Chapter 8. Method of Least Squares

8.1 Statement of the linear least-square estimation Problem

Consider a physical phenomenon that is characterized by two sets of variables, $d(i)$ and $u(i)$. The variable $d(i)$ is observed at time i in response to the subset of variables $u(i)$, $u(i-1)$, $\ldots$, $u(i-M+1)$, applied as inputs. That is, $d(i)$ is a function of the input $u(i)$, $u(i-1)$, $\ldots$, $u(i-M+1)$. This functional relationship is hypothesized to be linear, such that

$$d(i) = \sum_{k=0}^{M-1} w_{ok}^{*}\, u(i-k) + e_o(i),$$

where w_{ok} are the unknown parameters of the model and $e_o(i)$ represents the measurement error. The unobservable measurement error $e_o(i)$ is white with zero mean and variance σ^2, such that

$$E[e_o(i)] = 0 \quad \forall i$$

and
$$E[e_o(i)\, e_o^*(k)] = \begin{cases} \sigma^2, & i = k \\ 0, & i \neq k. \end{cases}$$

Therefore, $E[d(i)] = \sum_{k=0}^{M-1} w_{ok}^* u(i-k)$,

where $u(i), u(i-1) \cdots u(i-M+1)$ are deterministic.

The problem we have to solve is to estimate the unknown parameters w_{ok} of the multiple linear regression model, given $u(i)$ and $d(i)$, $i = 1, 2, \ldots, N$. This can be achieved by applying a linear transversal filter to estimate the optimal weights $w_\emptyset, w_1, \ldots, w_{M-1}$ such that

$$e(i) = d(i) - y(i)$$

where $y(i) = \sum_{k=0}^{M-1} w_k^* u(i-k)$ is

the output of the filter.

In the method of least squares, we choose the tap weights w_k of the transversal filter so as to minimize a cost function that consists of the sum of error squares,

$$\mathcal{E}(w_0, \cdots w_{M-1}) = \sum_{i=i_1}^{i_2} |e(i)|^2,$$

where i_1 and i_2 define the index limits at which the error minimization occurs; this sum may also be viewed as an error energy.

8.2 Data Windowing

Given M as the number of tap weights used in the transversal filter model. A rectangular matrix constructed from the input data $u(1), u(2), \cdots u(N)$ may have four different kinds of forms:

1. The covariance method, $i_1 = M$, $i_2 = N$

$$\tilde{A}^H = \begin{bmatrix} u(M) & u(M+1) & \cdots & u(N) \\ u(M-1) & u(M) & \cdots & u(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ u(1) & u(2) & \cdots & u(N-M+1) \end{bmatrix}$$

2. The autocorrelation method, $i_1 = 1$, $i_2 = N+M-1$,

$$\tilde{A}^H = \begin{bmatrix} u(1) & u(2) & \cdots & u(M) & \cdots & u(N) & 0 & \cdots & 0 \\ 0 & u(1) & \cdots & u(M-1) & \cdots & u(N-1) & u(N) & \ddots & \vdots \\ \vdots & 0 & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \; u(1) & \cdots & u(N-M+1) & u(N-M) & \cdots & u(N) \end{bmatrix}$$

3. The prewindowing method, $i_1 = 1$, $i_2 = N+M-1$

$$\tilde{A}^H = \begin{bmatrix} u(1) & u(2) & \cdots & u(M) & u(M+1) & \cdots & u(N) \\ 0 & u(1) & \cdots & u(M-1) & u(M) & \cdots & u(N-1) \\ \vdots & 0 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots 0 & u(1) & u(2) & \cdots & u(N-M+1) \end{bmatrix}$$

4. Postwindowing method, $i_1 = M$, $i_2 = N+M-1$

$$\tilde{A}^H = \begin{bmatrix} u(M) & u(M+1) & \cdots & u(N) & 0 & \cdots & 0 \\ u(M-1) & u(M) & \cdots & u(N-1) & u(N) & 0 \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ & & & & & & 0 \\ u(1) & u(2) & \cdots & u(N-M+1) & u(N-M) & \cdots & u(N) \end{bmatrix}$$

Different data windowing methods will lead to different correlation matrix for the input. The only one yielding a Toeplitz correlation matrix is the autocorrelation method.

8.3 Principle of Orthogonality

Take the covariance method for example. The cost function for a least-square problem is stated as

$$\mathcal{E}(\omega_0, \omega_1, \ldots, \omega_{M-1}) = \sum_{i=M}^{N} |e(i)|^2$$

$$= \sum_{i=M}^{N} e(i)\, e^*(i),$$

where $e(i)$ is the estimation error.

Let the k^{th} tap-weight be expressed as

$$w_k = a_k + j b_k, \quad k = 0, 1, \dots, M-1.$$

Then
$$e(i) = d(i) - \sum_{k=0}^{M-1} (a_k - j b_k) u(i-k).$$

The gradient vector

$$\nabla_k \mathcal{E} = -\sum_{i=M}^{N} \left[e(i) \frac{\partial e^*(i)}{\partial a_k} + e^*(i) \frac{\partial e(i)}{\partial a_k} + j e(i) \frac{\partial e^*(i)}{\partial b_k} + j e^*(i) \frac{\partial e(i)}{\partial b_k} \right]$$

$$= -2 \sum_{i=M}^{N} u(i-k) e^*(i) \Big|_{e(i) = e_{min}(i)} = 0$$

$$k = 0, 1, \dots, M-1$$

Hence, the principle of the orthogonality is equivalent to

$$\sum_{i=M}^{N} u(i-k) e^*_{min}(i) = 0,$$

$$k = 0, 1, \dots M-1$$

Or it is stated as follows:

The minimum-error series $e_{min}(i)$ is orthogonal to the input series $u(i-k)$ for $k = 0, 1, \cdots M-$ when the filter is operating in its least square condition.

