]$$

where $E[w[n]] = 0$ and $\mathbf{C}[N] = E[\mathbf{w}[N]\mathbf{w}^T[N]] = \text{diag}(\sigma_0^2, \sigma_1^2, \dots, \sigma_N^2)$, i.e., noises are uncorrelated, where $\mathbf{w}[N] = [w[0], w[1], \dots, w[N]]^T$.

Assume we have collected N samples

$$\begin{aligned}\mathbf{x}[N-1] &= [x[0] \cdots x[N-1]]^T \\ \mathbf{H}[N-1] &= \begin{bmatrix} \mathbf{h}^T[0] \\ \mathbf{h}^T[1] \\ \vdots \\ \mathbf{h}^T[N-1] \end{bmatrix}\end{aligned}$$

$$\mathbf{C}[N-1] = E[\mathbf{w}[N-1]\mathbf{w}^T[N-1]] = \text{diag}(\sigma_0^2, \sigma_1^2, \dots, \sigma_{N-1}^2) = \begin{bmatrix} \sigma_0^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix}.$$

Estimate $\boldsymbol{\theta}$ to minimize

$$J(\boldsymbol{\theta}) = (\mathbf{x}[N-1] - \mathbf{H}[N-1]\boldsymbol{\theta})^T \mathbf{C}^{-1}[N-1] (\mathbf{x}[N-1] - \mathbf{H}[N-1]\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} (x[n] - \mathbf{h}^T[n]\boldsymbol{\theta})^2.$$

We already know the solution is

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

and we know

$$\begin{aligned} \mathbf{C}_{\hat{\boldsymbol{\theta}}} &= E \left[\left(\hat{\boldsymbol{\theta}}[N-1] - \boldsymbol{\theta} \right) \left(\hat{\boldsymbol{\theta}}[N-1] - \boldsymbol{\theta} \right)^T \right] \\ &= (\mathbf{H}^T[N-1]\mathbf{C}^{-1}[N-1]\mathbf{H}[N-1])^{-1} \\ &= \boldsymbol{\Sigma}[N-1]. \end{aligned}$$

What if we add another data sample?

$$\begin{aligned} \mathbf{x}[N] &= \begin{bmatrix} \mathbf{x}[N-1] \\ x[N] \end{bmatrix} \\ \mathbf{H}[N] &= \begin{bmatrix} \mathbf{H}[N-1] \\ \mathbf{h}^T[N] \end{bmatrix} \\ \mathbf{C}[N] &= \begin{bmatrix} \mathbf{C}[N-1] & 0 \\ 0 & \sigma_N^2 \end{bmatrix}. \end{aligned}$$

Now we solve to minimize

$$J(\boldsymbol{\theta}) = \sum_{n=0}^N \frac{1}{\sigma_n^2} (x[n] - \mathbf{h}^T[n]\boldsymbol{\theta})^2.$$

We know the solution is

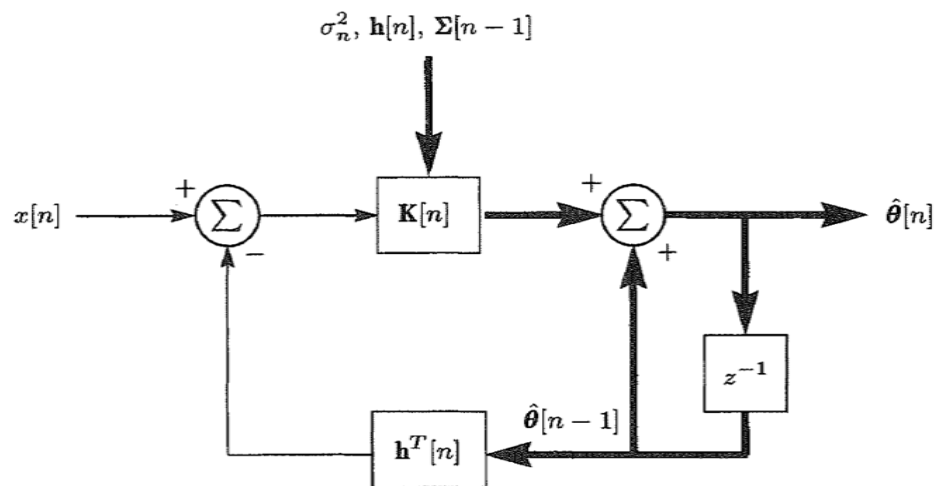
$$\begin{aligned} \hat{\boldsymbol{\theta}}[N] &= (\mathbf{H}^T[N]\mathbf{C}^{-1}[N]\mathbf{H}[N])^{-1} \mathbf{H}^T[N]\mathbf{C}^{-1}[N]\mathbf{x}[N] \\ &= \hat{\boldsymbol{\theta}}[N-1] + \bar{k}[N] \left(x[N] - \mathbf{h}^T[N]\hat{\boldsymbol{\theta}}[N-1] \right) \end{aligned} \quad (4)$$

