Matrix Analysis and Applications

Chapter 7: Least Squares and Orthogonal Projections

Instructor: Kai Lu

(http://seit.sysu.edu.cn/teacher/1801)

School of Electronics and Information Technology
Sun Yat-sen University

December 13, 2020

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Problem Formulation

Consider a linear system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{C}^{m}$ are given; $\boldsymbol{x} \in \mathbb{C}^{n}$ is unknown.

1) The system is said to be
   - **square** if $m = n$;
   - **overdetermined** if $m > n$, i.e., more equations than unknowns;
   - **underdetermined** if $m < n$, i.e., more unknowns than equations.

Problem Formulation

Consider a linear system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{C}^{m}$ are given; $\boldsymbol{x} \in \mathbb{C}^{n}$ is unknown.

1) The system is said to be
   - **square** if $m = n$;
   - **overdetermined** if $m > n$, i.e., more equations than unknowns;
   - **underdetermined** if $m < n$, i.e., more unknowns than equations.

2) Goal: find an $\boldsymbol{x}$ that solves or best approximates $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.
   - There are (many) cases for which there is no solution to (1), specifically, when $\boldsymbol{b} \notin \mathcal{R}(\boldsymbol{A})$;
   - There are also cases for which there exists more than one $\boldsymbol{x}$ that satisfies (1), specifically, when $\mathcal{N}(\boldsymbol{A}) \notin \{\boldsymbol{0}\}$.

Least Squares

Least squares (LS) problem: find an $\boldsymbol{x}$ that solves

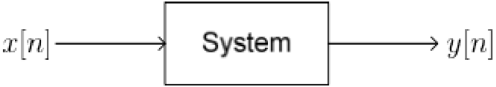
$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2. \tag{2}$$

1) Let $\boldsymbol{r} \triangleq \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ be the residual associated with $\boldsymbol{x}$. It is clear that the LS problem is in essence to find an $\boldsymbol{x}$ that has the smallest $\boldsymbol{r}$ in the 2-norm sense.

2) Has numerous applications, seen in almost all science and engineering disciplines.

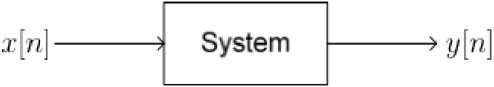
Application I: System Identification

Scenario: a discrete-time linear time-invariant system with known system input and output, and with unknown system impulse response.



Application I: System Identification

Scenario: a discrete-time linear time-invariant system with known system input and output, and with unknown system impulse response.



Model:

$$y[n] = \sum_{l=0}^{L-1} h_l x[n-l] + v[n],\ n = 0, 1, 2, \cdots \tag{3}$$

where $\{h_l\}_{l=0}^{L-1}, h_l \in \mathbb{R}$, is the system impulse response, $x[n] \in \mathbb{R}$ the system input at time index n, $y[n] \in \mathbb{R}$ the system output, $v[n] \in \mathbb{R}$ noise.

Problem: to estimate $\{h_l\}_{l=0}^{L-1}$ from $\{x[n], y[n]\}$.

Applications: channel estimation in digital communications, and many others.

Application I: System Identification (cont'd)

Let

$$\boldsymbol{x}[n] \triangleq \begin{bmatrix} x[n] \\ x[n-1] \\ \vdots \\ x[n-L+1] \end{bmatrix} \in \mathbb{R}^L, \;\; \boldsymbol{h} \triangleq \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{L-1} \end{bmatrix} \tag{4}$$

and write the model as

$$y[n] = \boldsymbol{x}^T[n]\boldsymbol{h} + v[n]. \tag{5}$$

Application I: System Identification (cont'd)

Let

$$\boldsymbol{x}[n] \triangleq \begin{bmatrix} x[n] \\ x[n-1] \\ \vdots \\ x[n-L+1] \end{bmatrix} \in \mathbb{R}^L, \ \boldsymbol{h} \triangleq \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{L-1} \end{bmatrix} \tag{4}$$

and write the model as

$$y[n] = \boldsymbol{x}^T[n]\boldsymbol{h} + v[n]. \tag{5}$$

Formulation: identify $\boldsymbol{h}$ by solving

$$\min_{\boldsymbol{h}\in\mathbb{R}^L} \sum_{n=L-1}^{N} \left|\boldsymbol{x}^T[n]\boldsymbol{h} - y[n]\right|^2, \tag{6}$$

where N is the data length, i.e., the minimum sum square error is sought.

Application I: System Identification (cont'd)

Let

$$\boldsymbol{X} \triangleq \begin{bmatrix} \boldsymbol{x}^T[L-1] \\ \boldsymbol{x}^T[L] \\ \vdots \\ \boldsymbol{x}^T[N] \end{bmatrix}, \ \boldsymbol{y} \triangleq \begin{bmatrix} y[L-1] \\ y[L] \\ \vdots \\ y[N] \end{bmatrix}, \tag{7}$$

then, the minimum sum squared-error formulation can be written as an LS problem:

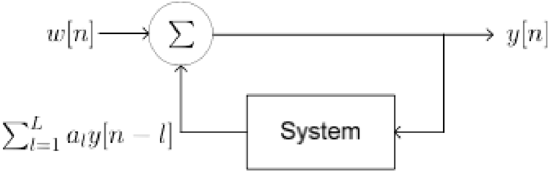
$$\min_{\boldsymbol{h} \in \mathbb{R}^L} \|\boldsymbol{X}\boldsymbol{h} - \boldsymbol{y}\|_2^2. \tag{8}$$

Application II: Linear Prediction

Autoregressive Model: a real-valued time series $y[n]$ modeled as

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + \cdots + a_L y[n-L] + \omega[n], \ n = 0, 1, 2, \cdots \quad (9)$$

for some coefficients $\{a_l\}_{l=1}^{L}$, where $\omega[n]$ is noise or modeling error.

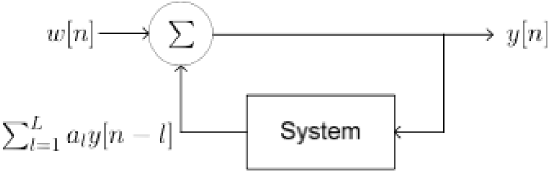


Application II: Linear Prediction

Autoregressive Model: a real-valued time series $y[n]$ modeled as

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + \cdots + a_L y[n-L] + \omega[n],\ n = 0, 1, 2, \cdots \quad (9)$$

for some coefficients $\{a_l\}_{l=1}^{L}$, where $\omega[n]$ is noise or modeling error.



Problem: to estimate $\{a_l\}_{l=1}^{L}$ from $\{y[n]\}$.

Applications: time-series prediction, speech analysis and coding, spectral estimation $\cdots$

Application II: Linear Prediction (cont'd)

Let $\boldsymbol{y}[n] \triangleq \begin{bmatrix} y[n-1] & \cdots & y[n-L] \end{bmatrix}^T$, $\boldsymbol{a} \triangleq \begin{bmatrix} a_1 & \cdots & a_L \end{bmatrix}^T$, a linear prediction formulation is given by

$$\min_{\boldsymbol{a}\in\mathbb{R}^L} \sum_{n=L}^{N} \left| \boldsymbol{y}^T[n]\boldsymbol{a} - y[n] \right|^2 . \tag{10}$$

Application II: Linear Prediction (cont'd)

Let $\boldsymbol{y}[n] \triangleq \begin{bmatrix} y[n-1] & \cdots & y[n-L] \end{bmatrix}^T$, $\boldsymbol{a} \triangleq \begin{bmatrix} a_1 & \cdots & a_L \end{bmatrix}^T$, a linear prediction formulation is given by

$$\min_{\boldsymbol{a}\in\mathbb{R}^L} \sum_{n=L}^{N} \left| \boldsymbol{y}^T[n]\boldsymbol{a} - y[n] \right|^2 . \tag{10}$$

Moreover, let

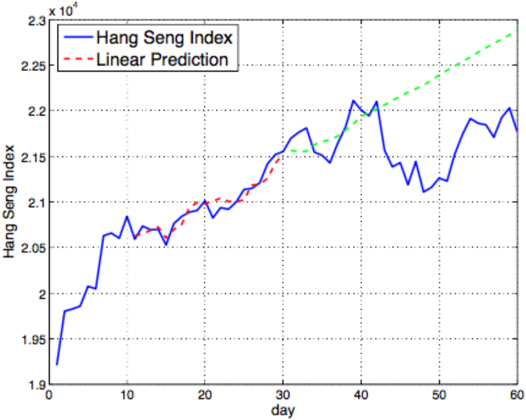
$$\boldsymbol{Y} \triangleq \begin{bmatrix} \boldsymbol{y}^T[L] \\ \vdots \\ \boldsymbol{y}^T[N] \end{bmatrix}, \ \tilde{\boldsymbol{y}} \triangleq \begin{bmatrix} y[L] \\ \vdots \\ y[N] \end{bmatrix},$$

the linear prediction formulation can be written as an LS problem:

$$\min_{\boldsymbol{a}\in\mathbb{R}^L} \|\boldsymbol{Y}\boldsymbol{a} - \tilde{\boldsymbol{y}}\|_2^2. \tag{11}$$

Application II: Linear Prediction (cont'd)

A toy demonstration of linear prediction:

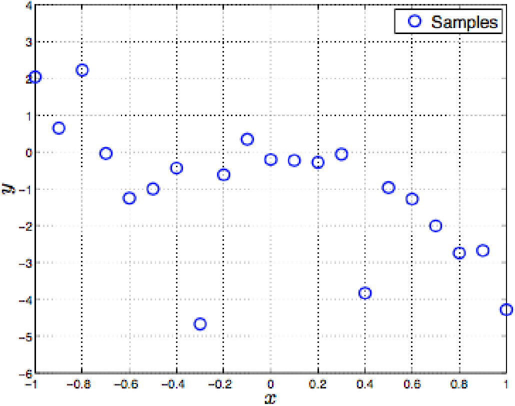


blue – Hang Seng Index during a certain time period.
red – training phase; the line is $\sum_{l=1}^{L} a_l y[n-1]$; $\boldsymbol{a}$ is obtained by LS; $L = 10$.

green – prediction phase; the line is $\tilde{y}[n] = \sum_{l=1}^{L} a_l \tilde{y}[n-l]$; the same $\boldsymbol{a}$ in the training phase.

Application III: Data Fitting

Aim: given a set of input-output data pairs (x_i, y_i), $x_i, y_i \in \mathbb{R}$, $i = 1, \cdots, m$, find a good function $f(x)$ that fits the data set well (what do you mean by “good”?)