※ Collary:

Let $\hat{w}_0, \hat{w}_1, \cdots, \hat{w}_{M-1}$ denote the special values of the tap weights $w_0, w_1, w_2, \cdots, w_{M-1}$ that result from the optimization of the transversal filter in its least-squares condition.

The output is

$$y_{min}(i) = \sum_{k=0}^{M-1} \hat{w}_k^* u(i-k).$$

This output provides a least-squares estimation of the desired response and it's a linear estimation such that

$$\hat{d}(i \mid \vec{u}_i) \equiv y_{min}(i)$$

or $\hat{d}(i|\vec{u}_i) = \sum_{k=0}^{M-1} \hat{\omega}_k^* u(i-k)$.

$$\Rightarrow \sum_{i=M}^{N} \left[\sum_{k=0}^{M-1} \hat{\omega}_k^* u(i-k)\right] e_{min}^*(i) = 0$$

$$\Rightarrow \sum_{i=M}^{N} \hat{d}(i|\vec{u}_i)\, e_{min}^*(i) = 0.$$

It states that the least-squares estimate of the desired response $\hat{d}(i|\vec{u}_i)$ is orthogonal to the minimum error series $e_{min}(i)$.

8.4 Minimum Sum of Error Squares

The desired response can be decomposed as:

$$d(i) = \hat{d}(i|\vec{u}_i) + e_{min}(i)$$

If we define

$$\mathcal{E}_d \equiv \sum_{i=M}^{N} |d(i)|^2$$

$$\mathcal{E}_{est} \equiv \sum_{i=M}^{N} |\hat{d}(i|\vec{u}_i)|^2$$

$$\mathcal{E}_{min} \equiv \sum_{i=M}^{N} |e_{min}(i)|^2$$

then $\mathcal{E}_d = \mathcal{E}_{est} + \mathcal{E}_{min}$.

8.5 Normal Equations and Linear Least-Squares Filters

According to Eq. (8.4), the minimum estimation error $e_{min}(i)$ can be achieved as

$$e_{min}(i) = d(i) - \sum_{t=0}^{M-1} \hat{w}_i^* u(i-t).$$

Substituting Eq. (8.27) into Eq. (8.15), we obtain M simultaneous equations:

$$\sum_{t=0}^{M-1} \hat{w}_t \sum_{i=M}^{N} u(i-k)\, u^*(i-t)$$

$$= \sum_{i=M}^{N} u(i-k)\, d^*(i), \quad k = 0, \ldots, M-1$$

It can be shown that those equations are equivalat to $\sum_{t=0}^{M-1} \hat{w}_t\, \phi(t,k) = z(-k)$,

$k = 0, 1, \ldots, M-1,$

(Nomal Equations)

where
$$\phi(t,k) = \sum_{i=M}^{N} u(i-k)\, u^*(i-t), \quad 0 \le (t,k) \le M-1$$

and
$$z(-k) = \sum_{i=M}^{N} u(i-k)\, d^*(i), \quad 0 \le k \le M-1.$$

※ Matrix Formulation of the Normal Equations

Eq. (8.31) can be written as

$$\tilde{\Phi}\, \hat{\vec{w}} = \vec{z},$$

where
$$\tilde{\Phi} = \begin{bmatrix} \phi(0,0) & \phi(1,0) & \cdots & \phi(M-1,0) \\ \phi(0,1) & \phi(1,1) & \cdots & \phi(M-1,1) \\ \vdots & & & \vdots \\ \phi(0,M-1) & \phi(1,M-1) & \cdots & \phi(M-1,M-1) \end{bmatrix}$$

and $\hat{\vec{w}} = [\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_{M-1}]^T$, $\vec{z} = [z(0), z(-1), \ldots z(-M+1)]^T$.

Therefore the tap-weight vector of the linear least squares filter is

$$\hat{\vec{w}} = \tilde{\Phi}^{-1} \vec{z},$$

where $\tilde{\Phi}$ is assumed to be nonsingular.

※ Minimum Sum of Error Squares

$$\mathcal{E}_{est} = \sum_{i=M}^{N} |\hat{d}(i|\vec{u}_i)|^2$$

$$= \sum_{i=M}^{N} \sum_{t=0}^{M-1} \sum_{k=0}^{M-1} \hat{w}_t \hat{w}_k^* u(i-k) u^*(i-t)$$

$$= \sum_{t=0}^{M-1} \sum_{k=0}^{M-1} \hat{w}_t \hat{w}_k^* \sum_{i=M}^{N} u(i-k)\, u^*(i-t),$$

$$= \sum_{t=0}^{M-1} \sum_{k=0}^{M-1} \hat{w}_k^* \phi(t,k) \hat{w}_t$$

$$= \hat{\vec{w}}^H \underbrace{\tilde{\Phi} \hat{\vec{w}}}_{= \vec{z}} = \vec{z}^H \hat{\vec{w}}$$

Therefore $\mathcal{E}_{min} = \mathcal{E}_d - \mathcal{E}_{est}$

$$= \mathcal{E}_d - \vec{z}^H \hat{\vec{w}}$$

$$= \mathcal{E}_d - \vec{z}^H \tilde{\Phi}^{-1} \vec{z}.$$

8.6 Time-average correlation matrix $\tilde{\Phi}$

The time-average correlation matrix $\tilde{\Phi}$ can be written as

$$\tilde{\Phi} = \sum_{i=M}^{N} \vec{u}(i)\,\vec{u}^H(i),$$

where $\vec{u}(i) = [u(i), u(i-1), \cdots u(i-M+1)]^T$.

Property 1. The correlation matrix $\tilde{\Phi}$ is Hermitian.

