

Lecture 9: Linear Least Squares Filtering

Overview

- Linear LS estimation problem;
- Normal equations and LS filters;
- Properties of Least-Squares estimates;
- Singular value decomposition; Pseudoinverse

Reference : Chapter 8 from *S. Haykin- Adaptive Filtering Theory - Prentice Hall, 2002*.

Linear LS estimation problem

Problem statement

- Given the set of input samples $\{u(1), u(2), \dots, u(N)\}$ and the set of desired response $\{d(1), d(2), \dots, d(N)\}$
- In the family of linear filters computing their output according to

$$y(n) = \sum_{k=0}^{M-1} w_k u(n-k), \quad n = 0, 1, 2, \dots \quad (1)$$

- Find the parameters $\{w_0, w_1, \dots, w_{M-1}\}$ such as to minimize the sum of error squares

$$\mathcal{E}(w_0, w_1, \dots, w_{M-1}) = \sum_{i=i_1}^{i_2} [e(i)^2] = \sum_{i=i_1}^{i_2} \left[d(i) - \sum_{k=0}^{M-1} w_k u(i-k) \right]^2$$

where the error signal is

$$e(i) = d(i) - y(i) = d(i) - \sum_{k=0}^{M-1} w_k u(i-k)$$

□

Data windows

Using the vector notations:

$$\begin{aligned}\underline{u}(n) &= [u(n) \ u(n-1) \ u(n-2) \ \dots \ u(n-M+1)]^T \\ \underline{w} &= [w_0 \ w_1 \ \dots \ w_{M-1}]^T\end{aligned}\tag{2}$$

we can write the filter output at time instant i

$$y(i) = \sum_{k=0}^{M-1} w_k u(i-k) = [u(i) \ u(i-1) \ u(i-2) \ \dots \ u(i-M+1)] \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_{M-1} \end{bmatrix} = \underline{u}(i)^T \underline{w}$$

The criterion $\mathcal{E}(w_0, w_1, \dots, w_{M-1})$ will make use of the following errors:

$$\begin{bmatrix} e(i_1) \\ e(i_1+1) \\ \dots \\ e(i_2) \end{bmatrix} = \begin{bmatrix} d(i_1) \\ d(i_1+1) \\ \dots \\ d(i_2) \end{bmatrix} - \begin{bmatrix} y(i_1) \\ y(i_1+1) \\ \dots \\ y(i_2) \end{bmatrix} = \begin{bmatrix} d(i_1) \\ d(i_1+1) \\ \dots \\ d(i_2) \end{bmatrix} - \begin{bmatrix} u(i_1) & u(i_1-1) & u(i_1-2) & \dots & u(i_1-M+1) \\ u(i_1+1) & u(i_1) & u(i_1-1) & \dots & u(i_1-M+2) \\ \dots & \dots & \dots & \dots & \dots \\ u(i_2) & u(i_2-1) & u(i_2-2) & \dots & u(i_2-M+1) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_{M-1} \end{bmatrix}$$

Making use of available data in LS criterion: Selecting the limits i_1 and i_2

There are four ways of selecting the limits i_1 and i_2 and making use of simplifying assumptions:

- **Covariance method:** Uses *only* available data: $i_1 = M$ and $i_2 = N$

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

- **Autocorrelation (Pre- and Post-windowing) method:** Uses unavailable data: $i_1 = 1$ and $i_2 = N + M - 1$. Assumes input data prior to $u(1)$ and after $u(N)$ are zero

$$A = \begin{bmatrix} u(1) & 0 & 0 & \dots & 0 \\ u(2) & u(1) & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot \\ u(M) & u(M-1) & u(M-2) & \dots & u(1) \\ u(M+1) & u(M) & u(M-1) & \dots & u(2) \\ \cdot & \cdot & \cdot & \dots & \cdot \\ u(N) & u(N-1) & u(N-2) & \dots & u(N-M+1) \\ 0 & u(N) & u(N-1) & \dots & u(N-M+2) \\ \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & 0 & 0 & \dots & u(N) \end{bmatrix}$$