Application III: Data Fitting (cont'd)

Model: a noisy polynomial model

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1}, \tag{12}$$

$$y_i = f(x_i) + v_i,\ i = 1, 2, \cdots, m, \tag{13}$$

where $a_0, \cdots, a_{n-1}$ are the polynomial coefficients and are unknown; v_i is noise.

Application III: Data Fitting (cont'd)

Model: a noisy polynomial model

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1}, \tag{12}$$

$$y_i = f(x_i) + v_i,\ i = 1, 2, \cdots, m, \tag{13}$$

where $a_0, \cdots, a_{n-1}$ are the polynomial coefficients and are unknown; v_i is noise.

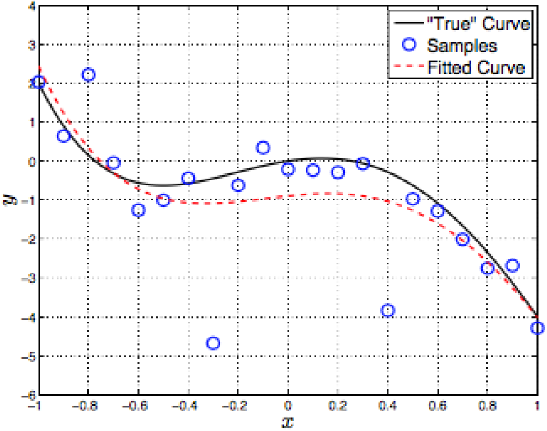
Problem: determine $f(x)$ by estimating $\boldsymbol{a} = [a_0, \cdots, a_{n-1}]^T$ from the data set. An LS fitting formulation:

$$\min_{\boldsymbol{a} \in \mathbb{R}^L} \sum_{i=1}^{m} |f(x_i) - y_i|^2 = \min_{\boldsymbol{a} \in \mathbb{R}^L} \|\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}\|_2^2, \tag{14}$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & \cdots & x_m^{n-1} \end{bmatrix},\ \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}. \tag{15}$$

Application III: Data Fitting (cont'd)



"True" curve –the true $f(\boldsymbol{x})$; the true model order is $n = 4$.

Fitted curve – the estimated $f(\boldsymbol{x})$ via LS, with the same order as the true $f(x)$.

Application III: Data Fitting (cont'd)

1) Given m data points (t_i, y_i), find n-vector $\boldsymbol{x}$ of parameters that gives 'best fit' to model function $f(t, \boldsymbol{x})$ in the sense of

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} (y_i - f(t_i, \boldsymbol{x}))^2 \,. \tag{16}$$

Application III: Data Fitting (cont'd)

1) Given m data points (t_i, y_i), find n-vector $\boldsymbol{x}$ of parameters that gives 'best fit' to model function $f(t, \boldsymbol{x})$ in the sense of

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} \left(y_i - f(t_i, \boldsymbol{x})\right)^2 . \tag{16}$$

2) Problem is linear if function f is linear in components of $\boldsymbol{x}$,

$$f(t, \boldsymbol{x}) \triangleq x_1\phi_1(t) + x_2\phi_2(t) + \cdots + x_n\phi_n(t), \tag{17}$$

where functions ϕ_j depend only upon t.

Application III: Data Fitting (cont'd)

1) Given m data points (t_i, y_i), find n-vector $\boldsymbol{x}$ of parameters that gives 'best fit' to model function $f(t, \boldsymbol{x})$ in the sense of

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} \left(y_i - f(t_i, \boldsymbol{x})\right)^2 . \tag{16}$$

2) Problem is linear if function f is linear in components of $\boldsymbol{x}$,

$$f(t, \boldsymbol{x}) \triangleq x_1\phi_1(t) + x_2\phi_2(t) + \cdots + x_n\phi_n(t), \tag{17}$$

where functions ϕ_j depend only upon t.

3) Problem can be written in matrix form as

$$\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}, \tag{18}$$

with $a_{ij} = \phi_j(t_i)$ and $b_i = y_i$.

Application III: Data Fitting (cont'd)

4) Polynomial fitting

$$f(t, \boldsymbol{x}) \triangleq x_1 + x_2 t + x_3 t^2 \cdots + x_n t^{n-1} \tag{19}$$

is linear, since polynomial linear in coefficients, though nonlinear in independent variable t.

5) Fitting sum of exponentials,

$$f(t, \boldsymbol{x}) \triangleq x_1 e^{x_1 t} + \cdots + x_n e^{x_n t} \tag{20}$$

is an example of nonlinear problem.

6) For now, we will consider only linear least squares problems.

Application III: Data Fitting (cont'd)

1) Fitting quadratic polynomial to five data points gives linear least squares problem

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cong \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \boldsymbol{b}. \tag{21}$$

2) Matrix whose columns (or rows) are successive powers of independent variable is called Vandermonde matrix.

Application III: Data Fitting–An Example

1) For data

t	-1.0	-0.5	0.0	0.5	1.0
y	1.0	0.5	0.0	0.5	2.0

overdetermined 5×3 linear system is

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} 1 & -1.0 & 1.0 \\ 1 & -0.5 & 0.25 \\ 1 & 0.0 & 0.0 \\ 1 & 0.5 & 0.25 \\ 1 & 1.0 & 1.0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cong \begin{bmatrix} 1.0 \\ 0.5 \\ 0.0 \\ 0.5 \\ 2.0 \end{bmatrix} = \boldsymbol{b}.$$

2) Solution, which we will see later how to compute, is

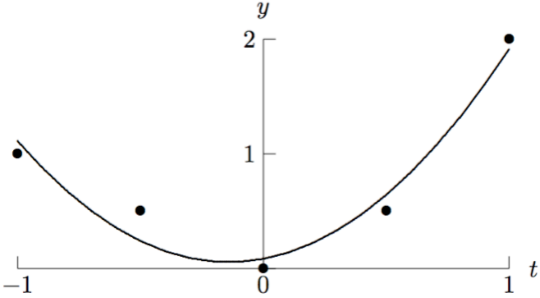
$$\boldsymbol{x} = \begin{bmatrix} 0.086 & 0.40 & 1.4 \end{bmatrix}^T,$$

so approximating polynomial is

$$f(t, \boldsymbol{x}) = 0.086 + 0.4t + 1.4t^2.$$

Application III: Data Fitting–An Example

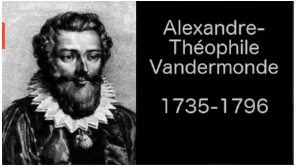
3) Resulting curve and original data points are shown in graph:



Interactive example (offline with Java)

A Short Bio of Vandermonde

Alexandre-Théophile Vandermonde was a French violinist, mathematician and chemist.



His name is now principally associated with determinant theory in mathematics but he became engaged with mathematics only around 1770 (at the age of 35). He was elected to the Académie des Sciences in 1771 after his first paper was read to the Academy in November 1770. His four (and all) mathematical papers include

1) Mémoire sur la rsolution des équations (1771),
2) Remarques sur des problèmes de situation (1771),
3) Mémoire sur des irrationnelles de différents ordres avec une application au cercle (1772),
4) Mémoire sur l'élimination (1772).

In one of his papers on the theory of music, *Système d'harmonie applicable à l'état actuel de la musique (1780)*, he put forward the idea that musicians should ignore all theory of music and rely solely on their trained ears when judging music. As a result, by the beginning of the nineteenth century the Académie des Sciences had moved music from the mathematical area to the arts area.

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
    - Overdetermined Case: the LS Problem
    - Underdetermined Case: The Minimum 2-Norm Problem
    - General Case: Pseudo-Inverse
    - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
    - Orthogonal Projector
    - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Table of Contents

Overdetermined Case: the LS Problem

Theorem 1

A vector $\boldsymbol{x}_{\mathrm{LS}} \in \mathbb{C}^n$ *is an optimal solution to the LS problem* $\min\limits_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2$ *if and only if it satisfies*

$$\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{x}_{LS} = \boldsymbol{A}^H\boldsymbol{b}. \tag{22}$$

Overdetermined Case: the LS Problem

Theorem 1

A vector $\boldsymbol{x}_{\mathrm{LS}} \in \mathbb{C}^n$ *is an optimal solution to the LS problem* $\min_{\boldsymbol{x} \in \mathbb{C}^n} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$ *if and only if it satisfies*

$$\boldsymbol{A}^H \boldsymbol{A} \boldsymbol{x}_{LS} = \boldsymbol{A}^H \boldsymbol{b}. \tag{22}$$

- (22) is called the **normal equations**;
- (22) is a square positive semi-definite (PSD) linear system.

Overdetermined Case: the LS Problem

Theorem 1

A vector $\boldsymbol{x}_{\mathrm{LS}} \in \mathbb{C}^n$ *is an optimal solution to the LS problem* $\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2$ *if and only if it satisfies*

$$\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{x}_{LS} = \boldsymbol{A}^H\boldsymbol{b}. \tag{22}$$

- (22) is called the **normal equations**;
- (22) is a square positive semi-definite (PSD) linear system.

Result: Suppose that $\boldsymbol{A}$ has full column rank (which also implies $m \geq n$). Then, $\boldsymbol{A}^H\boldsymbol{A}$ is positive definite (PD), and the LS solution is uniquely given by

$$\boldsymbol{x}_{\mathrm{LS}} = \underbrace{(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H}_{\boldsymbol{A}^\dagger}\boldsymbol{b} = \boldsymbol{A}^\dagger\boldsymbol{b}. \tag{23}$$

Proving the LS Optimality Condition: Use Opt. Principles

There is more than one way to prove Theorem 1.

Alternative 1: use results in convex optimization. Some facts in such a context:

Fact 1: consider an unconstrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}} f(\boldsymbol{x}), \tag{24}$$

where $f : \mathbb{R}^n \to \mathbb{R}$. If f is convex and differentiable, then a vector $\boldsymbol{x}^* \in \mathbb{R}^n$ is an optimal solution to the above problem if and only if

$$\Delta f(\boldsymbol{x}^*) = \mathbf{0}. \tag{25}$$

Fact 2: let $f : \mathbb{R}^n \to \mathbb{R}$. be a twice differentiable function. The function f is convex if and only if (by recalling the Hessian matrix)

$$\Delta^2 f(\boldsymbol{x}) \succeq \mathbf{0}. \tag{26}$$

Proving the LS Optimality Condition: Use Opt. Principles (cont'd)

Proof of Theorem 1: Consider the real-valued case, and let

$$\begin{aligned} f(\boldsymbol{x}) &\triangleq \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 \\ &= (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) \\ &= \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}-2\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b}+\boldsymbol{b}^T\boldsymbol{b}. \end{aligned} \tag{27}$$

1) it can be verified that

$$\Delta f(\boldsymbol{x}) = 2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}-2\boldsymbol{A}^T\boldsymbol{b}, \tag{28}$$

$$\Delta^2 f(\boldsymbol{x}) = 2\boldsymbol{A}^T\boldsymbol{A}. \tag{29}$$

2) $f(\boldsymbol{x})$ is convex since $\Delta^2 f(\boldsymbol{x}) = 2\boldsymbol{A}^T\boldsymbol{A}$ is always PSD;

3) by Fact 1, $\Delta f(\boldsymbol{x}_{\mathrm{LS}}) = 2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_{\mathrm{LS}}-2\boldsymbol{A}^T\boldsymbol{b}=\boldsymbol{0}$, which leads to (22).

