I225E Statistical Signal Processing

10. Least Squares Estimation

MAWALIM and UNOKI

candylim@jaist.ac.jp and unoki@jaist.ac.jp

School of Information Science

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
HOKURIKU 1990

Least Squares Estimation

What if the data distribution is unknown?

$\Longrightarrow$ **Least Squares Estimation**

[Features]

1. No probabilistic assumption on data
2. Assumption only on signal model
3. No guaranty for optimality
4. Performance cannot be assessed without assumption about probabilistic structure of data
5. Widely applied, due to its ease of implementation

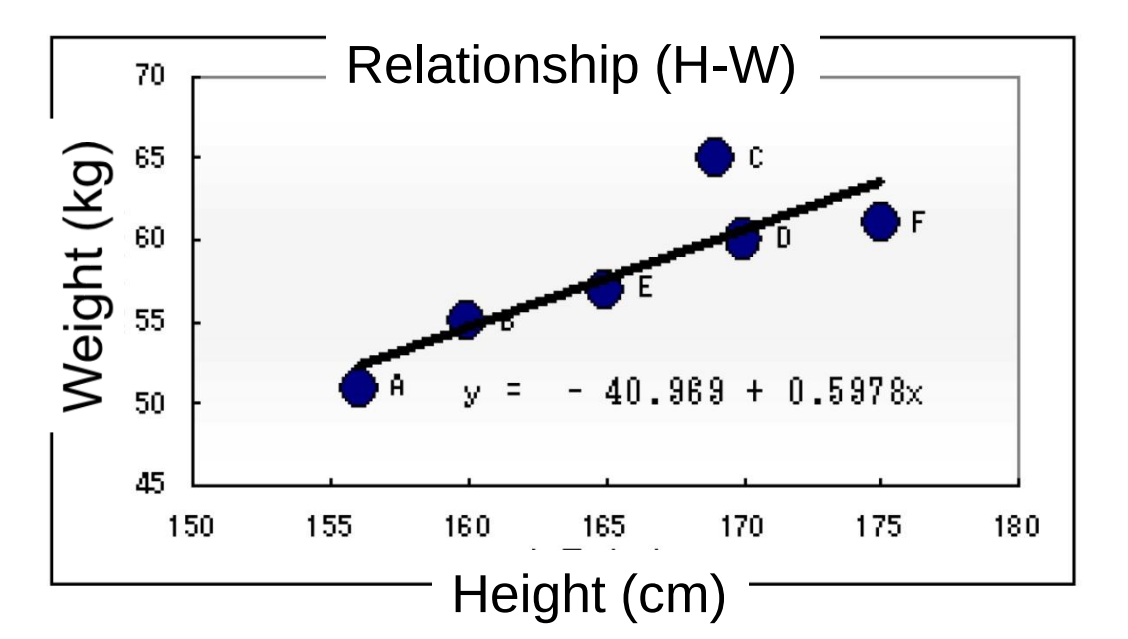
Least Squares Estimation

- Linear least squares
- Orthogonality principle
- Matrix inversion lemma
- Order-recursive least squares
- Sequential least squares

Idea is simple:

With respect to observed data $s[n]$ and their estimation $\hat{s}[n;\theta]$, find coefficients θ of the estimation function in such a way that they minimize the squared error:

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



3. Linear Least Squares

Estimation function constructed by linear summation of observed data:

$$\hat{s}[n;\theta] = \sum_{m=1}^{p} H_{nm}\theta_m$$

In vector formula, $\hat{\boldsymbol{s}} = \boldsymbol{H}\theta$

- Linear coefficients: $\theta = \left[\theta_1, \theta_2, \cdots, \theta_p\right]^T$
- Estimation: $\hat{\boldsymbol{s}} = [\hat{s}[0], \hat{s}[1], \cdots, \hat{s}[N-1]]^T$
- Observation: $\boldsymbol{s} = [s[0], s[1], \cdots, s[N-1]]^T$
- Observation matrix: $\boldsymbol{H} = |H_{nm}|_{n=0,\cdots,N-1;m=1,\cdots,p}$