- **Prewindowing method:** Uses unavailable data: $i_1 = 1$ and $i_2 = N$. Assumes input data prior to $u(1)$ are zero

$$A = \begin{bmatrix} u(1) & 0 & 0 & \dots & 0 \\ u(2) & u(1) & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot \\ u(M) & u(M-1) & u(M-2) & \dots & u(1) \\ u(M+1) & u(M) & u(M-1) & \dots & u(2) \\ \cdot & \cdot & \cdot & \dots & \cdot \\ u(N) & u(N-1) & u(N-2) & \dots & u(N-M+1) \end{bmatrix}$$

- **Post-windowing method:** Uses unavailable data: $i_1 = M$ and $i_2 = N + M - 1$. Assumes input data after $u(N)$ are zero

$$A = \begin{bmatrix} u(M) & u(M-1) & u(M-2) & \dots & u(1) \\ u(M+1) & u(M) & u(M-1) & \dots & u(2) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ u(N) & u(N-1) & u(N-2) & \dots & u(N-M+1) \\ 0 & u(N) & u(N-1) & \dots & u(N-M+2) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & u(N) \end{bmatrix}$$

Principle of orthogonality for LS filters When the minimum value of the criterion will be attained, the gradient of criterion with respect to parameter vector will be zero:

$$\nabla_{\underline{w}} \mathcal{E}(\underline{w}) = \nabla_{\underline{w}} \sum_{i=i_1}^{i_2} [e(i)^2] = 2 \sum_{i=i_1}^{i_2} e(i) \nabla_{\underline{w}} e(i) = 0$$

which can be written for each component of the gradient vector

$$\nabla_k \mathcal{E}(\underline{w}) = 2 \sum_{i=i_1}^{i_2} e(i) \nabla_k e(i) = 2 \sum_{i=i_1}^{i_2} e(i) \frac{\partial}{\partial w_k} [d(i) - \sum_{l=0}^{M-1} w_l u(i-l)] = 2 \sum_{i=i_1}^{i_2} e(i) u(i-k) = 0$$

$$\begin{aligned} & \sum_{i=i_1}^{i_2} e(i) u(i-k) = \\ & = \begin{bmatrix} e(i_1) & e(i_1+1) & e(i_1+2) & \dots & e(i_2) \end{bmatrix} \begin{bmatrix} u(i_1-k) & u(i_1-k+1) & u(i_1-k+2) & \dots & u(i_2-k) \end{bmatrix}^T = 0 \end{aligned}$$

Principle of orthogonality for LS filters

$$\sum_{i=i_1}^{i_2} e_o(i)u(i-k) = 0 \quad k = 0, 1, \dots, M-1$$

The minimum error time series is orthogonal to the input time series shifted backward with k units, for $k = 0, 1, 2, \dots, M-1$

$$\sum_{i=i_1}^{i_2} e_o(i)y_o(i) = \sum_{i=i_1}^{i_2} e_o(i) \sum_{l=0}^{M-1} \hat{w}_l u(i-l) = \sum_{l=0}^{M-1} \hat{w}_l \sum_{i=i_1}^{i_2} e_o(i)u(i-l) = 0$$

Corollary of principle of orthogonality

$$\sum_{i=i_1}^{i_2} e_o(i)y_o(i) = 0$$

The minimum error time series is orthogonal to the optimal LS filter output time series

Normal equations and Linear Least Squares filters

Rearranging the orthogonality equations we have for all $k = 0, 1, \dots, M - 1$

$$\begin{aligned} \sum_{i=i_1}^{i_2} e_o(i)u(i-k) &= 0 \\ \sum_{i=i_1}^{i_2} [d(i) - \sum_{l=0}^{M-1} \hat{w}_l u(i-l)]u(i-k) &= 0 \\ \sum_{i=i_1}^{i_2} d(i)u(i-k) &= \sum_{i=i_1}^{i_2} \sum_{l=0}^{M-1} \hat{w}_l u(i-l)u(i-k) \\ \sum_{i=i_1}^{i_2} d(i)u(i-k) &= \sum_{l=0}^{M-1} \hat{w}_l \sum_{i=i_1}^{i_2} u(i-l)u(i-k) \end{aligned}$$