Note: the complex-valued case can be solved by the same way.

Proving the LS Optimality Condition: Use Projection Theorem

Alternative 2: use the projection theorem.

Projection onto subsets: Let $\boldsymbol{z} \in \mathbb{C}^m$ and $\mathcal{S} \subseteq \mathbb{C}^m$ be given. For any $\boldsymbol{z}$, if there exists a $\boldsymbol{y}$ that satisfies

$$\min_{\boldsymbol{y}} \|\boldsymbol{z} - \boldsymbol{y}\|_2^2, \tag{30}$$

denoted $\boldsymbol{z}_s$, then $\boldsymbol{z}_s \in \mathcal{S}$ is called a projection of $\boldsymbol{z}$ onto $\mathcal{S}$.

- $\mathcal{S}$ does not need to a subspace (although we will assume).

Proving the LS Optimality Condition: Use Projection Theorem

Alternative 2: use the projection theorem.

Projection onto subsets: Let $\boldsymbol{z} \in \mathbb{C}^m$ and $\mathcal{S} \subseteq \mathbb{C}^m$ be given. For any $\boldsymbol{z}$, if there exists a $\boldsymbol{y}$ that satisfies

$$\min_{\boldsymbol{y}} \|\boldsymbol{z} - \boldsymbol{y}\|_2^2, \tag{30}$$

denoted $\boldsymbol{z}_s$, then $\boldsymbol{z}_s \in \mathcal{S}$ is called a projection of $\boldsymbol{z}$ onto $\mathcal{S}$.

- $\mathcal{S}$ does not need to a subspace (although we will assume).

Notation: Suppose that a projection of $\boldsymbol{z}$ onto $\mathcal{S}$ exists and is unique (or there exists a unique solution to (30)). The following notation will be used

$$\Pi_{\mathcal{S}}(\boldsymbol{z}) \triangleq \arg\min_{\boldsymbol{y}} \|\boldsymbol{z} - \boldsymbol{y}\|_2^2, \text{ for all } \boldsymbol{y} \in \mathcal{S} \tag{31}$$

to denote the projection of $\boldsymbol{z}$ onto $\mathcal{S}$.

Proving the LS Optimality Condition: Use Projection Theorem (cont'd)

Theorem 2 (Complex Extension of Theorem 1)

Let $\mathcal{S} \subseteq \mathbb{C}^m$ be a subspace.

1) *For every $\boldsymbol{b} \in \mathbb{C}^m$, there exists a unique projection of $\boldsymbol{b}$ onto $\mathcal{S}$;*

2) *Given $\boldsymbol{b} \in \mathbb{C}^m$, we have $\boldsymbol{b}_s = \Pi_{\mathcal{S}}(\boldsymbol{b})$ if and only if*

$$\boldsymbol{z}^H(\boldsymbol{b}_s - \boldsymbol{b}) = 0, \text{ for all } \boldsymbol{z} \in \mathcal{S}. \tag{32}$$

Proving the LS Optimality Condition: Use Projection Theorem (cont'd)

Theorem 2 (Complex Extension of Theorem 1)

Let $\mathcal{S} \subseteq \mathbb{C}^m$ be a subspace.

1) *For every $\boldsymbol{b} \in \mathbb{C}^m$, there exists a unique projection of $\boldsymbol{b}$ onto $\mathcal{S}$;*

2) *Given $\boldsymbol{b} \in \mathbb{C}^m$, we have $\boldsymbol{b}_s = \Pi_{\mathcal{S}}(\boldsymbol{b})$ if and only if*

$$\boldsymbol{z}^H(\boldsymbol{b}_s - \boldsymbol{b}) = 0, \text{ for all } \boldsymbol{z} \in \mathcal{S}. \tag{32}$$

Proof of Theorem 1: By letting $\mathcal{S} = \mathcal{R}(\boldsymbol{A})$, $\boldsymbol{b}_s = \boldsymbol{A}\boldsymbol{x}_{\text{LS}}$, $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$, and applying Theorem 2, we have

$$\boldsymbol{x}^H\boldsymbol{A}^H(\boldsymbol{A}\boldsymbol{x}_{\text{LS}} - \boldsymbol{b}) = 0, \text{ for all } \boldsymbol{x} \in \mathbb{C}^n \Longleftrightarrow \boldsymbol{A}^H(\boldsymbol{A}\boldsymbol{x}_{\text{LS}} - \boldsymbol{b}) = 0 \Longrightarrow (22).$$

Further Implication: The LS residual $\boldsymbol{r}_{\text{LS}} \triangleq \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{\text{LS}}$ is orthogonal to any $\boldsymbol{z} = \mathcal{R}(\boldsymbol{A})$.

Proving the LS Optimality Condition: Use Projection Theorem (cont'd)

- Geometric relationships among $\boldsymbol{b}$, $\boldsymbol{r}$, and span($\boldsymbol{A}$) (note that $\mathrm{span}(\boldsymbol{A}) = \mathcal{S} = \mathcal{R}(\boldsymbol{A})$)

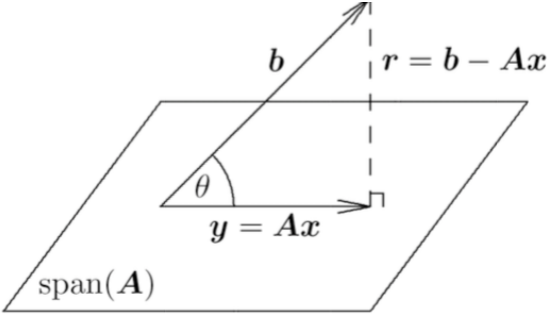

$\boldsymbol{b}$
$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$
θ
$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$
$\mathrm{span}(\boldsymbol{A})$

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Underdetermined Case: The Minimum 2-Norm Problem

Suppose that there exists a solution to $\boldsymbol{A}\boldsymbol{x} \ = \boldsymbol{b}$, but $\mathcal{N}(\boldsymbol{A}) \neq \{\mathbf{0}\}$.

- $\mathcal{N}(\boldsymbol{A}) \neq \{\mathbf{0}\}$ happens for underdetermined systems, or when $\boldsymbol{A}$ is rank-deficient.

Minimum 2-norm Problem: find an $\boldsymbol{x}$ that solves

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{x}\|_2^2, \text{ s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{33}$$

Underdetermined Case: The Minimum 2-Norm Problem

Suppose that there exists a solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, but $\mathcal{N}(\boldsymbol{A}) \neq \{\boldsymbol{0}\}$.

- $\mathcal{N}(\boldsymbol{A}) \neq \{\boldsymbol{0}\}$ happens for underdetermined systems, or when $\boldsymbol{A}$ is rank-deficient.

Minimum 2-norm Problem: find an $\boldsymbol{x}$ that solves

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{x}\|_2^2, \text{ s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{33}$$

Theorem 3

Suppose that $\boldsymbol{A}$ has full row rank. The optimal solution to the minimum 2-norm problem is unique and is given by

$$\boldsymbol{x}^* = \underbrace{\boldsymbol{A}^H(\boldsymbol{A}\boldsymbol{A}^H)^{-1}}_{\boldsymbol{A}^\dagger}\boldsymbol{b}. \tag{34}$$

Proof of Theorem 3

Step 1 (general solution): Let $\boldsymbol{A} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{V}_1^H$ denote the thin SVD of $\boldsymbol{A}$, where $\boldsymbol{U} \in \mathbb{C}^{m\times m}$, $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{m\times m}$, $\boldsymbol{V}_1 \in \mathbb{C}^{m\times m}$. The equation $\boldsymbol{A}\boldsymbol{x} \ = \boldsymbol{b}$ can be rewritten as

$$\boldsymbol{V}_1^H\boldsymbol{x} = \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}. \tag{35}$$

Now, let $\boldsymbol{V}_2 \in \mathbb{C}^{n\times(n-m)}$ such that $[\boldsymbol{V}_1, \boldsymbol{V}_2]$ is unitary. We can represent any $\boldsymbol{x} \in \mathbb{C}^n$ by

$$\boldsymbol{x} = \boldsymbol{V}_1\boldsymbol{\alpha}_1 + \boldsymbol{V}_2\boldsymbol{\alpha}_2, \tag{36}$$

where $\boldsymbol{\alpha}_1 \in \mathbb{C}^m, \boldsymbol{\alpha}_2 \in \mathbb{C}^{n-m}$. Substituting (36) into (35) yields

$$\boldsymbol{\alpha}_1 = \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}. \tag{37}$$

Also, $\boldsymbol{\alpha}_2$ does not affect the equality in (35). Hence, we conclude that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is satisfied by

$$\boldsymbol{x} = \boldsymbol{V}_1\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b} + \boldsymbol{V}_2\boldsymbol{\alpha}_2, \tag{38}$$

for any $\boldsymbol{\alpha}_2$.

Proof of Theorem 3 (cont'd)

Step 2 (norm minimization): Moreover, any x satisfying (38) has

$$\|\boldsymbol{x}\|_2^2 = \left\|\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}\right\|_2^2 + \|\boldsymbol{\alpha}_2\|_2^2 \geq \left\|\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}\right\|_2^2. \tag{39}$$

and the above equality holds if and only if $\boldsymbol{\alpha}_2 = \boldsymbol{0}$. It follows that the minimum 2-norm solution is given by $\boldsymbol{x}^* = \boldsymbol{V}_1\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}$ in a unique sense.

Proof of Theorem 3 (cont'd)

Step 2 (norm minimization): Moreover, any x satisfying (38) has

$$\|\boldsymbol{x}\|_2^2 = \left\|\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}\right\|_2^2 + \|\boldsymbol{\alpha}_2\|_2^2 \geq \left\|\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}\right\|_2^2. \tag{39}$$

and the above equality holds if and only if $\boldsymbol{\alpha}_2 = \mathbf{0}$. It follows that the minimum 2-norm solution is given by $\boldsymbol{x}^* = \boldsymbol{V}_1\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}$ in a unique sense.