Least squares error is

$$J(\theta) = \sum_{n=0}^{N-1}(s[n] - \hat{s}[n;\theta])^2$$
$$= (\boldsymbol{s} - \boldsymbol{H}\theta)^T(\boldsymbol{s} - \boldsymbol{H}\theta)$$
$$= \boldsymbol{s}^T\boldsymbol{s} - \boldsymbol{s}^T\boldsymbol{H}\theta - \theta^T\boldsymbol{H}^T\boldsymbol{s} + \theta^T\boldsymbol{H}^T\boldsymbol{H}\theta$$
$$= \boldsymbol{s}^T\boldsymbol{s} - 2\theta^T\boldsymbol{H}^T\boldsymbol{s} + \theta^T\boldsymbol{H}^T\boldsymbol{H}\theta$$

The least squares solution is

$$\frac{\partial J}{\partial \theta} = -2\boldsymbol{H}^T s + 2\boldsymbol{H}^T\boldsymbol{H}\theta = 0$$
$$\Rightarrow \qquad \hat{\theta} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{s}$$

If $\boldsymbol{H} \in \Re^{N\times p}$ is of full rank p, $\boldsymbol{H}^T\boldsymbol{H}$ is invertible and the least squares solution can be obtained.

The least squares error is

$$J_{min} = J(\hat{\theta})$$
$$= \left(\boldsymbol{s} - \boldsymbol{H}\hat{\theta}\right)^T \left(\boldsymbol{s} - \boldsymbol{H}\hat{\theta}\right)$$
$$= (\boldsymbol{s} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{s})^T(\boldsymbol{s} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{s})$$
$$= \boldsymbol{s}^T(\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T)^T(\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T)\boldsymbol{s}$$
$$= \boldsymbol{s}^T(\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T)\boldsymbol{s}$$
$$(\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T)^2 = (\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T)$$

This can be further simplified as

$$J_{min} = \boldsymbol{s}^T\boldsymbol{s} - \boldsymbol{s}^T\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{s}$$
$$= \boldsymbol{s}^T\left(\boldsymbol{s} - \boldsymbol{H}\hat{\theta}\right).$$

4. Orthogonality Principle

Decomposing the observation matrix $\boldsymbol{H}$ into column vectors $\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_p}$.

$$s = \left[\boldsymbol{h}_1\ \boldsymbol{h}_2 \cdots \boldsymbol{h}_p\right] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} = \sum_{i=1}^{p} \theta_i \boldsymbol{h}_i$$

Error vector is

$$\epsilon = \boldsymbol{s} - \boldsymbol{H}\theta = s - \sum_{i=1}^{p} \theta_i \boldsymbol{h}_i$$

The least squares error and the local minimum condition are

$$J(\theta) = \|\epsilon\|^2 = \|\boldsymbol{s} - H\theta\|^2 = \left\|\boldsymbol{s} - \sum_{i=1}^{p} \theta_i \boldsymbol{h}_i\right\|^2$$

$$\frac{\partial J}{\partial \theta_i} = -2\boldsymbol{h}_i^T\left(\boldsymbol{s} - \sum_{i=1}^{p} \theta_i \boldsymbol{h}_i\right)$$

$$= -2\boldsymbol{h}_i^T \epsilon = 0$$

$$\rightarrow \ \epsilon \perp \boldsymbol{h}_i$$

If coefficients θ_i are chosen in such a way that error vector ϵ is orthogonal to observation vectors $\boldsymbol{h_1}, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_p$, the least squares error J is minimum.

Inversion of partitioned matrix

- $(p+q) \times (p+q)$ matrix $\mathbf{M}$ partitioned as:

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}, \mathbf{A} \in \mathrm{R}^{p \times p}, \mathbf{B} \in \mathrm{R}^{p \times q}, \mathbf{C} \in \mathrm{R}^{q \times p}, \mathbf{D} \in \mathrm{R}^{q \times q}$$

- Inversion of partitioned matrix:

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{pmatrix}^{-1}$$

- (Exercise) Prove by:

$$\mathbf{M}^{-1}\mathbf{M} = \mathbf{M}\mathbf{M}^{-1} = \mathbf{I}_{p+q}$$

Matrix inversion lemma

- Suppose matrices:

$$\mathbf{A} \in \mathrm{R}^{p \times p}, \mathbf{B} \in \mathrm{R}^{p \times q}, \mathbf{C} \in \mathrm{R}^{q \times q}, \mathbf{D} \in \mathrm{R}^{q \times p}$$