Proof: It can easily be proved through Eq. (8.42).

Property 2. The correlation matrix $\tilde{\Phi}$ is nonnegative definite, that is $\vec{x}^H \tilde{\Phi} \vec{x} \geq 0$ for any nontrivial M-by-1 $\vec{x}$.

Proof: $\vec{x}^H \tilde{\Phi} \vec{x} = \sum_{i=M}^{N} \vec{x}^H \vec{u}(i)\, \vec{u}^H(i) \vec{x}$

$$= \sum_{i=M}^{N} [\vec{x}^H \vec{u}(i)] [\vec{x}^H \vec{u}(i)]^H$$

$$= \sum_{i=M}^{N} |\vec{x}^H \vec{u}(i)|^2 \geq 0.$$

Property 3. The correlation matrix $\tilde{\Phi}$ is nonsingular iff its determinant is not zero

Property 4. The eigenvalues of the correlation matrix $\tilde{\Phi}$ are all real and nonnegative.

Proof: It can follow Property 1 and 2 to verify this.

Property 5: The correlation matrix is the product of two rectangular matrices that are Hermitian adjoint of each other.

Proof: According to Eq. (8.42),

$$\tilde{\Phi} = [\vec{u}(M), \vec{u}(M+1), \ldots, \vec{u}(N)] \begin{bmatrix} \vec{u}^H(M) \\ \vec{u}^H(M-1) \\ \vdots \\ \vec{u}^H(N) \end{bmatrix}$$

For convenience, we define the data matrix as

$$\tilde{A}^H = [\vec{u}(M), \vec{u}(M+1), \ldots, \vec{u}(N)]$$

$$= \begin{bmatrix} u(M) & u(M+1) & \cdots & u(N) \\ u(M-1) & u(M) & \cdots & u(N-1) \\ \vdots & \vdots & & \vdots \\ u(1) & u(2) & \cdots & u(N-M+1) \end{bmatrix}$$

Then $\tilde{\Phi} = \tilde{A}^H \tilde{A}$, where $\tilde{A}^H$ is an M-by-$(N-M+1)$ rectangular Toeplitz matrix and $\tilde{A}$ itself is likewise an $(N-M+1)$-by-M rectangular Toeplitz matrix. $\tilde{A}$ can have different forms as described in Section 8.2. Here we only show the covariance method.

8.7 Reformulation of the Normal Equations in Terms of Data Matrices

First, we can define a desired data vector $\vec{d}$ such that

$$\vec{d}^H \equiv [d(M), d(M+1), \ldots, d(N)]$$

From Eq. (8.30), (8.33), (8.44), (8.46),

We have the cross-correlation vector as

$$\vec{z} = \tilde{A}^H \vec{d}.$$

From Eq. (8.35), (8.45), (8.47), we have

$$\tilde{A}^H \tilde{A} \vec{w} = \tilde{A}^H \vec{d},$$

which are the normal equations in matrix form.

When the cost function $\mathcal{E}$ is minimized,

the optimal tap-weight vector is

$$\hat{\vec{w}} = (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \vec{d}.$$

The minimum squared error can be obtained as

$$\mathcal{E}_{min} = \vec{d}^H \vec{d} - \vec{d}^H \tilde{A} (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \vec{d}.$$

X. Projection Operator

The least-squares estimate of $\vec{d}$ is given by

$$\hat{\vec{d}} = \tilde{A}\,\hat{\vec{w}}$$

$$= \tilde{A}(\tilde{A}^H\tilde{A})^{-1}\tilde{A}^H\vec{d}$$

We can denote $\tilde{P} \equiv \tilde{A}(\tilde{A}^H\tilde{A})^{-1}\tilde{A}^H$ as the projection operator onto the linear space spanned by the columns of $\tilde{A}$.

In addition, the matrix difference

$$\tilde{I} - \tilde{A}(\tilde{A}^H\tilde{A})^{-1}\tilde{A}^H = \tilde{I} - \tilde{P}$$ is

the orthogonal complement operator, such that

$$\vec{e}_{min} = \vec{d} - \hat{\vec{d}} = (\tilde{I} - \tilde{P})\vec{d}$$

Example: A linear least-squares filter with two taps ($M=2$) and a real-valued input series consisting of four samples ($N=4$). Hence $N-M+1=3$. The input data matrix $\tilde{A}$ (covariance method) can be written as

$$\tilde{A} = \begin{bmatrix} u(2) & u(1) \\ u(3) & u(2) \\ u(4) & u(3) \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 1 & 2 \\ -1 & 1 \end{bmatrix}$$

The desired data vector is

$$\vec{d} = \begin{bmatrix} d(2) \\ d(3) \\ d(4) \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ \frac{1}{34} \end{bmatrix}$$

Hence $\tilde{P} = \tilde{A}\,(\tilde{A}^T\tilde{A})^{-1}\,\tilde{A}^T$

$$= \frac{1}{35}\begin{bmatrix} 26 & 15 & -3 \\ 15 & 10 & 5 \\ -3 & 5 & 34 \end{bmatrix}$$

$$\tilde{I} - \tilde{P} = \frac{1}{35}\begin{bmatrix} 9 & -15 & 3 \\ -15 & 25 & -5 \\ -3 & -5 & 1 \end{bmatrix}$$