Step 3 (equivalence): The last step is to show that $\boldsymbol{x}^* = \boldsymbol{V}_1\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H\boldsymbol{b}$ can be re-expressed as $\boldsymbol{x}^* = \boldsymbol{A}^H\left(\boldsymbol{A}\boldsymbol{A}^H\right)^{-1}\boldsymbol{b}$. This step is easy since

$$\begin{aligned}\boldsymbol{A}^H\left(\boldsymbol{A}\boldsymbol{A}^H\right)^{-1} &= \boldsymbol{V}_1\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}^H\left(\boldsymbol{U}\tilde{\boldsymbol{\Sigma}}^2\boldsymbol{U}^H\right)^{-1}\\ &= \boldsymbol{V}_1\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}^H\left(\boldsymbol{U}\tilde{\boldsymbol{\Sigma}}^{-2}\boldsymbol{U}^H\right)\\ &= \boldsymbol{V}_1\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^H. \end{aligned} \tag{40}$$

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

General Case: Pseudo-Inverse

Pseudo-inverse: Let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\Sigma}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^H \\ \boldsymbol{V}_2^H \end{bmatrix} = \boldsymbol{U}_1 \tilde{\boldsymbol{\Sigma}} \boldsymbol{V}_1^H$ be the thin SVD of $\boldsymbol{A}$. The pseudo-inverse of $\boldsymbol{A}$ is defined as

$$\boldsymbol{A}^\dagger \triangleq \boldsymbol{V}_1 \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{U}_1^H. \tag{41}$$

General Case: Pseudo-Inverse

Pseudo-inverse: Let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\Sigma}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^H \\ \boldsymbol{V}_2^H \end{bmatrix} = \boldsymbol{U}_1 \tilde{\boldsymbol{\Sigma}} \boldsymbol{V}_1^H$ be the thin SVD of $\boldsymbol{A}$. The pseudo-inverse of $\boldsymbol{A}$ is defined as

$$\boldsymbol{A}^\dagger \triangleq \boldsymbol{V}_1 \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{U}_1^H. \tag{41}$$

Results for pseudo-inverse:

1) **Moore-Penrose conditions**: given $\boldsymbol{A} \in \mathbb{C}^{m \times n}$, a matrix $\boldsymbol{C} \in \mathbb{C}^{n \times m}$ is said to satisfy the Moore-Penrose conditions if
   - $\boldsymbol{ACA} = \boldsymbol{A}$ ($\boldsymbol{AC}$ need not be the general identity matrix, but it maps all column vectors of $\boldsymbol{A}$ to themselves)
   - $\boldsymbol{CAC} = \boldsymbol{C}$ ($\boldsymbol{A}$ is a weak inverse for the multiplicative semigroup)
   - $(\boldsymbol{AC})^H = \boldsymbol{AC}$ ($\boldsymbol{AC}$ is Hermitian)
   - $(\boldsymbol{CA})^H = \boldsymbol{CA}$ ($\boldsymbol{CA}$ is also Hermitian)

2) The pseudo-inverse $\boldsymbol{A}^\dagger$ satisfies the Moore-Penrose conditions.
   - $\boldsymbol{A}^\dagger \boldsymbol{y}$ is an LS solution for any $\boldsymbol{A}$ (even if $\boldsymbol{A}$ is not of full column rank);
   - $\boldsymbol{A}^\dagger \boldsymbol{y}$ is the minimum 2-norm solution if $\boldsymbol{A}$ has full row rank.

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Average Performance of LS

The Standard Linear Model: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{x}$ is the 'true' result; $\boldsymbol{A}$ is assumed to have full column rank; $\boldsymbol{v}$ is noise with zero mean and covariance $\gamma^2\boldsymbol{I}$.

Average Performance of LS

The Standard Linear Model: $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{v}$, where $\boldsymbol{x}$ is the 'true' result; $\boldsymbol{A}$ is assumed to have full column rank; $\boldsymbol{v}$ is noise with zero mean and covariance $\gamma^2\boldsymbol{I}$.
Aim: analyze the mean-square error (MSE) $\mathbb{E}\left\{\|\boldsymbol{x}_{\mathrm{LS}}-\boldsymbol{x}\|_2^2\right\}$.

Solution: By noting $\boldsymbol{x}_{\mathrm{LS}}=\boldsymbol{A}^\dagger\boldsymbol{y}\xlongequal{\boldsymbol{A}^\dagger\boldsymbol{A}=(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H\boldsymbol{A}=\boldsymbol{I}}\boldsymbol{x}+\boldsymbol{A}^\dagger\boldsymbol{v}$, we have

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{x}_{\mathrm{LS}}-\boldsymbol{x}\|_2^2\right\}&=\mathbb{E}\left\{\left\|\boldsymbol{A}^\dagger\boldsymbol{v}\right\|_2^2\right\}\\&=\mathbb{E}\left\{\mathrm{tr}\left[\boldsymbol{A}^\dagger\boldsymbol{v}\boldsymbol{v}^H\left(\boldsymbol{A}^\dagger\right)^H\right]\right\}\\&=\gamma^2\mathrm{tr}\left[\boldsymbol{A}^\dagger\left(\boldsymbol{A}^\dagger\right)^H\right]\\&=\gamma^2\mathrm{tr}\left[\left(\boldsymbol{A}^H\boldsymbol{A}\right)^{-1}\right]\\&=\gamma^2\sum_{i=1}^{n}\frac{1}{\sigma_i^2(\boldsymbol{A})},\end{aligned}\tag{42}$$

where $\sigma_i(\boldsymbol{A})$ denotes the i^{th} singular value of $\boldsymbol{A}$.

Average Performance of LS

The Standard Linear Model: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{x}$ is the 'true' result; $\boldsymbol{A}$ is assumed to have full column rank; $\boldsymbol{v}$ is noise with zero mean and covariance $\gamma^2\boldsymbol{I}$.
Aim: analyze the mean-square error (MSE) $\mathbb{E}\left\{\|\boldsymbol{x}_{\rm LS} - \boldsymbol{x}\|_2^2\right\}$.

Solution: By noting $\boldsymbol{x}_{\rm LS} = \boldsymbol{A}^\dagger \boldsymbol{y} \xlongequal{\boldsymbol{A}^\dagger\boldsymbol{A} = (\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H\boldsymbol{A} = \boldsymbol{I}} \boldsymbol{x} + \boldsymbol{A}^\dagger\boldsymbol{v}$, we have

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{x}_{\rm LS} - \boldsymbol{x}\|_2^2\right\} &= \mathbb{E}\left\{\left\|\boldsymbol{A}^\dagger\boldsymbol{v}\right\|_2^2\right\}\\ &= \mathbb{E}\left\{\mathrm{tr}\left[\boldsymbol{A}^\dagger\boldsymbol{v}\boldsymbol{v}^H\left(\boldsymbol{A}^\dagger\right)^H\right]\right\}\\ &= \gamma^2\mathrm{tr}\left[\boldsymbol{A}^\dagger\left(\boldsymbol{A}^\dagger\right)^H\right]\\ &= \gamma^2\mathrm{tr}\left[\left(\boldsymbol{A}^H\boldsymbol{A}\right)^{-1}\right]\\ &= \gamma^2\sum_{i=1}^{n}\frac{1}{\sigma_i^2(\boldsymbol{A})}, \end{aligned} \tag{42}$$

where $\sigma_i(\boldsymbol{A})$ denotes the $i^{\rm th}$ singular value of $\boldsymbol{A}$.

Observation: the MSE becomes very large if some $\sigma(\boldsymbol{A})$'s are close to zero.

Gauss–Markov Theorem

The standard linear model:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}, \tag{43}$$

where $\boldsymbol{A} \in \mathcal{R}^{m\times n}$ is of full column rank, $\mathbb{E}[\boldsymbol{v}] = \boldsymbol{0}$ and $\mathrm{Cov}[\boldsymbol{v}] = \gamma^2 \boldsymbol{I}$.

Theorem 4 (Gauss–Markov Theorem)

Recalling the standard linear model given by (43)*, the minimum variance linear unbiased estimator for x_i is given by the i^{th} component of $\hat{x}_i$ in the vector*

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{A}^{\dagger}\boldsymbol{y}. \tag{44}$$

In other words, the best linear unbiased estimator for $\boldsymbol{x}$ is the least squares solution of $\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{y}$.

Proof.

See pp. 448-449 of [R1].[1] □

[1] [R1] Carl D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, 2001.

Appendix: Some Measures to Evaluate Estimators

Definition 8.1: An **estimator** $\hat{\Theta}$ is a function of the observation vector $X = (X_1, \cdots, X_n)^T$ that estimates θ but is not a function of θ.

Definition 8.2: An estimator $\hat{\Theta}$ for θ is said to be **unbiased** if and only if $\mathbb{E}[\hat{\Theta}] = \theta$. The bias in estimating θ with $\hat{\Theta}$ is

$$\mathbb{E}\left[\left|\hat{\Theta} - \theta\right|\right]. \tag{45}$$

Definition 8.3: An estimator $\hat{\Theta}$ is said to be a **linear estimator** of θ if it is a linear function of the observation vector $\mathbf{X} \triangleq (X_1, \cdots, X_n)^T$, that is,

$$\hat{\Theta} = \mathbf{b}^T\mathbf{X}. \tag{46}$$

The vector $\mathbf{b}$ is an $n \times 1$ vector of coefficients that do not depend on $\mathbf{X}$.

Appendix: Some Measures to Evaluate Estimators (cont'd)

Definition 8.4: Let $\hat{\Theta}_n$ be an estimator computed from n samples $X_1, \cdots, X_n$ for every $n \geq 1$. Then $\hat{\Theta}_n$ is said to be **consistent** if

$$\lim_{n\to\infty} \Pr\left[\left|\hat{\Theta} - \theta\right| > \epsilon\right] = 0, \text{ for every } \epsilon > 0. \tag{47}$$

The condition in (47) is referred to as **convergence in probability**.

Definition 8.5: An estimator $\hat{\Theta}$ is called **minimum-variance unbiased** if

$$\mathbb{E}\left[\left(\hat{\Theta} - \mathbb{E}\left[\hat{\Theta}\right]\right)^2\right] = \mathbb{E}\left[\left(\hat{\Theta} - \theta\right)^2\right] \leq \mathbb{E}\left[\left(\hat{\Theta}' - \theta\right)^2\right], \tag{48}$$

where $\hat{\Theta}'$ is any other estimator and $\mathbb{E}\left[\hat{\Theta}'\right] = \mathbb{E}\left[\hat{\Theta}\right] = \theta$.

Definition 8.6: An estimator $\hat{\Theta}$ is called a **minimum mean-square error** (MMSE) estimator if

$$\mathbb{E}\left[\left(\hat{\Theta} - \theta\right)^2\right] \leq \mathbb{E}\left[\left(\hat{\Theta}' - \theta\right)^2\right], \tag{49}$$

where $\hat{\Theta}'$ is any other estimator.