- Matrix inversion lemma

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

- (Exercise): Prove the lemma by showing:

$$\left(\mathbf{A}+\mathbf{BCD}\right)\left(\mathbf{A}+\mathbf{BCD}\right)^{-1} = \left(\mathbf{A}+\mathbf{BCD}\right)^{-1}\left(\mathbf{A}+\mathbf{BCD}\right) = \mathbf{I}_p$$

5. Order-Recursive Least Squares

Denote observation matrix as $\boldsymbol{H}(k)$ $(\in R^{N\times k})$ and observed data as $\boldsymbol{s}$ $(\in R^{N})$. kth order least square estimation based on $\boldsymbol{H}(k)$ is given by

$$\hat{s}[n] = \sum_{m=1}^{k} H_{nm}(k)\theta_m(k).$$

Suppose a column is added to the observation matrix

$$\boldsymbol{H}(k+1) = [\boldsymbol{H}(k) \quad \boldsymbol{h}_{k+1}] = [N\times k \quad N\times 1]$$

Then, $(k+1)$th order least square estimation

$$\hat{s}[n] = \sum_{m=1}^{k+1} H_{nm}(k+1)\theta_m(k+1)$$

can be updated via the following recursive formula

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

Here, $\boldsymbol{P}_k^{\perp} = \boldsymbol{I} - \boldsymbol{H}(k)\{\boldsymbol{H}^T(k)\boldsymbol{H}(k)\}^{-1}\boldsymbol{H}^T(k)$.

The minimum least square error is updated as

$$J_{min}(k+1) = J_{min}(k) - \frac{\left(\boldsymbol{h}_{k+1}^T\boldsymbol{P}_k^{\perp}\boldsymbol{s}\right)^2}{\boldsymbol{h}_{k+1}^T\boldsymbol{P}_k^{\perp}\boldsymbol{h}_{k+1}}$$

Moreover, inverse matrix, denoted as $\boldsymbol{D}_k = (\boldsymbol{H}^T(k)\boldsymbol{H}(k))^{-1}$, $\boldsymbol{D}_k$ can be updated as follows

$$\boldsymbol{D}_{k+1} = \begin{bmatrix} \boldsymbol{D}_k + \dfrac{\boldsymbol{D}_k \boldsymbol{H}^T(k)\boldsymbol{h}_{k+1}\boldsymbol{h}_{k+1}^T\boldsymbol{H}(k)\boldsymbol{D}_k}{\boldsymbol{h}_{k+1}^T\boldsymbol{P}_k^{\perp}\boldsymbol{h}_{k+1}} & -\dfrac{\boldsymbol{D}_k \boldsymbol{H}^T(k)\boldsymbol{h}_{k+1}}{\boldsymbol{h}_{k+1}^T\boldsymbol{P}_k^{\perp}\boldsymbol{h}_{k+1}} \\ -\dfrac{\boldsymbol{h}_{k+1}^T\boldsymbol{H}_k\boldsymbol{D}_k}{\boldsymbol{h}_{k+1}^T\boldsymbol{P}_k^{\perp}\boldsymbol{h}_{k+1}} & \dfrac{1}{\boldsymbol{h}_{k+1}^T\boldsymbol{P}_k^{\perp}\boldsymbol{h}_{k+1}} \end{bmatrix}$$

[Procedure]