$$\hat{\vec{d}} = \tilde{P}\vec{d}$$

$$= \begin{bmatrix} 1.91 \\ 1.15 \\ 0 \end{bmatrix}$$

and $$\vec{e}_{min} = (\tilde{I} - \tilde{P})\vec{d}$$

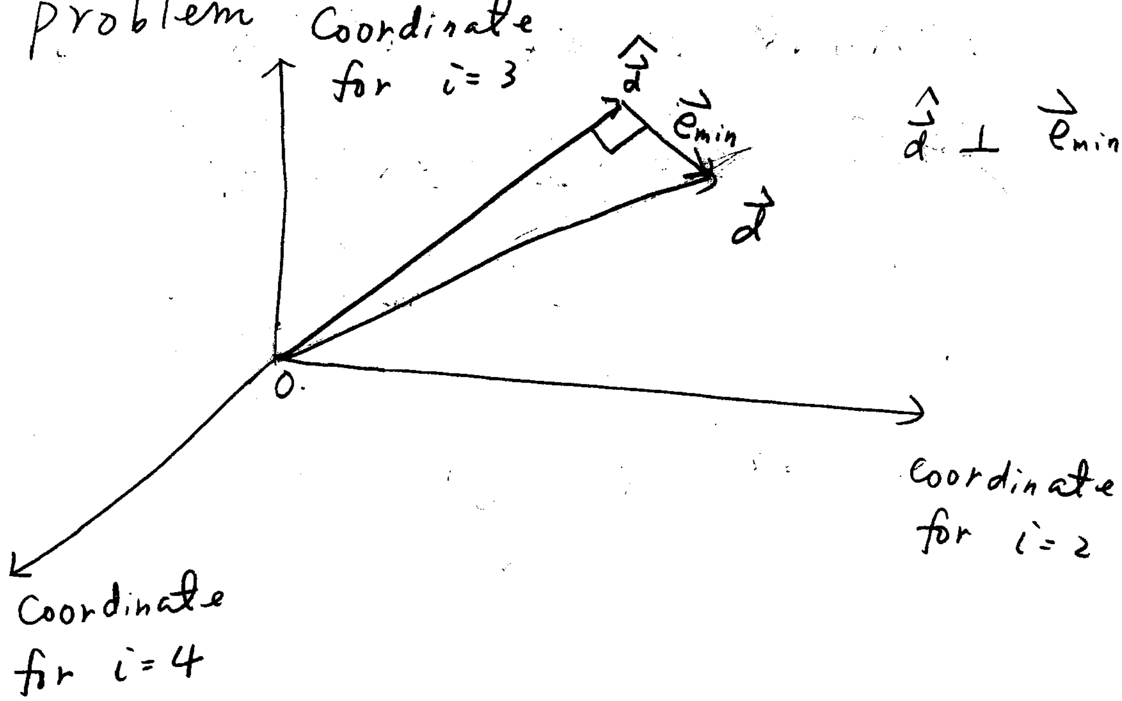
$$= \tilde{d} - \hat{\vec{d}}$$

$$= \begin{bmatrix} 0.09 \\ -0.15 \\ 0.03 \end{bmatrix}$$

$$\|\vec{e}_{min}\|^2 = \vec{e}_{min}^{\,T}\vec{e}_{min} = 0.0315 \cdots\cdots>$$

total squared error

※ Geometric representation of Least-squares problem



$\hat{\vec{d}} \perp \vec{e}_{min}$

X. Uniqueness Theorem

The least-squares estimate $\hat{\vec{w}}$ is unique if and only if the nullity of the data matrix $\widetilde{A}$ equals zero

Proof: Let $\widetilde{A}$ be a K-by-M matrix as defined in Eq. (8.44) such that $K = N-M+1$. The null space of matrix $\widetilde{A}$, written as $\mathcal{N}(\widetilde{A})$, as the space of all vectors $\vec{x}$ such that $\widetilde{A}\vec{x} = \vec{0}$. We define the nullity of $\widetilde{A}$ or $\text{null}(\widetilde{A})$ as the dimension of the null space $\mathcal{N}(\widetilde{A})$.

In general, $\text{null}(\widetilde{A}) \neq \text{null}(\widetilde{A}^H)$.

We may expect a unique solution to the linear least-squares problem only when the data matrix $\widetilde{A}$ has M linearly independent columns, that is, the data matrix $\tilde{A}$ is of full column rank. This implies that the matrix $\tilde{A}$ has at least as many rows as columns, $(N-M+1) \geq M$.

Or in other words, $\tilde{A}\hat{\vec{w}} = \vec{d}$ is an over determined minimization problem. Thus, $\tilde{A}^H\tilde{A}$ is nonsingular and the least-squares estimate has the unique value given in Eq. (8.48)

8.8 Properties of Least-squares Estimator

Property 1. The least-squares estimate $\hat{\vec{w}}$ is unbiased, provided that the measurement error process $e_o(i)$ has zero mean.

Proof: According to Eq. (8.44) & (8.46),

$$\vec{d} = \tilde{A}\vec{w_0} + \vec{\varepsilon}_0,$$

where $\vec{\varepsilon}_0^{\,H} = [\, e_0(M),\, e_0(M+1) \,..\, e_0(N) \,]$

Substituting Eq. (8.52) into Eq. (8.48), the least-squares estimate can be expressed as

$$\hat{\vec{w}} = (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \tilde{A}\, \vec{w}_0 + (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \vec{\varepsilon}_0$$

$$= \vec{w}_0 + (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \vec{\varepsilon}_0 .$$

Hence $E[\hat{\vec{w}}] = \vec{w}_0 + (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \underbrace{E[\vec{\varepsilon}_0]}_{= \vec{0}}$

$$= \vec{w}_0$$

Property 2. If the measurement error process $e_0(i)$ is white process with zero mean and variance σ^2, the covariance matrix of the least-squares estimate $\hat{\vec{w}}$ equals $\sigma^2 \tilde{\Phi}^{-1}$

Proof:

$$\operatorname{cov}[\hat{\vec{w}}] = E[(\hat{\vec{w}} - \vec{w}_0)(\hat{\vec{w}} - \vec{w}_0)^H]$$

$$= E[(\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \vec{\varepsilon}_0 \vec{\varepsilon}_0^H \tilde{A} (\tilde{A}^H \tilde{A})^{-1}]$$

$$= (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \underbrace{E[\vec{\varepsilon}_0 \vec{\varepsilon}_0^H]}_{\sigma^2 \tilde{I}} \tilde{A} (\tilde{A}^H \tilde{A})^{-1}$$

$$= \sigma^2 \cancel{(\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H \tilde{A}} (\tilde{A}^H \tilde{A})^{-1}$$

$$= \sigma^2 (\tilde{A}^H \tilde{A})^{-1} = \sigma^2 \tilde{\Phi}^{-1}$$

Property 3. If the measurement error series $e_0(i)$ is white process with zero mean and variance σ^2, the least squares estimate $\hat{\vec{w}}$ is the minimum-variance linear unbiased estimate.