and denoting

$$\begin{aligned} \Phi(l, k) &= \sum_{i=i_1}^{i_2} u(i-l)u(i-k) = \Phi(k, l) \\ \psi(k) &= \sum_{i=i_1}^{i_2} d(i)u(i-k) \end{aligned}$$

we obtain the system of equations

$$\sum_{l=0}^{M-1} \hat{w}_l \Phi(l, k) = \psi(k), \quad k = 0, 1, \dots, M - 1$$

$$\left\{ \begin{array}{l} \Phi(0,0)\hat{w}_0 + \Phi(1,0)\hat{w}_1 + \dots + \Phi(M-1,0)\hat{w}_{M-1} = \psi(0) \\ \Phi(0,1)\hat{w}_0 + \Phi(1,1)\hat{w}_1 + \dots + \Phi(M-1,1)\hat{w}_{M-1} = \psi(1) \\ \dots = \dots \\ \Phi(0,M-1)\hat{w}_0 + \Phi(1,M-1)\hat{w}_1 + \dots + \Phi(M-1,M-1)\hat{w}_{M-1} = \psi(M-1) \end{array} \right.$$

and using the vector notation

$$\underline{\psi} = [\psi(0) \ \psi(1) \ \psi(2) \ \dots \ \psi(M-1)]^T$$

we may rewrite the normal equations:

$$\begin{bmatrix} \Phi(0,0) & \Phi(1,0) & \Phi(2,0) & \dots & \Phi(M-1,0) \\ \Phi(0,1) & \Phi(1,1) & \Phi(2,1) & \dots & \Phi(M-1,1) \\ \Phi(0,2) & \Phi(1,2) & \Phi(2,2) & \dots & \Phi(M-1,2) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \Phi(0,M-1) & \Phi(1,M-1) & \Phi(2,M-1) & \dots & \Phi(M-1,M-1) \end{bmatrix} \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \\ \hat{w}_2 \\ \vdots \\ \vdots \\ \hat{w}_{M-1} \end{bmatrix} = \begin{bmatrix} \psi(0) \\ \psi(1) \\ \psi(2) \\ \vdots \\ \vdots \\ \psi(M-1) \end{bmatrix}$$

or in compact notations

$$\Phi \underline{\hat{w}} = \underline{\psi}$$

$$\underline{\hat{w}} = [\Phi]^{-1} \underline{\psi}$$

Minimum sum of Error Squares

$$\begin{aligned}
\mathcal{E}(\hat{w}) &= \sum_{i=i_1}^{i_2} [e_o(i)^2] = \sum_{i=i_1}^{i_2} e_o(i)(d(i) - y_o(i)) = \sum_{i=i_1}^{i_2} e_o(i)d(i) - \sum_{i=i_1}^{i_2} e_o(i)y_o(i) = \sum_{i=i_1}^{i_2} (d(i) - y_o(i))d(i) \\
&= \sum_{i=i_1}^{i_2} (d(i))^2 - \sum_{i=i_1}^{i_2} \sum_{l=0}^{M-1} \hat{w}_l u(i-l)d(i) = \sum_{i=i_1}^{i_2} (d(i))^2 - \sum_{l=0}^{M-1} \hat{w}_l \sum_{i=i_1}^{i_2} u(i-l)d(i) = \sum_{i=i_1}^{i_2} (d(i))^2 - \sum_{l=0}^{M-1} \hat{w}_l \psi(l) \\
&= \sum_{i=i_1}^{i_2} (d(i))^2 - \underline{\hat{w}}^T \underline{\psi} = \sum_{i=i_1}^{i_2} (d(i))^2 - \underline{\hat{w}}^T \Phi \underline{\hat{w}} = \sum_{i=i_1}^{i_2} (d(i))^2 - \underline{\psi}^T [\Phi]^{-1} \underline{\psi}
\end{aligned}$$