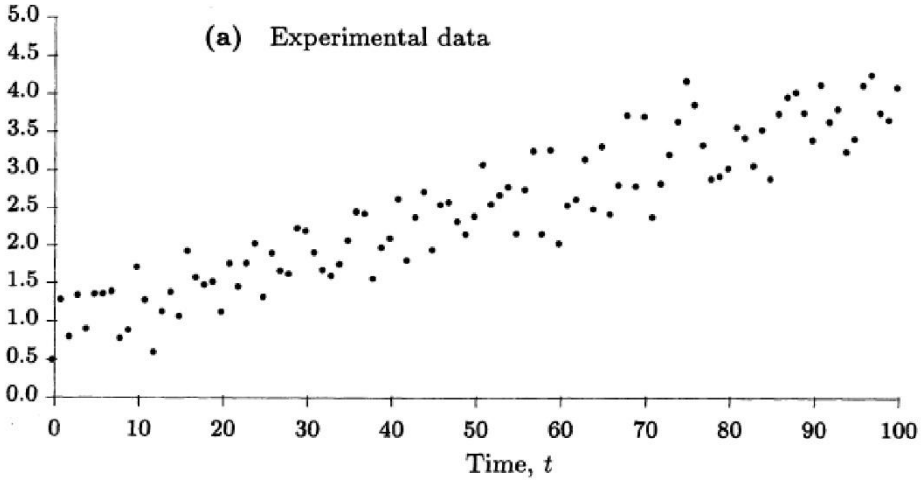
1. If $k = 1$,

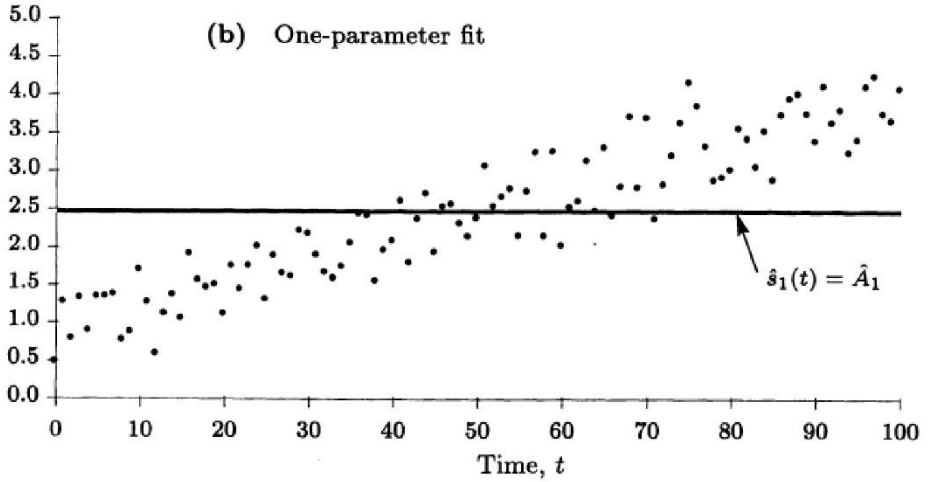
$$\hat{\theta}(1) = \{\boldsymbol{H}^T(1)\boldsymbol{H}(1)\}^{-1}\boldsymbol{H}^T(1)\boldsymbol{s}$$

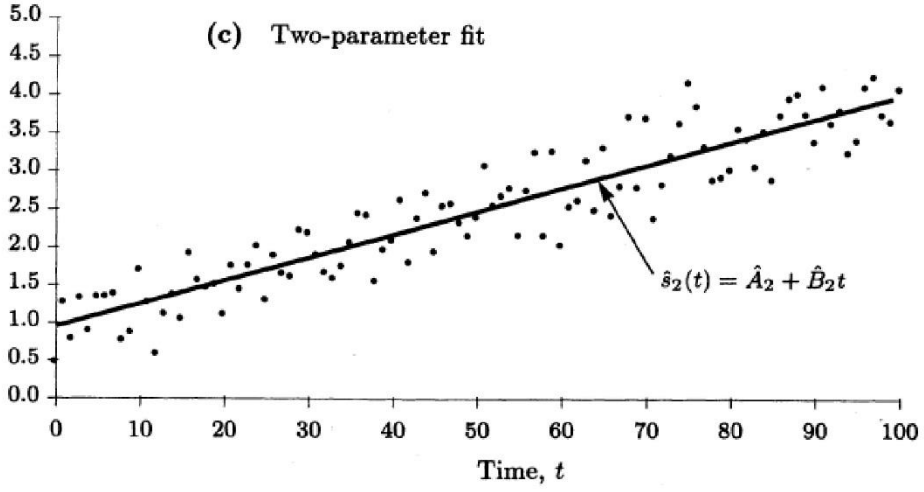
$$J_{min}(1) = \{\boldsymbol{s} - \boldsymbol{H}(1)\hat{\theta}(1)\}^T\{s - H(1)\hat{\theta}(1)\}$$