Proof: Consider any linear unbiased estimator that is defined by

$$\check{\vec{w}} = \tilde{B}\vec{d},$$

where $\tilde{B}$ is an M-by-$(N-M+1)$ matrix. Substituting Eq. (8.52) into Eq. (8.58), we get

$$\check{\vec{w}} = \tilde{B}\tilde{A}\vec{w}_0 + \tilde{B}\vec{\varepsilon}_0$$

$$E[\check{\vec{w}}] = \tilde{B}\tilde{A}\vec{w}_0 + \tilde{B}\underbrace{E[\vec{\varepsilon}_0]}_{=\vec{0}}$$

$$= \tilde{B}\tilde{A}\vec{w}_0$$

Since the linear estimate $\check{\vec{w}}$ is unbiased, then $\tilde{B}$ has to satisfy the condition:

$$\tilde{B}\,\tilde{A} = \tilde{I}.$$

Therefore, Eq. (8.59) can be rewritten as

$$\check{\vec{w}} = \vec{w}_0 + \tilde{B}\,\vec{\varepsilon}_0$$

The covariance matrix of the estimate is

$$\begin{aligned} \mathrm{cov}(\check{\vec{w}}) &= E\left[(\check{\vec{w}} - \vec{w}_0)(\check{\vec{w}} - \vec{w}_0)^H\right] \\ &= E\left[\tilde{B}\,\vec{\varepsilon}_0\,\vec{\varepsilon}_0^{\,H}\,\tilde{B}^H\right] \\ &= \sigma^2\,\tilde{B}\,\tilde{B}^H \end{aligned}$$

Define a new matrix $\tilde{\Psi}$ such that

$$\tilde{\Psi} \equiv \tilde{B} - (\tilde{A}^H\tilde{A})^{-1}\tilde{A}^H$$

$\tilde{\Psi}\tilde{\Psi}^H \geq 0$ ($\tilde{\Psi}\tilde{\Psi}^H$ is always nonnegative definite)

$$\tilde{\Psi}\tilde{\Psi}^H = [\tilde{B} - (\tilde{A}^H\tilde{A})^{-1}\tilde{A}^H][\tilde{B}^H - \tilde{A}(\tilde{A}^H\tilde{A})^{-1}]$$

$$= \tilde{B}\tilde{B}^H - \underbrace{\tilde{B}\tilde{A}}_{\tilde{I}}(\tilde{A}^H\tilde{A})^{-1} - (\tilde{A}^H\tilde{A})^{-1}\underbrace{\tilde{A}^H\tilde{B}^H}_{\tilde{I}} + (\tilde{A}^H\tilde{A})^{-1}$$

$$= \tilde{B}\tilde{B}^H - (\tilde{A}^H\tilde{A})^{-1}$$

Since $\tilde{\Psi}\tilde{\Psi}^H$ is nonnegative definite,

$\vec{x}^H\tilde{\Psi}\tilde{\Psi}^H\vec{x} \geq 0$ for any nontrivial column vector $\vec{x}$.

$$\Rightarrow \quad \vec{x}^H\tilde{B}\tilde{B}^H\vec{x} \geq \vec{x}^H(\tilde{A}^H\tilde{A})^{-1}\vec{x}$$

$$\Rightarrow \quad \sigma^2 \operatorname{diag}[\tilde{B}\tilde{B}^H] \geq \sigma^2 \operatorname{diag}[(\tilde{A}^H\tilde{A})^{-1}]$$

$$\Rightarrow \quad E[\|\check{\vec{w}}\|^2] \geq E[\|\hat{\vec{w}}\|^2]$$

$\therefore$ The variance of $\hat{\vec{w}}$ is always less than that of any other linear unbiased estimate $\check{\vec{w}}$.

Property 4. If the measurement error series $e_o(i)$ is white process with zero mean and variance σ^2, the least-squares estimate $\hat{\vec{w}}$ achieves the Cramer-Rao lower bound (CR-bound) among all unbiased estimates.

Proof: Let $f_E(\vec{\varepsilon}_o)$ denote the joint probability density function of the measurement error vector $\vec{\varepsilon}_o$. Let $\check{\vec{w}}$ denote any unbiased estimate of the unknown parameter vector $\vec{w}_o$. Then from the Cramer-Rao inequality

$$\operatorname{cov}(\check{\vec{w}}) \geq \tilde{J}^{-1},$$

where $\operatorname{cov}(\check{\vec{w}}) = E\left[(\check{\vec{w}} - \vec{w}_o)(\check{\vec{w}} - \vec{w}_o)^H\right]$

and $$\tilde{J} = E\left[\left(\frac{\partial \ln f_E(\vec{\varepsilon}_o)}{\partial \vec{w}_o^{*}}\right)\left(\frac{\partial \ln f_E(\vec{\varepsilon}_o)}{\partial \vec{w}_o^{T}}\right)\right]$$

is called the Fisher's information matrix.