Compact forms using data matrices

$$A = \begin{bmatrix} u(i_1) & u(i_1-1) & u(i_1-2) & \dots & u(i_1-M+1) \\ u(i_1+1) & u(i_1) & u(i_1-1) & \dots & u(i_1-M+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u(i_2) & u(i_2-1) & u(i_2-2) & \dots & u(i_2-M+1) \end{bmatrix} = \begin{bmatrix} \underline{u}(i_1)^T \\ \underline{u}(i_1+1)^T \\ \vdots \\ \underline{u}(i_2)^T \end{bmatrix}$$

$$\begin{aligned}
A^T A &= \begin{bmatrix} u(i_1) & u(i_1+1) & u(i_1+2) & \dots & u(i_2) \\ u(i_1-1) & u(i_1) & u(i_1+1) & \dots & u(i_2-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u(i_1-M+1) & u(i_1-M+2) & u(i_1-M+3) & \dots & u(i_2-M+1) \end{bmatrix} \begin{bmatrix} u(i_1) & u(i_1-1) & u(i_1-2) & \dots & u(i_1-M+1) \\ u(i_1+1) & u(i_1) & u(i_1-1) & \dots & u(i_1-M+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u(i_2) & u(i_2-1) & u(i_2-2) & \dots & u(i_2-M+1) \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=i_1}^{i_2} u(i)^2 & \sum_{i=i_1}^{i_2} u(i)u(i-1) & \sum_{i=i_1}^{i_2} u(i)u(i-2) & \dots & \sum_{i=i_1}^{i_2} u(i)u(i-M+1) \\ \sum_{i=i_1}^{i_2} u(i-1)u(i) & \sum_{i=i_1}^{i_2} u(i-1)^2 & \sum_{i=i_1}^{i_2} u(i-1)u(i-2) & \dots & \sum_{i=i_1}^{i_2} u(i-1)u(i_1-M+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=i_1}^{i_2} u(i-M+1)u(i) & \sum_{i=i_1}^{i_2} u(i-M+1)u(i-1) & \sum_{i=i_1}^{i_2} u(i-M+1)u(i-2) & \dots & \sum_{i=i_1}^{i_2} u(i-M+1)^2 \end{bmatrix} = \Phi
\end{aligned}$$

$$\Phi = A^T A = \begin{bmatrix} \underline{u}(i_1) & \underline{u}(i_1+1) & \dots & \underline{u}(i_2) \end{bmatrix} \begin{bmatrix} \underline{u}(i_1)^T \\ \underline{u}(i_1+1)^T \\ \vdots \\ \underline{u}(i_2)^T \end{bmatrix} = \sum_{i=i_1}^{i_2} \underline{u}(i)\underline{u}(i)^T$$

$$A^T \underline{d} = \begin{bmatrix} u(i_1) & u(i_1 + 1) & u(i_1 + 2) & \dots & u(i_2) \\ u(i_1 - 1) & u(i_1) & u(i_1 + 1) & \dots & u(i_2 - 1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u(i_1 - M + 1) & u(i_1 - M + 2) & u(i_1 - M + 3) & \dots & u(i_2 - M + 1) \end{bmatrix} \begin{bmatrix} d(i_1) \\ d(i_1 + 1) \\ d(i_1 + 2) \\ \vdots \\ d(i_2) \end{bmatrix} = \begin{bmatrix} \sum_{i=i_1}^{i_2} u(i)d(i) \\ \sum_{i=i_1}^{i_2} u(i-1)d(i) \\ \sum_{i=i_1}^{i_2} u(i-2)d(i) \\ \vdots \\ \sum_{i=i_1}^{i_2} u(i-M+1)d(i) \end{bmatrix} = \underline{\psi}$$

$$\underline{\psi} = A^T \underline{d} = \begin{bmatrix} \underline{u}(i_1) & \underline{u}(i_1 + 1) & \dots & \underline{u}(i_2) \end{bmatrix} \begin{bmatrix} d(i_1) \\ d(i_1 + 1) \\ d(i_1 + 2) \\ \vdots \\ d(i_2) \end{bmatrix} = \sum_{i=i_1}^{i_2} \underline{u}(i)d(i)$$

Normal equations:

$$\begin{aligned} (A^T A) \hat{\underline{w}} &= (A^T \underline{d}) \\ \hat{\underline{w}} &= (A^T A)^{-1} A^T \underline{d} \end{aligned}$$