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Subspace

A nonempty subset $\mathcal{S}$ of $\mathbb{R}^m$ is called a **subspace** of $\mathbb{R}^m$ if, for any $\alpha, \beta \in \mathbb{R}$,

$$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S} \Longrightarrow \alpha \boldsymbol{x} + \beta \boldsymbol{y} \in \mathcal{S}. \tag{50}$$

Subspace

A nonempty subset $\mathcal{S}$ of $\mathbb{R}^m$ is called a **subspace** of $\mathbb{R}^m$ if, for any $\alpha, \beta \in \mathbb{R}$,

$$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S} \Longrightarrow \alpha \boldsymbol{x} + \beta \boldsymbol{y} \in \mathcal{S}. \tag{50}$$

Basic properties for a subspace $\mathcal{S} \subseteq \mathbb{R}^m$:

- Any subspace must include the origin (i.e., $\alpha = \beta = 0$).
- The sum of any two vectors in $\mathcal{S}$ also lies in $\mathcal{S}$, i.e., $\boldsymbol{x} + \boldsymbol{y} \in \mathcal{S}$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ (i.e., $\alpha = \beta = 1$).
- Any scalar multiplication of any vector in $\mathcal{S}$ also lies in $\mathcal{S}$, i.e., $\alpha \boldsymbol{x} \in \mathcal{S}$ for any $\alpha \in \mathbb{R}, \boldsymbol{x} \in \mathcal{S}$ (i.e., $\alpha = 0$ or $\beta = 0$).
- Any linear combination of $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n \in \mathcal{S}$, i.e., $\sum_{i=1}^{n} \alpha_i \boldsymbol{a}_i$, where $\boldsymbol{\alpha} \in \mathbb{R}^n$, also lies in $\mathcal{S}$ (straightforward extension).

Span

Given a collection of vectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n \in \mathbb{R}^m$, the **span** of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is defined as

$$\operatorname{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} = \left\{ \boldsymbol{y} = \sum_{i=1}^{n} \beta_i \boldsymbol{a}_i \,|\boldsymbol{\beta} \in \mathbb{R}^n \right\}. \tag{51}$$

Span

Given a collection of vectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n \in \mathbb{R}^m$, the **span** of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is defined as

$$\operatorname{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} = \left\{ \boldsymbol{y} = \sum_{i=1}^{n} \beta_i \boldsymbol{a}_i \,|\boldsymbol{\beta} \in \mathbb{R}^n \right\}. \tag{51}$$

- For any $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n$, $\operatorname{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is a subspace, i.e., $\mathcal{R}(\boldsymbol{A})$.
- Span is commonly used to represent or characterize a subspace, e.g.,

  $$\mathbb{R}^m = \operatorname{span}\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_m\} \tag{52}$$

  where $\boldsymbol{e}_i = [0, \cdots, 0, 1, 0, \cdots, 0]^T \in \mathbb{R}^m$ with the nonzero element being at the i^{th} entry, called **unit vectors**,
- The converse also holds, i.e., for every subspace $\mathcal{S} \subseteq \mathbb{R}^m$, there exists a positive integer n and a collection of vectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n \in \mathbb{R}^m$ such that $\mathcal{S} = \operatorname{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$.

Linearly Independent

A set of vectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n \subset \mathbb{R}^m$ is said to be **linearly independent** if

$$\sum_{i=1}^{n} \beta_i \boldsymbol{a}_i = \mathbf{0},\ \boldsymbol{\beta} \in \mathbb{R}^n \Longrightarrow \boldsymbol{\beta} = \mathbf{0}. \tag{53}$$

Otherwise, it is called **linearly dependent**.

- If $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is linearly independent, then any $\boldsymbol{a}_j$ cannot be a linear combination of the set of the other vectors $\{\boldsymbol{a}_i\}_{i\in\{1,\cdots,n\},\, i\neq j}$,
- If $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is linearly dependent, then there exists a vector a_j such that it is a linear combination of the other vectors.

Maximal Linearly Independent Subset

A subset $\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}\}$, where $i_j \in \{1, 2, \cdots, n\}$ for all $j \in [1, k]$ and $k \in [1, n]$, is called a **maximal linearly independent subset** of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$, if it is linearly independent and is not contained by any other linearly independent subset of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$.

Maximal Linearly Independent Subset

A subset $\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}\}$, where $i_j \in \{1, 2, \cdots, n\}$ for all $j \in [1, k]$ and $k \in [1, n]$, is called a **maximal linearly independent subset** of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$, if it is linearly independent and is not contained by any other linearly independent subset of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$.

- $\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}\}$ is a maximal linearly independent subset of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ if and only if $\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}, \boldsymbol{a}_j\}$ is linearly dependent for any $j \in \{1, \cdots, n\} \setminus \{i_1, \cdots, i_k\}$.
- A linearly independent $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is a non-redundant or sufficiently different set of vectors.
- A maximal linearly independent $\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}\}$ is an irreducible (and non-redundant) set of vectors for representing the whole vector set $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$.
- For any maximal linearly independent subset $\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}\}$ of $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$, we have $\mathrm{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} = \mathrm{span}\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}\}$.

Basis and Dimension

1) **Basis**: given a subspace $\mathcal{S} \subseteq \mathbb{R}^m$ with $\mathcal{S} \neq \{\mathbf{0}\}$, a set of vectors $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\} \subseteq \mathbb{R}^m$ is called a **basis** for $\mathcal{S}$ if $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\}$ is linearly independent and $\mathcal{S} = \operatorname{span}\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\}$.

   - A subspace may have more than one basis.
   - All bases for a subspace $\mathcal{S}$ have the same number of elements; i.e., if $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\}$ and $\{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_l\}$ are both bases for $\mathcal{S}$, then $k = l$.

Basis and Dimension

1) **Basis**: given a subspace $\mathcal{S} \subseteq \mathbb{R}^m$ with $\mathcal{S} \neq \{\mathbf{0}\}$, a set of vectors $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\} \subseteq \mathbb{R}^m$ is called a **basis** for $\mathcal{S}$ if $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\}$ is linearly independent and $\mathcal{S} = \text{span}\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\}$.

   - A subspace may have more than one basis.
   - All bases for a subspace $\mathcal{S}$ have the same number of elements; i.e., if $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k\}$ and $\{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_l\}$ are both bases for $\mathcal{S}$, then $k = l$.

2) **Dimension**: Given a subspace $\mathcal{S} \subseteq \mathbb{R}^m$ with $\mathcal{S} \neq \{\mathbf{0}\}$, the **dimension** of $\mathcal{S}$ is defined as the number of elements of a basis for $\mathcal{S}$.

   - The dimension of the subspace $\{\mathbf{0}\}$ is defined as zero.
   - The notation $\dim \mathcal{S}$ is used to denote the dimension of $\mathcal{S}$.

Examples:

- $\dim \mathbb{R}^m = m$.
- If $\{\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}\}$ is a maximal linearly independent subset, then $\dim \{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} = k$.

Projections onto Subspaces

Theorem 5

Let $\mathcal{S}$ be a subspace of $\mathbb{R}^m$.

1) *For every $\boldsymbol{y} \in \mathbb{R}^m$, there exists a unique vector $\boldsymbol{y}_s \in \mathbb{R}^m$ that minimizes $\|\boldsymbol{z} - \boldsymbol{y}\|_2$ over all $\boldsymbol{z} \in \mathcal{S}$. The vector $\boldsymbol{y}_s$ is thus the* **projection** *of $\boldsymbol{y}$ onto $\mathcal{S}$, and we denote it as*

$$\Pi_S(\boldsymbol{y}) \triangleq \arg\min_{\boldsymbol{z} \in \mathcal{S}} \|\boldsymbol{z} - \boldsymbol{y}\|_2^2. \tag{54}$$

2) *Given $\boldsymbol{y} \in \mathbb{R}^m$, we have $\boldsymbol{y}_s = \Pi_S(\boldsymbol{y})$ if and only if $\boldsymbol{y}_s \in \mathcal{S}$ and*

$$\boldsymbol{z}^T(\boldsymbol{y} - \boldsymbol{y}_s) = 0, \text{ for all } \boldsymbol{z} \in \mathcal{S}. \tag{55}$$

Projections onto Subspaces (cont'd)

Given a subspace $\mathcal{S} \subseteq \mathbb{R}^m$ and a vector $\boldsymbol{y} \in \mathbb{R}^m$, a **projection** of $\boldsymbol{y}$ onto $\mathcal{S}$ is a vector in $\mathcal{S}$ that is closest to $\boldsymbol{y}$ in terms of the Euclidean distance.

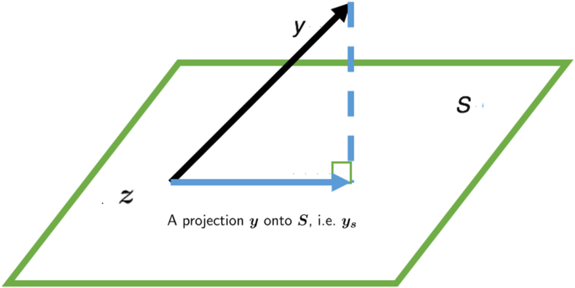

y
S
z
A projection $\boldsymbol{y}$ onto $\boldsymbol{S}$, i.e. $\boldsymbol{y}_s$

Figure 1: A projection of $\boldsymbol{y}$ onto $\mathcal{S}$.

Complementary Subspaces

Subspaces $\mathcal{X}$, $\mathcal{Y}$ of a space $\mathcal{V}$ are said to be complementary whenever

$$\mathcal{V} = \mathcal{X} + \mathcal{Y} \text{ and } \mathcal{X} \cap \mathcal{Y} = \{\mathbf{0}\}, \tag{56}$$

in which case $\mathcal{V}$ is said to be the direct sum of $\mathcal{X}$ and $\mathcal{Y}$, and this is denoted by writing $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$.

Complementary Subspaces

Subspaces $\mathcal{X}$, $\mathcal{Y}$ of a space $\mathcal{V}$ are said to be complementary whenever

$$\mathcal{V} = \mathcal{X} + \mathcal{Y} \text{ and } \mathcal{X} \cap \mathcal{Y} = \{\mathbf{0}\}, \tag{56}$$

in which case $\mathcal{V}$ is said to be the direct sum of $\mathcal{X}$ and $\mathcal{Y}$, and this is denoted by writing $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$.

For a vector space $\mathcal{V}$ with subspaces $\mathcal{X}$, $\mathcal{Y}$ having respective bases $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$, the following statements are equivalent.

- $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$.
- For each $\boldsymbol{v} \in \mathcal{V}$, there are unique vectors $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ such that $\boldsymbol{v} = \boldsymbol{x} + \boldsymbol{y}$.
- $\mathcal{B}_{\mathcal{X}} \cap \mathcal{B}_{\mathcal{Y}} = \phi$ and $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\boldsymbol{V}$.