Since $e_o(i)$ is assumed to be complex with zero mean and variance σ^2:

$$f_E(\vec{\mathcal{E}}_o) = \frac{1}{(\pi\sigma^2)^{(N-M+1)}} \exp\left[-\frac{1}{\sigma^2}\sum_{i=M}^{N} |e_o(i)|^2\right]$$

$$\ln\left[f_E(\vec{\mathcal{E}}_o)\right] = -(N-M+1)\ln(\pi\sigma^2) - \frac{1}{\sigma^2}\sum_{i=M}^{N}|e_o(i)|^2$$

$$= -(N-M+1)\ln(\pi\sigma^2) - \frac{1}{\sigma^2}\vec{\mathcal{E}}_o^H\vec{\mathcal{E}}_o$$

According to Eq. (8.52), $\vec{\mathcal{E}}_o = \vec{d} - \tilde{A}\vec{w}_o$

$$\therefore \ln\left[f_E(\vec{\mathcal{E}}_o)\right] = -(N-M+1)\ln(\pi\sigma^2) - \frac{1}{\sigma^2}\vec{d}^H\vec{d} + \frac{1}{\sigma^2}\vec{w}_o^H\tilde{A}^H\vec{d} + \frac{1}{\sigma^2}\vec{d}^H\tilde{A}\vec{w}_o - \frac{1}{\sigma^2}\vec{w}_o^H\tilde{A}^H\tilde{A}\vec{w}_o$$

$$\frac{\partial \ln\left[f_E(\vec{\mathcal{E}}_o)\right]}{\partial \vec{w}_o^*} = \frac{1}{\sigma^2}\tilde{A}^H(\vec{d} - \tilde{A}\vec{w}_o)$$

$$= \frac{1}{\sigma^2}\tilde{A}^H\vec{\mathcal{E}}_o$$

Substituting Eq. (8.70) into Eq. (8.65), we get

$$J = \frac{1}{\sigma^4} E[\tilde{A}^H \vec{\varepsilon}_0 \vec{\varepsilon}_0^H \tilde{A}]$$

$$= \frac{1}{\sigma^4} \tilde{A}^H E[\vec{\varepsilon}_0 \vec{\varepsilon}_0^H] \tilde{A}$$

$$= \frac{1}{\sigma^2} \tilde{A}^H \tilde{A} = \frac{1}{\sigma^2} \tilde{\Phi}$$

Therefore, $\mathrm{cov}(\check{\vec{w}}) \geq J^{-1} = \sigma^2 \tilde{\Phi}^{-1}$

8.11 Singular-Value Decomposition

Given the data matrix $\tilde{A}$, there are two unitary matrices $\tilde{V}$ and $\tilde{U}$, such that

$$\tilde{U}^H \tilde{A} \tilde{V} = \begin{bmatrix} \tilde{\Sigma} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix},$$

where $\tilde{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_W)$

and W is the rank of $\tilde{A}$

Proof:

Case 1: Overdetermined System

When $k > M$, the eigen values of $\tilde{A}^H \tilde{A}$ are $\sigma_1^2, \sigma_2^2, \cdots, \sigma_M^2$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_W > 0$ and $\sigma_{W+1}, \sigma_{W+2} \cdots \sigma_M$ are all zero with $1 \leq W \leq M$.

$$\tilde{V}^H \tilde{A}^H \tilde{A} \tilde{V} = \begin{bmatrix} \tilde{\Sigma}^2 & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix}$$

Let the eigen vector matrix $\tilde{V}$ be partitioned as

$$\tilde{V} = [\tilde{V}_1, \tilde{V}_2],$$

where $\tilde{V}_1 = [\vec{V}_1, \vec{V}_2, \cdots \vec{V}_W]$ is an M-by-W matrix and

$$\tilde{V}_2 = [\vec{V}_{W+1}, \vec{V}_{W+2}, \cdots \vec{V}_M]$$

is an M-by-$(M-W)$ matrix, with

$$\tilde{V}_1^H \tilde{V}_2 = \tilde{0}$$

$$\tilde{V}_1^H \tilde{A}^H \tilde{A} \tilde{V}_1 = \tilde{\Sigma}^2$$

$$\tilde{\Sigma}^{-1} \tilde{V}_1^H \tilde{A}^H \tilde{A} \tilde{V}_1 \tilde{\Sigma}^{-1} = \tilde{I}$$

$$\tilde{V}_2^H \tilde{A}^H \tilde{A} \tilde{V}_2 = \tilde{0}$$

$$\Rightarrow \tilde{A} \tilde{V}_2 = \tilde{0}$$

We define a new K-by-W matrix

$$\tilde{U}_1 = \tilde{A} \tilde{V}_1 \tilde{\Sigma}^{-1}$$

Then $\tilde{U}_1^H \tilde{U}_1 = \tilde{I}$,

which means that the columns of the matrix $\tilde{U}_1$ are orthonormal with respect to each other.

Choose $\tilde{U}_2$ such that $\tilde{U}_1^H \tilde{U}_2 = \tilde{0}$

Then
$$\tilde{U}^H \tilde{A} \tilde{V} = \begin{bmatrix} \tilde{U}_1^H \\ \tilde{U}_2^H \end{bmatrix} \tilde{A} [\tilde{V}_1, \tilde{V}_2]$$

$$= \begin{bmatrix} \tilde{U}_1^H \tilde{A} \tilde{V}_1 & \tilde{U}_1^H \tilde{A} \tilde{V}_2 \\ \tilde{U}_2^H \tilde{A} \tilde{V}_1 & \tilde{U}_2^H \tilde{A} \tilde{V}_2 \end{bmatrix}$$

$$= \begin{bmatrix} (\tilde{\Sigma}^{-1} \tilde{V}_1^H \tilde{A}^H) \tilde{A} \tilde{V}_1 & \tilde{U}_1^H (\tilde{0}) \\ \tilde{U}_2^H (\tilde{U}_1 \tilde{\Sigma}) & \tilde{U}_2^H (\tilde{0}) \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\Sigma} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix}$$

Case 2. Underdetermined System.