Minimum sum of error squares

$$\mathcal{E}(\hat{\underline{w}}) = \sum_{i=i_1}^{i_2} (d(i))^2 - \underline{\psi}^T [\Phi]^{-1} \underline{\psi} = \underline{d}^T \underline{d} - \underline{d}^T A (A^T A)^{-1} A^T \underline{d}$$

Projection operator Denote the time series provided by the output of LS filter

$$\hat{\underline{y}} = \begin{bmatrix} \hat{y}(i_1) & \hat{y}(i_1 + 1) & \hat{y}(i_1 + 2) & \dots & \hat{y}(i_2) \end{bmatrix}^T$$

$$\hat{\underline{y}} = A\hat{\underline{w}} = A(A^T A)^{-1}A^T \underline{d}$$

The matrix

$$P = A(A^T A)^{-1}A^T$$

is the projector operator onto the linear space spanned by the columns of the data matrix A .

Properties of Least-Squares estimates

Property 1 The least squares estimate $\hat{\underline{w}}$ is unbiased, provided that the measurement error process $\underline{\varepsilon}_o$ has zero mean.

Proof When discussing about unbiasedness, we assume the data was generated by a "true" parameter vector \underline{w}_o , and corrupted by the error vector $\underline{\varepsilon}_o$, therefore the model of the data is

$$\underline{d} = A\underline{w}_o + \underline{\varepsilon}_o$$

and the LS estimate can be written

$$\begin{aligned}\hat{\underline{w}} &= (A^T A)^{-1} (A^T \underline{d}) = (A^T A)^{-1} A^T (A\underline{w}_o + \underline{\varepsilon}_o) \\ &= \underline{w}_o + (A^T A)^{-1} A^T \underline{\varepsilon}_o\end{aligned}$$

Since by hypothesis $E\underline{\varepsilon}_o = 0$,

$$E\hat{\underline{w}} = \underline{w}_o + E(A^T A)^{-1} A^T \underline{\varepsilon}_o = \underline{w}_o + (A^T A)^{-1} A^T E\underline{\varepsilon}_o = \underline{w}_o$$

Property 2 When the measurement error process $\varepsilon_o(i)$ is white with zero mean and variance σ^2 , the covariance matrix of the LS estimate $\hat{\underline{w}}$ equals $\sigma^2(A^T A)^{-1}$.

Proof Under the mentioned hypothesis on $\varepsilon_o(i)$, the vector $\underline{\varepsilon}_o$ has zero mean and covariance matrix

$$E(\underline{\varepsilon}_o \underline{\varepsilon}_o^T) = \sigma^2 I$$

Now the covariance matrix of $\hat{\underline{w}}$ is

$$\begin{aligned}\text{cov}(\hat{\underline{w}}) &= E(\hat{\underline{w}} - \underline{w}_o)(\hat{\underline{w}} - \underline{w}_o)^T = E(A^T A)^{-1} A^T \underline{\varepsilon}_o \underline{\varepsilon}_o^T A (A^T A)^{-1} \\ &= (A^T A)^{-1} A^T E[\underline{\varepsilon}_o \underline{\varepsilon}_o^T] A (A^T A)^{-1} = (A^T A)^{-1} A^T \sigma^2 A (A^T A)^{-1} = \sigma^2 (A^T A)^{-1}\end{aligned}$$

Property 3 When the measurement error process $\varepsilon_o(i)$ is white with zero mean and variance σ^2 , the LS estimate \hat{w} is the best linear unbiased estimate (BLUE).