Orthogonal Complements

Given a subspace $\mathcal{S}$ in $\mathbb{R}^m$, the **orthogonal complement** of $\mathcal{S}$ is defined as the set

$$\mathcal{S}_{\perp} = \left\{ \boldsymbol{y} \in \mathbb{R}^m \,|\, \boldsymbol{z}^T \boldsymbol{y} = 0, \text{ for all } \boldsymbol{z} \in \mathcal{S} \right\}. \tag{57}$$

- any $\boldsymbol{z} \in \mathcal{S}$, $\boldsymbol{y} \in \mathcal{S}_{\perp}$ are orthogonal, i.e., $\mathcal{S}_{\perp}$ consists of all vectors that are orthogonal to all vectors of $\mathcal{S}$;
- $\mathcal{S} \bigcap \mathcal{S}_{\perp} = \{\mathbf{0}\}$, i.e., except for point $\mathbf{0}$, the sets $\mathcal{S}$ and $\mathcal{S}_{\perp}$ are non-intersecting.

Orthogonal Complements

Given a subspace $\mathcal{S}$ in $\mathbb{R}^m$, the **orthogonal complement** of $\mathcal{S}$ is defined as the set

$$\mathcal{S}_\perp = \left\{ \boldsymbol{y} \in \mathbb{R}^m \,|\, \boldsymbol{z}^T \boldsymbol{y} = 0, \text{ for all } \boldsymbol{z} \in \mathcal{S} \right\}. \tag{57}$$

- any $\boldsymbol{z} \in \mathcal{S}$, $\boldsymbol{y} \in \mathcal{S}_\perp$ are orthogonal, i.e., $\mathcal{S}_\perp$ consists of all vectors that are orthogonal to all vectors of $\mathcal{S}$;
- $\mathcal{S} \bigcap \mathcal{S}_\perp = \{\mathbf{0}\}$, i.e., except for point $\mathbf{0}$, the sets $\mathcal{S}$ and $\mathcal{S}_\perp$ are non-intersecting.

Theorem 6

Let $\mathcal{S}$ be a subspace of $\mathbb{R}^m$.

1) *For every $\boldsymbol{y} \in \mathbb{R}^m$, there exists a unique 2-tuple $(\boldsymbol{y}_s, \boldsymbol{y}_c) \in \mathcal{S} \times \mathcal{S}_\perp$ such that $\boldsymbol{y} = \boldsymbol{y}_s + \boldsymbol{y}_c$. Also, such $(\boldsymbol{y}_s, \boldsymbol{y}_c)$ is given by $\boldsymbol{y}_s = \Pi_S(\boldsymbol{y})$, $\boldsymbol{y}_c = \boldsymbol{y} - \Pi_S(\boldsymbol{y})$.*
2) *The projection of $\boldsymbol{y}$ onto $\mathcal{S}_\perp$ is given by $\Pi_{S_\perp}(\boldsymbol{y}) = \boldsymbol{y} - \Pi_S(\boldsymbol{y})$.*

Orthogonal Complements (cont'd)

Given two subsets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^m$, denote $\mathcal{X} + \mathcal{Y} = \{\boldsymbol{x} + \boldsymbol{y} \mid \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}\}$. By recalling Theorem 6, we have

Property 1

If $\mathcal{S}$ is a subspace in $\mathbb{R}^m$, i.e., $\mathcal{S} \subseteq \mathbb{R}^m$, we have

- $\mathcal{S} + \mathcal{S}_\perp = \mathbb{R}^m$.
- $\dim \mathcal{S} + \dim \mathcal{S}_\perp = m$.
- $(\mathcal{S}_\perp)_\perp = \mathcal{S}$.

Orthogonal Complements (cont'd)

Given two subsets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^m$, denote $\mathcal{X} + \mathcal{Y} = \{\boldsymbol{x} + \boldsymbol{y} \,|\, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}\}$. By recalling Theorem 6, we have

Property 1

If $\mathcal{S}$ is a subspace in $\mathbb{R}^m$, i.e., $\mathcal{S} \subseteq \mathbb{R}^m$, we have

- $\mathcal{S} + \mathcal{S}_\perp = \mathbb{R}^m$.
- $\dim \mathcal{S} + \dim \mathcal{S}_\perp = m$.
- $(\mathcal{S}_\perp)_\perp = \mathcal{S}$.

Application: Recall that the nullspace of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is denoted by $\mathcal{N}(\boldsymbol{A}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$. Please verify that the dimension of $\mathcal{N}(\boldsymbol{A})$ is given by $\dim \mathcal{N}(\boldsymbol{A}) = n - \operatorname{rank}(\boldsymbol{A})$.

Orthogonal Complements (cont'd)

Given two subsets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^m$, denote $\mathcal{X} + \mathcal{Y} = \{\boldsymbol{x} + \boldsymbol{y} \,|\, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}\}$. By recalling Theorem 6, we have

Property 1

If $\mathcal{S}$ is a subspace in $\mathbb{R}^m$, i.e., $\mathcal{S} \subseteq \mathbb{R}^m$, we have

- $\mathcal{S} + \mathcal{S}_\perp = \mathbb{R}^m$.
- $\dim \mathcal{S} + \dim \mathcal{S}_\perp = m$.
- $(\mathcal{S}_\perp)_\perp = \mathcal{S}$.

Application: Recall that the nullspace of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is denoted by $\mathcal{N}(\boldsymbol{A}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$. Please verify that the dimension of $\mathcal{N}(\boldsymbol{A})$ is given by $\dim \mathcal{N}(\boldsymbol{A}) = n - \operatorname{rank}(\boldsymbol{A})$.

Proof.

a) $\mathcal{N}(\boldsymbol{A}) = \mathcal{R}(\boldsymbol{A}^T)_\perp \xRightarrow{\text{Property 1}} n = \dim \mathcal{R}(\boldsymbol{A}^T) + \dim \mathcal{R}(\boldsymbol{A}^T)_\perp$
b) $\dim \mathcal{R}(\boldsymbol{A}^T) = \operatorname{rank}(\boldsymbol{A}^T) = \operatorname{rank}(\boldsymbol{A})$. □

Table of Contents

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Orthogonal Projections

Let $\mathcal{S} = \mathcal{R}(\boldsymbol{A})$ for some $\boldsymbol{A} \in \mathbb{C}^{m \times n}$, and consider the corresponding projections

$$\Pi_{\mathcal{S}}(\boldsymbol{y}) \triangleq \arg\min_{\boldsymbol{z} \in \mathcal{S}} \|\boldsymbol{z} - \boldsymbol{y}\|_2^2. \tag{58}$$

Suppose that $\boldsymbol{A}$ has full column rank. Then,

$$\Pi_{\mathcal{S}}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{x}_{LS} = \underbrace{\boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H}_{\boldsymbol{P}_{\boldsymbol{A}}}\boldsymbol{y}. \tag{59}$$

Orthogonal Projections

Let $\mathcal{S} = \mathcal{R}(\boldsymbol{A})$ for some $\boldsymbol{A} \in \mathbb{C}^{m \times n}$, and consider the corresponding projections

$$\Pi_{\mathcal{S}}(\boldsymbol{y}) \triangleq \arg\min_{\boldsymbol{z} \in \mathcal{S}} \|\boldsymbol{z} - \boldsymbol{y}\|_2^2. \tag{58}$$

Suppose that $\boldsymbol{A}$ has full column rank. Then,

$$\Pi_{\mathcal{S}}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{x}_{LS} = \underbrace{\boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H}_{\boldsymbol{P}_{\boldsymbol{A}}}\boldsymbol{y}. \tag{59}$$

Orthogonal Projector: Given a full column-rank $\boldsymbol{A} \in \mathbb{C}^{m \times n}$, let

$$\boldsymbol{P}_{\boldsymbol{A}} \triangleq \boldsymbol{A}\underbrace{(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H}_{\boldsymbol{A}^\dagger} = \boldsymbol{A}\boldsymbol{A}^\dagger, \tag{60}$$

which is called the orthogonal projector onto $\mathcal{R}(\boldsymbol{A})$.

Orthogonal Projections (cont'd)

Theorem 7 (Complex Extension of Theorem 6)

Let $\mathcal{S} \subseteq \mathbb{C}^m$ be a subspace. Every $\boldsymbol{y} \in \mathbb{C}^m$ can be decomposed as

$$\boldsymbol{y} = \boldsymbol{y}_s + \boldsymbol{y}_c, \tag{61}$$

where $\boldsymbol{y}_s \in \mathcal{S}$ and $\boldsymbol{y}_c \in \mathcal{S}_\perp$. Also, $\boldsymbol{y}_s$ and $\boldsymbol{y}_c$ are uniquely given by

$$\boldsymbol{y}_s = \Pi_{\mathcal{S}}(\boldsymbol{y}), \tag{62}$$

$$\boldsymbol{y}_c = \Pi_{\mathcal{S}_\perp}(\boldsymbol{y}) = \boldsymbol{y} - \Pi_{\mathcal{S}}(\boldsymbol{y}). \tag{63}$$

Orthogonal Projections (cont'd)

Theorem 7 (Complex Extension of Theorem 6)

Let $\mathcal{S} \subseteq \mathbb{C}^m$ be a subspace. Every $\boldsymbol{y} \in \mathbb{C}^m$ can be decomposed as

$$\boldsymbol{y} = \boldsymbol{y}_s + \boldsymbol{y}_c, \tag{61}$$

where $\boldsymbol{y}_s \in \mathcal{S}$ and $\boldsymbol{y}_c \in \mathcal{S}_\perp$. Also, $\boldsymbol{y}_s$ and $\boldsymbol{y}_c$ are uniquely given by

$$\boldsymbol{y}_s = \Pi_{\mathcal{S}}(\boldsymbol{y}), \tag{62}$$

$$\boldsymbol{y}_c = \Pi_{\mathcal{S}_\perp}(\boldsymbol{y}) = \boldsymbol{y} - \Pi_{\mathcal{S}}(\boldsymbol{y}). \tag{63}$$

Now, let $\mathcal{S} = \mathcal{R}(\boldsymbol{A})$, $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ being of full column rank. Then,

$$\Pi_{\mathcal{S}_\perp}(\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_{\mathrm{LS}} = \left(\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H\right)\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{y}. \tag{64}$$

Observation: Let $\boldsymbol{P}_{\boldsymbol{A}}^\perp \triangleq \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}}$. The matrix $\boldsymbol{P}_{\boldsymbol{A}}^\perp$ is the orthogonal projector onto $\mathcal{R}(\boldsymbol{A})_\perp$.