$K < M$, the eigenvalues of $\tilde{A}\tilde{A}^H$ can be denoted as $\sigma_1^2, \sigma_2^2, \cdots \sigma_K^2$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_W > 0$ and $\sigma_{W+1} = \sigma_{W+2} = \cdots = 0$ with $1 \leq W \leq K$. The eigenvectors of $\tilde{A}\tilde{A}^H$ are $\vec{u}_1, \vec{u}_2, \cdots, \vec{u}_K$ such that

$$\tilde{U}^H \tilde{A}\tilde{A}^H \tilde{U} = \begin{bmatrix} \tilde{\Sigma}^2 & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix}$$

Let the unitary matrix $\tilde{U}$ be partitioned as

$$\tilde{U} = [\tilde{U}_1, \tilde{U}_2],$$

where $\tilde{U}_1 = [\vec{u}_1, \vec{u}_2, \cdots \vec{u}_W]$ and

$$\tilde{U}_2 = [\vec{u}_{W+1}, \vec{u}_{W+2}, \cdots \vec{u}_K]$$

and $\tilde{U}_1^H \tilde{U}_2 = \tilde{0}$

Therefore,

$$\tilde{U}_1^H \tilde{A} \tilde{A}^H \tilde{U}_1 = \tilde{\Sigma}^2$$

$$\tilde{\Sigma}^{-1} \tilde{U}_1^H \tilde{A} \tilde{A}^H \tilde{U}_1 \tilde{\Sigma}^{-1} = \tilde{I}$$

$$\tilde{U}_2^H \tilde{A} \tilde{A}^H \tilde{U}_2 = \tilde{0}$$

$$\tilde{A}^H \tilde{U}_2 = \tilde{0}$$

$$\tilde{V}_1 = \tilde{A}^H \tilde{U}_1 \tilde{\Sigma}^{-1}$$

$$\tilde{V}_1^H \tilde{V}_1 = \tilde{I}$$

We can construct $\tilde{V}_2$ such that $\tilde{V}_2^H \tilde{V}_1 = \tilde{0}$ and a matrix $\tilde{V} = [\tilde{V}_1, \tilde{V}_2]$.

Then

$$\tilde{U}^H \tilde{A} \tilde{V} = \begin{bmatrix} \tilde{U}_1^H \\ \tilde{U}_2^H \end{bmatrix} \tilde{A} \begin{bmatrix} \tilde{V}_1 & \tilde{V}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{U}_1^H \tilde{A} \tilde{V}_1 & \tilde{U}_1^H \tilde{A} \tilde{V}_2 \\ \tilde{U}_2^H \tilde{A} \tilde{V}_1 & \tilde{U}_2^H \tilde{A} \tilde{V}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{U}_1^H \tilde{A} (\tilde{A}^H \tilde{U}_1 \tilde{\Sigma}^{-1}) & (\tilde{\Sigma} \tilde{V}_1^H) \tilde{V}_2 \\ (\tilde{0}) \tilde{V}_1 & (\tilde{0}) \tilde{V}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\Sigma} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix}$$

Terminology in SVD

The numbers $\sigma_1, \sigma_2, \cdots, \sigma_W$ constituting the diagonal matrix $\tilde{\Sigma}$, are called the singular values of the matrix $\tilde{A}$. The columns of the unitary matrix $\tilde{V}$, i.e., $\vec{v}_1, \vec{v}_2, \cdots \vec{v}_M$ are the right singular vectors of $\tilde{A}$ and the columns of the second unitary matrix $\tilde{u}$, i.e., $\vec{u}_1, \vec{u}_2, \cdots \vec{u}_K$, are the left singular vectors of $\tilde{A}$.

From Eq. (8.103),

$$\tilde{A}\tilde{V} = \tilde{u} \begin{bmatrix} \tilde{\Sigma} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix}.$$

Therefore $\tilde{A}\vec{v}_i = \sigma_i \vec{u}_i, \quad i = 1, 2, \cdots W$

and $\tilde{A}\vec{v}_i = \vec{0}, \quad i = W+1, \cdots K$

Hence $\tilde{A} = \sum_{i=1}^{W} \sigma_i \vec{u}_i \vec{v}_i^H$.

Since $\tilde{V}\tilde{V}^H = \tilde{I}$,

$$\tilde{U}^H \tilde{A} = \begin{bmatrix} \tilde{\Sigma} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix} \tilde{V}^H$$

Therefore $\tilde{A}^H \vec{u}_i = \sigma_i \vec{v}_i$, $i = 1, 2, \ldots, W$

and $\tilde{A}^H \vec{u}_i = \vec{0}$, $i = W+1, \ldots M$.

$$\tilde{A}^H = \sum_{i=1}^{W} \sigma_i \vec{v}_i \vec{u}_i^H$$

Example: Let $\tilde{A}$ be a k-by-M data matrix with rank W. Then $\tilde{A}$ is said to be full rank if $W = \min(k, M)$. Otherwise, the matrix $\tilde{A}$ is rank deficient. The rank W, is simply the number of nonzero Singular values of matrix $\tilde{A}$.