Proof Consider any unbiased estimator \tilde{w}

$$\tilde{w} = B\underline{d}$$

where B is an $M \times (N - m + 1)$ matrix, such that $E\tilde{w} = \underline{w}_o$, i.e.

$$E\tilde{w} = EB\underline{d} = EB(A\underline{w}_o + \underline{\varepsilon}_o) = BA\underline{w}_o + EB\underline{\varepsilon}_o = \underline{w}_o$$

therefore for the unbiasedness of \tilde{w} it is necessary that

$$BA = I$$

The covariance matrix of $\tilde{w} = BA\underline{w}_o + B\underline{\varepsilon}_o$ is

$$\text{cov}(\tilde{w}) = E(\tilde{w} - \underline{w}_o)(\tilde{w} - \underline{w}_o)^T = EB\underline{\varepsilon}_o\underline{\varepsilon}_o^T B^T = \sigma^2 BB^T$$

We show now that $\text{cov}(\tilde{w}) \geq \text{cov}(\hat{w})$. Consider the matrix $\Psi = B - (A^T A)^{-1} A^T$ and the product

$$\begin{aligned} \Psi\Psi^T &= (B - (A^T A)^{-1} A^T)(B - (A^T A)^{-1} A^T)^T = \\ &= BB^T - (A^T A)^{-1} A^T B^T - BA(A^T A)^{-1} + (A^T A)^{-1} A^T A(A^T A)^{-1} = BB^T - (A^T A)^{-1} \end{aligned}$$

But $\Psi\Psi^T$ is a semi-definite positive matrix (because $x^T \Psi\Psi^T x = \|\Psi^T x\|^2 \geq 0$, therefore $BB^T - (A^T A)^{-1} \geq 0$, or $\text{cov}(\tilde{w}) \geq \text{cov}(\hat{w})$), which finishes the proof of the property 3.

One can also show that:

Property 4 When the measurement error process $\varepsilon_o(i)$ is white and Gaussian, with zero mean, the LS estimate \hat{w} achieves the Cramer-Rao lower bound for unbiased estimators. Equivalently, it is said that for white Gaussian noise process the least squares is a minimum variance unbiased estimate (MVUE).

Least squares estimation using SVD (singular value decomposition)

There are mainly two forms of the normal equations:

$$\hat{\underline{w}} = \Phi^{-1} \underline{\psi}$$

which involves Φ , the time averaged correlation matrix of the input vector, and $\underline{\psi}$ which is the time averaged cross-correlation vector.

$$\hat{\underline{w}} = (A^T A)^{-1} A^T \underline{d}$$

which preserves the expression of Φ and $\underline{\psi}$ as functions of the data matrices.

The second form shows also that one can use the pseudoinverse (or Moore-Penrose generalized inverse) $A^+ = (A^T A)^{-1} A^T$ of the matrix A to express the LS estimate $\hat{\underline{w}} = A^+ \underline{d}$.

In the following we discuss numerical stable ways to compute the estimate $\hat{\underline{w}} = A^+ \underline{d}$.

We start from the system of linear equations

$$A \hat{\underline{w}} = \underline{d}$$

in which A is a $K \times M$ matrix, \underline{d} is a $K \times 1$ vector, and $\hat{\underline{w}}$ is a $M \times 1$ vector.

The SVD Theorem

Given the data matrix A there are two unitary matrices V and U such that

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

where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_W)$ and $\sigma_1 \geq \sigma_2 \geq \dots, \sigma_W > 0$ and $W \leq M$ is the rank of A .

Pseudoinverse

The pseudoinverse of the matrix A is

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

where $\Sigma^{-1} = \text{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \dots, \sigma_W^{-1})$. The expanded form is

$$A^+ = \sum_{i=1}^W \underline{v}_i \underline{u}_i^T \frac{1}{\sigma_i}$$

where \underline{v}_i are the columns of V and \underline{u}_i are the columns of U .

1. **Overdetermined system** If $K > M$ we assume that the rank $W = M$, and the inverse $(A^T A)^{-1}$ exists. Then, the pseudoinverse is given by

$$A^+ = (A^T A)^{-1} A^T$$

2. **Underdetermined system** If $K < M$ we assume that the rank $W = K$, and the inverse $(AA^T)^{-1}$ exists. Then, the pseudoinverse is given by

$$A^+ = A^T(AA^T)^{-1}$$

Minimum norm LS solution

When $\text{null}(A) \neq \emptyset$, (i.e. there is a nonzero vector \underline{y} such that $A\underline{y} = 0$) the solution of $A\underline{\hat{w}} = \underline{d}$ is nonunique. The pseudoinverse

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

provides the solution $\underline{\hat{w}} = A^+\underline{d}$ of minimum norm.