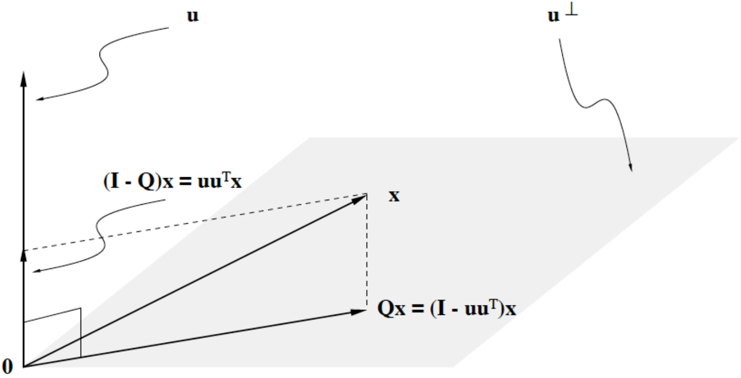


Figure 2: An illustration of orthogonal complement and orthogonal projection.

Note: Given $\boldsymbol{u}$ (a normalized line vector passing origin in $\mathbb{R}^3$), the matrix $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T$ is the orthogonal projector onto $\boldsymbol{u}^\perp$ in the sense that $\boldsymbol{Q}$ maps each $\boldsymbol{x} \in \mathbb{R}^3$ to its orthogonal projection in $\boldsymbol{u}^\perp$ (the plane containing origin that is perpendicular to $\boldsymbol{u}$).

Two Simple Applications of Orthogonal Projections

Application 1: signal (or data) subspace projection

- model: $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{v}$, where $\boldsymbol{s}$ is desired, and $\boldsymbol{v}$ is noise;
- $\boldsymbol{s}$ is unknown, but it is known to lie in a low-dimensional subspace $\mathcal{R}(\boldsymbol{A})$ for some full column-rank $\boldsymbol{A} \in \mathbb{C}^{m\times n}$;
- a solution: do projection $\tilde{\boldsymbol{s}} = \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{y}$, and use $\tilde{\boldsymbol{s}}$ as a "cleaned" version of $\boldsymbol{y}$.

Two Simple Applications of Orthogonal Projections

Application 1: signal (or data) subspace projection

- model: $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{v}$, where $\boldsymbol{s}$ is desired, and $\boldsymbol{v}$ is noise;
- $\boldsymbol{s}$ is unknown, but it is known to lie in a low-dimensional subspace $\mathcal{R}(\boldsymbol{A})$ for some full column-rank $\boldsymbol{A} \in \mathbb{C}^{m \times n}$;
- a solution: do projection $\tilde{\boldsymbol{s}} = \boldsymbol{P}_{\boldsymbol{A}} \boldsymbol{y}$, and use $\tilde{\boldsymbol{s}}$ as a "cleaned" version of $\boldsymbol{y}$.

Application 2: nulling of unwanted components

- model: $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{u} + \boldsymbol{v}$, where $\boldsymbol{s}$ is desired, $\boldsymbol{u}$ is an *unwanted* component (e.g., interference), and $\boldsymbol{v}$ is noise;
- $\boldsymbol{u}$ is unknown, but it is known to lie in a low-dimensional subspace $\mathcal{R}(\boldsymbol{B})$ for some full column-rank $\boldsymbol{B} \in \mathbb{C}^{m \times n}$;
- a solution: do projection $\boldsymbol{z} = \boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{y}$, which yields $\boldsymbol{z} = \boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{s} + \boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{v}$ and is free from the unwanted component.

Orthogonal Projections and SVD

1) **Orthogonal projector**: the notion of orthogonal projectors can be extended to general $\boldsymbol{A}$:

$$\boldsymbol{x}_{LS} = \boldsymbol{A}^{\dagger}\boldsymbol{b} \Longrightarrow \boldsymbol{A}\boldsymbol{x}_{LS} = \boldsymbol{A}\boldsymbol{A}^{\dagger}\boldsymbol{b} = \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{b}, \tag{65}$$

where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^{\dagger}$ is the **projector** onto $\mathcal{R}(\boldsymbol{A})$.

Orthogonal Projections and SVD

1) **Orthogonal projector**: the notion of orthogonal projectors can be extended to general $\boldsymbol{A}$:

$$\boldsymbol{x}_{LS} = \boldsymbol{A}^{\dagger}\boldsymbol{b} \Longrightarrow \boldsymbol{A}\boldsymbol{x}_{LS} = \boldsymbol{A}\boldsymbol{A}^{\dagger}\boldsymbol{b} = \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{b}, \tag{65}$$

where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^{\dagger}$ is the **projector** onto $\mathcal{R}(\boldsymbol{A})$.

2) **How to compute $\boldsymbol{P}_{\boldsymbol{A}}$?** Let $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{\Sigma}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_1^H \\ \boldsymbol{V}_2^H\end{bmatrix} = \boldsymbol{U}_1\tilde{\boldsymbol{\Sigma}}\boldsymbol{V}_1^H$ be the SVD and thin SVD of $\boldsymbol{A}$. Then it is easy to verify (try) that

$$\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{U}_1\boldsymbol{U}_1^H, \tag{66}$$

$$\boldsymbol{P}_{\boldsymbol{A}}^{\perp} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{U}_2\boldsymbol{U}_2^H. \tag{67}$$

Orthogonal Projections and SVD

1) **Orthogonal projector**: the notion of orthogonal projectors can be extended to general $\boldsymbol{A}$:

$$\boldsymbol{x}_{LS} = \boldsymbol{A}^{\dagger}\boldsymbol{b} \Longrightarrow \boldsymbol{A}\boldsymbol{x}_{LS} = \boldsymbol{A}\boldsymbol{A}^{\dagger}\boldsymbol{b} = \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{b}, \tag{65}$$

where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^{\dagger}$ is the **projector** onto $\mathcal{R}(\boldsymbol{A})$.

2) **How to compute $\boldsymbol{P}_{\boldsymbol{A}}$?** Let $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{\Sigma}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_1^H\\ \boldsymbol{V}_2^H\end{bmatrix} = \boldsymbol{U}_1\tilde{\boldsymbol{\Sigma}}\boldsymbol{V}_1^H$ be the SVD and thin SVD of $\boldsymbol{A}$. Then it is easy to verify (try) that

$$\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{U}_1\boldsymbol{U}_1^H, \tag{66}$$

$$\boldsymbol{P}_{\boldsymbol{A}}^{\perp} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{U}_2\boldsymbol{U}_2^H. \tag{67}$$

3) **Properties** (some are directly followed from basic SVD properties):

- $\boldsymbol{P}_{\boldsymbol{A}}^H = \boldsymbol{P}_{\boldsymbol{A}}$ (i.e., Hermitian)
- $\boldsymbol{P}_{\boldsymbol{A}}^2 = \boldsymbol{P}_{\boldsymbol{A}}$ (i.e., idempotent)
- The eigenvalues of $\boldsymbol{P}_{\boldsymbol{A}}$ are either 0 or 1
- $\mathcal{R}(\boldsymbol{P}_{\boldsymbol{A}}) = \mathcal{R}(\boldsymbol{A})$

Orthogonal Projections and SVD

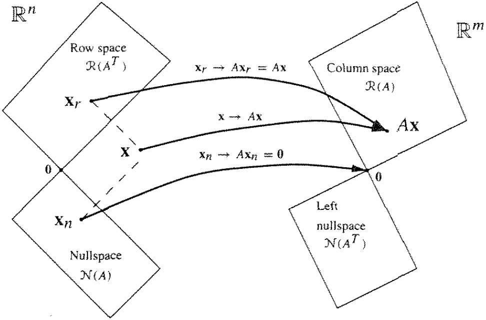


Figure 3: The four fundamental subspaces of a matrix operator.

Orthogonal Projections and SVD

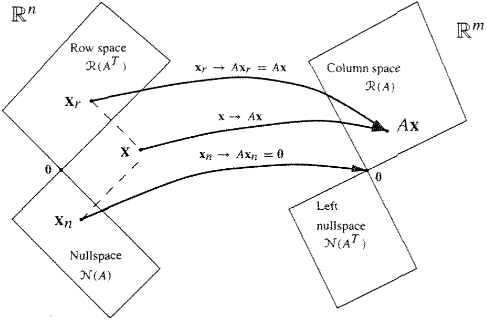


Figure 3: The four fundamental subspaces of a matrix operator.

- Projection onto $\mathcal{R}(\boldsymbol{A})$: $\boldsymbol{P_A} = \boldsymbol{U}_1\boldsymbol{U}_1^H$
- Projection onto $\mathcal{N}(\boldsymbol{A}^T)$: $\boldsymbol{P_A^\perp} = \boldsymbol{I} - \boldsymbol{P_A} = \boldsymbol{U}_2\boldsymbol{U}_2^H$
- Projection onto $\mathcal{R}(\boldsymbol{A}^T)$: $\boldsymbol{P_{A^T}} = \boldsymbol{V}_1\boldsymbol{V}_1^H$
- Projection onto $\mathcal{N}(\boldsymbol{A})$: $\boldsymbol{P_{A^T}^\perp} = \boldsymbol{I} - \boldsymbol{P_{A^T}} = \boldsymbol{V}_2\boldsymbol{V}_2^H$

Table of Contents

Angle Between Complementary Subspaces

1) Angle between nonzero vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (law of cosines):

$$\cos\theta \triangleq \frac{\boldsymbol{v}^T\boldsymbol{u}}{\|\boldsymbol{u}\|_2\,\|\boldsymbol{v}\|_2}. \tag{68}$$

Angle Between Complementary Subspaces

1) Angle between nonzero vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (law of cosines):

$$\cos\theta \triangleq \frac{\boldsymbol{v}^T\boldsymbol{u}}{\|\boldsymbol{u}\|_2\,\|\boldsymbol{v}\|_2}. \tag{68}$$

2) Angle between complementary subspaces: When $\mathbb{R}^n = \mathcal{R} \oplus \mathcal{N}$ with $\mathcal{R} \neq \mathbf{0} \neq \mathcal{N}$, the angle (a.k.a. the **minimal angle**) between $\mathcal{R}$ and $\mathcal{N}$, $0 < \theta \leq \pi/2$, is defined as

$$\cos\theta \triangleq \max_{\boldsymbol{u}\in\mathcal{R},\boldsymbol{v}\in\mathcal{N}} \frac{\boldsymbol{v}^T\boldsymbol{u}}{\|\boldsymbol{u}\|_2\,\|\boldsymbol{v}\|_2} = \max_{\substack{\boldsymbol{u}\in\mathcal{R},\boldsymbol{v}\in\mathcal{N}\\ \|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^T\boldsymbol{u}. \tag{69}$$

Angle Between Complementary Subspaces

1) Angle between nonzero vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (law of cosines):

$$\cos\theta \triangleq \frac{\boldsymbol{v}^T\boldsymbol{u}}{\|\boldsymbol{u}\|_2\,\|\boldsymbol{v}\|_2}. \tag{68}$$

2) Angle between complementary subspaces: When $\mathbb{R}^n = \mathcal{R} \oplus \mathcal{N}$ with $\mathcal{R} \neq \mathbf{0} \neq \mathcal{N}$, the angle (a.k.a. the **minimal angle**) between $\mathcal{R}$ and $\mathcal{N}$, $0 < \theta \leq \pi/2$, is defined as

$$\cos\theta \triangleq \max_{\boldsymbol{u}\in\mathcal{R},\boldsymbol{v}\in\mathcal{N}} \frac{\boldsymbol{v}^T\boldsymbol{u}}{\|\boldsymbol{u}\|_2\,\|\boldsymbol{v}\|_2} = \max_{\substack{\boldsymbol{u}\in\mathcal{R},\boldsymbol{v}\in\mathcal{N}\\ \|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^T\boldsymbol{u}. \tag{69}$$

Question: But how to numerically compute $\cos\theta$?