Consider a numerical value ε, such that each element of $\widehat{A}$ that is accurate of within $\pm\varepsilon$. Let $\widetilde{B}$ denote the approximate value of $\widehat{A}$ so obtained. We define ε - rank of matrix $\widehat{A}$ as

$$\operatorname{rank}(\widetilde{A}, \varepsilon) = \min_{\|\widetilde{A} - \widetilde{B}\| < \varepsilon} \operatorname{rank}(\widetilde{B}),$$

where $\|\widetilde{A} - \widetilde{B}\|$ is the spectral norm of the largest singular value of $(\widetilde{A} - \widetilde{B})$. The K-by-M matrix $\widetilde{A}$ is said to be numerically rank deficient if $\operatorname{rank}(\widetilde{A}, \varepsilon) < \min(K, M)$.

8.12 Pseudoinverse

The pseudo inverse matrix of $\widetilde{A}$ is defined as $\widetilde{A}^{+} = \widehat{V} \begin{bmatrix} \widetilde{\Sigma}^{-1} & \widetilde{0} \\ \widetilde{0} & \widetilde{0} \end{bmatrix} \widehat{U}^{H}$

where $\widetilde{\Sigma}^{-1} = \operatorname{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \cdots \sigma_W^{-1})$

and $\widetilde{A} = \widehat{U} \begin{bmatrix} \widetilde{\Sigma} & \widetilde{0} \\ \widetilde{0} & \widetilde{0} \end{bmatrix} \widehat{V}^{H}$

$$\tilde{A}^{+} = \sum_{i=1}^{W} \frac{1}{\sigma_i} \vec{v}_i \vec{u}_i^{H}$$

Case 1: Overdetermined System $(K > M)$

Assume $W = M$. Hence $(\tilde{A}^H \tilde{A})^{-1}$ exists.

The pseudoinverse of the data matrix $\tilde{A}$ is then defined by $\tilde{A}^{+} = (\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H$.

According to Eq. (8.110) and (8.112),

$$(\tilde{A}^H \tilde{A})^{-1} = \tilde{V}_1 \tilde{\Sigma}^{-2} \tilde{V}_1^H$$

and $\tilde{A}^H = \tilde{V}_1 \tilde{\Sigma} \tilde{U}_1^H$

Therefore,

$$\begin{aligned}
(\tilde{A}^H \tilde{A})^{-1} \tilde{A}^H &= (\tilde{V}_1 \tilde{\Sigma}^{-2} \tilde{V}_1^H)(\tilde{V}_1 \tilde{\Sigma} \tilde{U}_1^H) \\
&= \tilde{V}_1 \tilde{\Sigma}^{-1} \tilde{U}_1^H \\
&= \tilde{V} \begin{bmatrix} \tilde{\Sigma}^{-1} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix} \tilde{U}^H \\
&= \tilde{A}^{+}
\end{aligned}$$

Case 2. Underdetermined System $(M > k)$

Assume $W = k$. Therefore $(\tilde{A}\tilde{A}^H)^{-1}$ exists.

$$\tilde{A}^+ = \tilde{A}^H(\tilde{A}\tilde{A}^H)^{-1}$$

According to Eq. (8.121) and (8.123),

$$(\tilde{A}\tilde{A}^H)^{-1} = \tilde{U}_1 \tilde{\Sigma}^{-2} \tilde{U}_1^H$$

$$\tilde{A}^H = \tilde{V}_1 \tilde{\Sigma} \tilde{U}_1^H$$

$$\tilde{A}^H(\tilde{A}\tilde{A}^H)^{-1} = (\tilde{V}_1 \tilde{\Sigma} \tilde{U}_1^H)(\tilde{U}_1 \tilde{\Sigma}^{-2} \tilde{U}_1^H)$$

$$= \tilde{V}_1 \tilde{\Sigma}^{-1} \tilde{U}_1^H$$

$$= \tilde{V} \begin{bmatrix} \tilde{\Sigma}^{-1} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix} \tilde{U}^H = \tilde{A}^+$$

8.13 Interpretation of SVD

Consider a k-by-M data matrix $\tilde{A}$. Assume that the system is overdetermined, and the simultaneous linear equations can be stated as:

$$\vec{y} = \tilde{A}\vec{x}, \text{ where } \|\vec{x}\| = 1$$

It reads as a linear transformation of $\hat{A}$ from $\vec{x}$ to $\vec{y}$. Solving that, we get $\vec{x} = \hat{A}^{+}\vec{y}$

$$= \tilde{V} \begin{bmatrix} \tilde{\Sigma}^{-1} & \tilde{0} \\ \tilde{0} & \tilde{0} \end{bmatrix} \hat{u}^{H} \vec{y}$$

$$= \sum_{i=1}^{W} \frac{1}{\sigma_i} \vec{v_i}\, \vec{u_i}^{H} \vec{y} = \sum_{i=1}^{W} \frac{(\vec{u_i}^{H} \vec{y})}{\sigma_i} \vec{v_i},$$

where $\text{rank}(\hat{A}) = W$.

$$\|\vec{x}\|^2 = \vec{x}^{H}\vec{x}$$

$$= \sum_{i=1}^{W} \frac{\|\vec{y}^{H}\vec{u_i}\|^2}{\sigma_i^2} = 1$$

Eq. (8.140) defines the locus traced out by the tip of vector $\vec{y}$ in a K-dimensional space. This equation describes a hyperellipsoid.

A complex scalar is defined as

$$\zeta_i = \vec{y}^H \vec{u}_i = \sum_{k=1}^{K} y_k^* u_{ik}, \quad i = 1, 2, \ldots, W$$

where ζ_i is referred to as the span of $\vec{u}_i$.

We may write $$\sum_{i=1}^{W} \frac{|\zeta_i|^2}{\sigma_i^2} = 1.$$

This is the equation of a hyperellipsoid with coordinates $|\zeta_1|, \ldots |\zeta_W|$ and with semiaxes whose lengths are singular values $\sigma_1, \ldots, \sigma_W$, respectively