where

$$\bar{k}[N] = \frac{\boldsymbol{\Sigma}[N-1]\mathbf{h}[N]}{\sigma_N^2 + \mathbf{h}^T[N]\boldsymbol{\Sigma}[N-1]\mathbf{h}[N]} \quad (5)$$

$$\boldsymbol{\Sigma}[N] = (\mathbf{I} - \bar{k}[N]\mathbf{h}^T[N]) \boldsymbol{\Sigma}[N-1]. \quad (6)$$

The (4) and (5) represent the **estimator update** and (6) represents the **covariance update**.
No matrix inversions are required for the sequential LS.



6 Appendix

6.1 Derivation of Sequential Least Squares

In this appendix we derive (4), (5), and (6).

$$\begin{aligned}
\hat{\boldsymbol{\theta}}[N] &= (\mathbf{H}^T[N] \mathbf{C}^{-1}[N] \mathbf{H}[N])^{-1} \mathbf{H}^T[N] \mathbf{C}^{-1}[N] \mathbf{x}[N] \\
&= \left([\mathbf{H}^T[N-1] \ \mathbf{h}[N]] \begin{bmatrix} \mathbf{C}^{-1}[N-1] & 0 \\ 0 & \sigma_N^2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{H}[N-1] \\ \mathbf{h}^T[N] \end{bmatrix} \right)^{-1} \\
&\quad \left([\mathbf{H}^T[N-1] \ \mathbf{h}[N]] \begin{bmatrix} \mathbf{C}^{-1}[N-1] & 0 \\ 0 & \sigma_N^2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}[N-1] \\ x[N] \end{bmatrix} \right) \\
&= \left(\mathbf{H}^T[N-1] \mathbf{C}^{-1}[N-1] \mathbf{H}[N-1] + \frac{1}{\sigma_n^2} \mathbf{h}[N] \mathbf{h}^T[N] \right)^{-1} \left(\mathbf{H}^T[N-1] \mathbf{C}^{-1}[N-1] \mathbf{x}[N-1] + \frac{1}{\sigma_n^2} \mathbf{h}[N] x[N] \right)
\end{aligned}$$

We substitute the covariance matrix of $\hat{\boldsymbol{\theta}}[N-1]$, $\boldsymbol{\Sigma}[N-1] = (\mathbf{H}^T[N-1] \mathbf{C}^{-1}[N-1] \mathbf{H}[N-1])^{-1}$ and

$$\hat{\boldsymbol{\theta}}[N] = \left(\boldsymbol{\Sigma}^{-1}[N-1] + \frac{1}{\sigma_n^2} \mathbf{h}[N] \mathbf{h}^T[N] \right)^{-1} \left(\mathbf{H}^T[N-1] \mathbf{C}^{-1}[N-1] \mathbf{x}[N-1] + \frac{1}{\sigma_n^2} \mathbf{h}[N] x[N] \right).$$