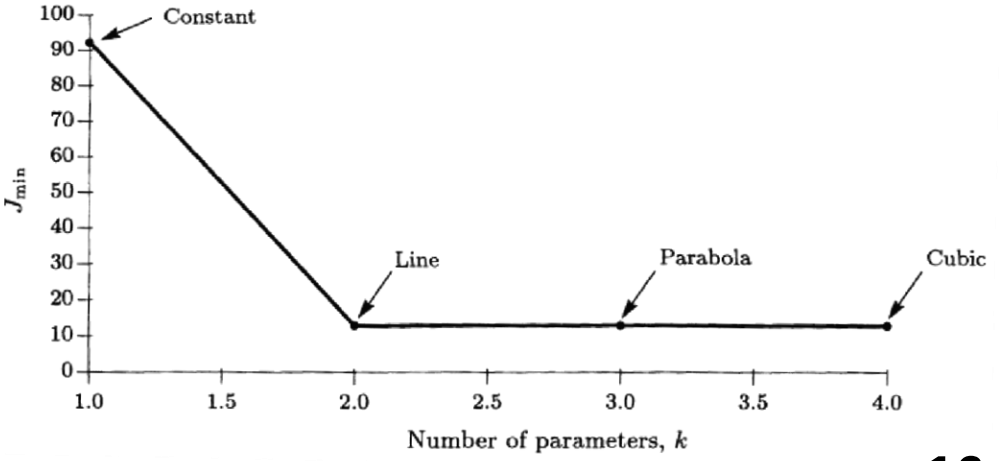
2. As the order is increased ($k \rightarrow k+1$), $\hat{\theta}(k), J_{min}(k), D_k$ are updated accordingly.

Example: Curve Fitting









6. Sequential Least Squares

For observation data $\boldsymbol{s}(n-1)$ $(\in R^n)$ and observation matrix $\boldsymbol{H}(n-1)$ $(\in R^{n\times p})$, consider the linear estimation:

$$\hat{\boldsymbol{s}} = \boldsymbol{H}(n-1)\theta.$$

We denote the least square solution as $\hat{\theta}(n-1)$. Suppose that additional data is observed as

$$s(n) = [s[0]\ s[1]\ \cdots s[n-1]\ s[n]]^T$$

$$H(n) = \begin{bmatrix} H(n-1) \\ h^T(n) \end{bmatrix} = \begin{bmatrix} n\times p \\ 1\times p \end{bmatrix}$$

Then, the least square solution $\hat{\theta}(n)$ can be updated sequentially as follows.

Estimator Update:

$$\hat{\theta}(n) = \hat{\theta}(n-1) + \boldsymbol{K}(n)\left\{s[n] - \boldsymbol{h}^T(n)\hat{\theta}(n-1)\right\}$$

$$\boldsymbol{K}(n) = \frac{\boldsymbol{\Sigma}(n-1)\boldsymbol{h}(n)}{\sigma_n^2 + \boldsymbol{h}^T(n)\boldsymbol{\Sigma}(n-1)\boldsymbol{h}(n)}$$

Variance Update:

$$\boldsymbol{\Sigma}(n) = \{\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{h}^T(n)\}\boldsymbol{\Sigma}(n-1)$$

Error Update:

$$J_{min}(n) = J_{min}(n-1) + \frac{\left\{s[n] - \boldsymbol{h}^T(n)\hat{\theta}(n-1)\right\}^2}{\sigma_n^2 + \boldsymbol{h}^T(n)\boldsymbol{\Sigma}(n-1)\boldsymbol{h}(n)}$$

- Noise $s - \hat{s}$ are independent; their covariance matrix has $\mathrm{diag}(\sigma_0^2, \sigma_1^2, \cdots, \sigma_n^2)$.
- $\boldsymbol{\Sigma}$ stands for covariance matrix of $\hat{\theta}(\boldsymbol{\Sigma} = \boldsymbol{C}_\theta)$.
- No need to compute inverse matrix.
- On-line process is possible.

Sequential least squares for DC level in white Gaussian noise