Notice that

$$\begin{aligned}
\boldsymbol{\Sigma}[N] &= (\mathbf{H}^T[N] \mathbf{C}^{-1}[N] \mathbf{H}[N])^{-1} \\
&= \left(\boldsymbol{\Sigma}^{-1}[N-1] + \frac{1}{\sigma_n^2} \mathbf{h}[N] \mathbf{h}^T[N] \right)^{-1} \\
&= \boldsymbol{\Sigma}[N-1] - \frac{\boldsymbol{\Sigma}[N-1] \mathbf{h}[N] \mathbf{h}^T[N] \boldsymbol{\Sigma}[N-1]}{\sigma_n^2 + \mathbf{h}^T[N] \boldsymbol{\Sigma}[N-1] \mathbf{h}[N]} \\
&= (\mathbf{I} - \bar{k}[N] \mathbf{h}^T[N]) \boldsymbol{\Sigma}[N-1]
\end{aligned} \tag{7}$$

where

$$\bar{k}[N] = \frac{\boldsymbol{\Sigma}[N-1] \mathbf{h}[N]}{\sigma_n^2 + \mathbf{h}^T[N] \boldsymbol{\Sigma}[N-1] \mathbf{h}[N]}$$

and the third equality follow from the matrix inversion lemma

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{DA}^{-1} \mathbf{B} + \mathbf{C}^{-1})^{-1} \mathbf{DA}^{-1}.$$

The (7) shows (6). Applying (7) and using $\hat{\boldsymbol{\theta}}[N-1] = \boldsymbol{\Sigma}[N-1] \mathbf{H}^T[N-1] \mathbf{C}^{-1}[N-1] \mathbf{x}[N-1]$, we have

$$\begin{aligned}
\hat{\boldsymbol{\theta}}[N] &= (\mathbf{I} - \bar{k}[N] \mathbf{h}^T[N]) \boldsymbol{\Sigma}[N-1] \left(\boldsymbol{\Sigma}^{-1}[N-1] \hat{\boldsymbol{\theta}}[N-1] + \frac{1}{\sigma_n^2} \mathbf{h}[N] x[N] \right) \\
&= \hat{\boldsymbol{\theta}}[N-1] - \bar{k}[N] \mathbf{h}^T[N] \hat{\boldsymbol{\theta}}[N-1] + \frac{1}{\sigma_n^2} \boldsymbol{\Sigma}[N-1] \mathbf{h}[N] x[N] - \frac{1}{\sigma_n^2} \bar{k}[N] \mathbf{h}^T[N] \boldsymbol{\Sigma}[N-1] \mathbf{h}[N] x[N].
\end{aligned}$$

Then, using the relation $\mathbf{\Sigma}[N-1]\mathbf{h}[N] = \bar{k}[N](\sigma_n^2 + \mathbf{h}^T[N]\mathbf{\Sigma}[N-1]\mathbf{h}[N])$, we have $\frac{1}{\sigma_n^2}\mathbf{\Sigma}[N-1]\mathbf{h}[N]x[N] = \frac{1}{\sigma_n^2}(\sigma_n^2\bar{k}[N] + \bar{k}[N]\mathbf{h}^T[N]\mathbf{\Sigma}[N-1]\mathbf{h}[N])$. Then,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}[N] &= \hat{\boldsymbol{\theta}}[N-1] - \bar{k}[N]\mathbf{h}^T[N]\hat{\boldsymbol{\theta}}[N-1] + \frac{1}{\sigma_n^2}\mathbf{\Sigma}[N-1]\mathbf{h}[N]x[N] - \frac{1}{\sigma_n^2}\bar{k}[N]\mathbf{h}^T[N]\mathbf{\Sigma}[N-1]\mathbf{h}[N]x[N] \\
&= \hat{\boldsymbol{\theta}}[N-1] - \bar{k}[N]\mathbf{h}^T[N]\hat{\boldsymbol{\theta}}[N-1] + \bar{k}[N]x[N] \\
&= \hat{\boldsymbol{\theta}}[N-1] + \bar{k}[N]\left(x[N] - \mathbf{h}^T[N]\hat{\boldsymbol{\theta}}[N-1]\right).
\end{aligned}$$