Angle Between Complementary Subspaces (cont'd)

Let $\boldsymbol{P}$ be the projector such that $R(\boldsymbol{P}) = \mathcal{R}$ and $N(\mathcal{P}) = \mathcal{N}$, and recall that $\|\boldsymbol{P}\|_2 = \max_{\|x\|=1} \|\boldsymbol{Px}\|_2$, which implies that $\|\boldsymbol{P}\|_2$ is the length of a longest vector in the image of the unit sphere under transformation by $\boldsymbol{P}$, as shown in the figure below.

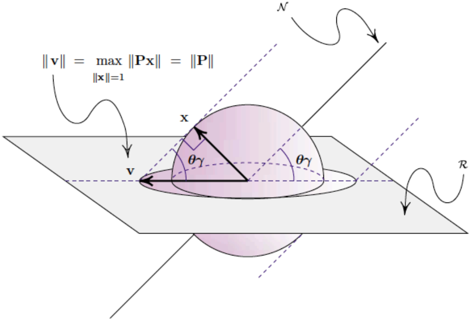


Figure 4: Angle between complementary subspaces.

Angle Between Complementary Subspaces (cont'd)

Let $\boldsymbol{P}$ be the projector such that $R(\boldsymbol{P}) = \mathcal{R}$ and $N(\mathcal{P}) = \mathcal{N}$, and recall that $\|\boldsymbol{P}\|_2 = \max_{\|x\|=1} \|\boldsymbol{P}\boldsymbol{x}\|_2$, which implies that $\|\boldsymbol{P}\|_2$ is the length of a longest vector in the image of the unit sphere under transformation by $\boldsymbol{P}$, as shown in the figure below.

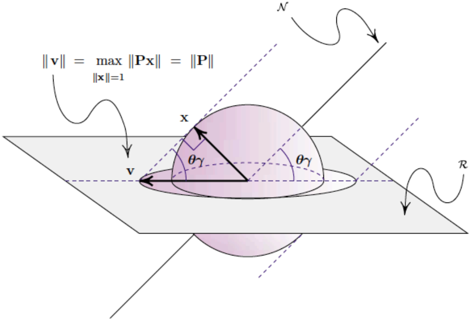


Figure 4: Angle between complementary subspaces.

$$\sin\theta = \frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{v}\|_2} = \frac{1}{\|\boldsymbol{v}\|_2} = \frac{1}{\|\boldsymbol{P}\|_2} = \frac{1}{\sigma_{\max}(\boldsymbol{P})}. \tag{70}$$

Minimal Angle Between Subspaces

3) **Minimal angle** between subspaces: the minimal angle between nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$, $0 < \theta_{\min} \leq \pi/2$, is defined as

$$\cos\theta_{\min} \triangleq \max_{\boldsymbol{u}\in\mathcal{M}, \boldsymbol{v}\in\mathcal{N}} \frac{\boldsymbol{v}^T\boldsymbol{u}}{\|\boldsymbol{u}\|_2 \, \|\boldsymbol{v}\|_2} = \max_{\substack{\boldsymbol{u}\in\mathcal{M}, \boldsymbol{v}\in\mathcal{N} \\ \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1}} \boldsymbol{v}^T\boldsymbol{u}. \tag{71}$$

Minimal Angle Between Subspaces

3) **Minimal angle** between subspaces: the minimal angle between nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$, $0 < \theta_{\min} \leq \pi/2$, is defined as

$$\cos\theta_{\min} \triangleq \max_{\boldsymbol{u}\in\mathcal{M}, \boldsymbol{v}\in\mathcal{N}} \frac{\boldsymbol{v}^T\boldsymbol{u}}{\|\boldsymbol{u}\|_2\,\|\boldsymbol{v}\|_2} = \max_{\substack{\boldsymbol{u}\in\mathcal{M}, \boldsymbol{v}\in\mathcal{N} \\ \|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^T\boldsymbol{u}. \tag{71}$$

- If $\boldsymbol{P}_{\mathcal{M}}$ and $\boldsymbol{P}_{\mathcal{N}}$ are the orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively, then

$$\cos\theta_{\min} = \|\boldsymbol{P}_{\mathcal{N}}\boldsymbol{P}_{\mathcal{M}}\|_2. \tag{72}$$

- If $\mathcal{M}$ and $\mathcal{N}$ are complementary subspaces, then $\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}$ is invertible and

$$\sin\theta_{\min} = \frac{1}{\left\|(\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}})^{-1}\right\|_2}. \tag{73}$$

The Directed Distance from $\mathcal{M}$ to $\mathcal{N}$

Remark 1 (Minimal angle between subspaces)

While the minimal angle works fine for complementary spaces, it may not convey much information about the separation between noncomplementary subspaces. For example, $\theta_{\min} = 0$ whenever $\mathcal{M}$ to $\mathcal{N}$ have a nontrivial intersection, but there nevertheless might be a nontrivial 'gap' between $\mathcal{M}$ to $\mathcal{N}$, see e.g., Fig. 5.

The Directed Distance from $\mathcal{M}$ to $\mathcal{N}$

Remark 1 (Minimal angle between subspaces)

While the minimal angle works fine for complementary spaces, it may not convey much information about the separation between noncomplementary subspaces. For example, $\theta_{\min} = 0$ whenever $\mathcal{M}$ to $\mathcal{N}$ have a nontrivial intersection, but there nevertheless might be a nontrivial 'gap' between $\mathcal{M}$ to $\mathcal{N}$, see e.g., Fig. 5.

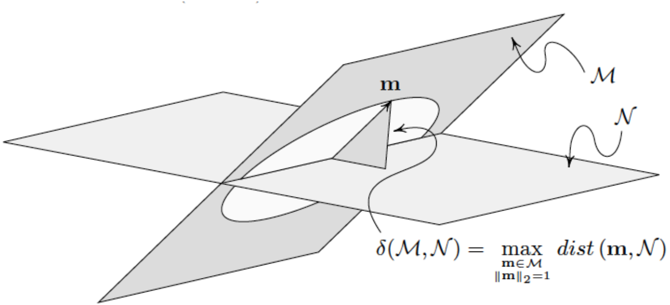

$\mathbf{m}$
$\mathcal{M}$
$\mathcal{N}$
$\delta(\mathcal{M},\mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M} \\ \|\mathbf{m}\|_2=1}} dist\,(\mathbf{m},\mathcal{N})$

Figure 5: Directed distance.

The Directed Distance from $\mathcal{M}$ to $\mathcal{N}$

Directed distance: Define the directed distance from $\mathcal{M}$ to $\mathcal{N}$ as

$$\delta(\mathcal{M},\mathcal{N}) \triangleq \max_{\substack{\boldsymbol{m}\in\mathcal{M}\\ \|\boldsymbol{m}\|_2=1}} \operatorname{dist}(\boldsymbol{m},\mathcal{N}) = \max_{\substack{\boldsymbol{m}\in\mathcal{M}\\ \|\boldsymbol{m}\|_2=1}} \|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}})\,\boldsymbol{m}\|_2, \tag{74}$$

where the **orthogonal distance** from $\boldsymbol{m}$ to $\mathcal{N}$ is defined as

$$\operatorname{dist}(\boldsymbol{m},\mathcal{N}) \triangleq \|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}})\,\boldsymbol{m}\|_2 = \|\boldsymbol{P}_{\mathcal{N}^\perp}\boldsymbol{m}\|_2 \leq \|\boldsymbol{P}_{\mathcal{N}^\perp}\|_2\,\|\boldsymbol{m}\|_2 = 1. \tag{75}$$

The Directed Distance from $\mathcal{M}$ to $\mathcal{N}$

Directed distance: Define the directed distance from $\mathcal{M}$ to $\mathcal{N}$ as

$$\delta(\mathcal{M},\mathcal{N}) \triangleq \max_{\substack{\boldsymbol{m}\in\mathcal{M} \\ \|\boldsymbol{m}\|_2=1}} \operatorname{dist}(\boldsymbol{m},\mathcal{N}) = \max_{\substack{\boldsymbol{m}\in\mathcal{M} \\ \|\boldsymbol{m}\|_2=1}} \|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}})\,\boldsymbol{m}\|_2, \qquad (74)$$

where the **orthogonal distance** from $\boldsymbol{m}$ to $\mathcal{N}$ is defined as

$$\operatorname{dist}(\boldsymbol{m},\mathcal{N}) \triangleq \|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}})\,\boldsymbol{m}\|_2 = \|\boldsymbol{P}_{\mathcal{N}^\perp}\boldsymbol{m}\|_2 \le \|\boldsymbol{P}_{\mathcal{N}^\perp}\|_2\,\|\boldsymbol{m}\|_2 = 1. \qquad (75)$$

Remark 2 (What does 'directed' mean?)

Notice that Fig. 5 is a bit misleading since $\delta(\mathcal{M},\mathcal{N}) = \delta(\mathcal{N},\mathcal{M})$ in this particular situation. However, $\delta(\mathcal{M},\mathcal{N})$ and $\delta(\mathcal{N},\mathcal{M})$ need not always agree – that's why the phrase directed distance is used. For example, if $\mathcal{M}$ is the XY-plane in $\mathcal{R}^3$ and $\mathcal{N} = \operatorname{span}\{(0,1,1)\}$, then $\delta(\mathcal{M},\mathcal{N}) = 1$ whereas $\delta(\mathcal{N},\mathcal{M}) = 1/\sqrt{2}$. Consequently, using orthogonal distance to gauge the degree of maximal separation between an arbitrary pair of subspaces requires that both values of δ be taken into account.

Gap Between Subspaces

The **gap** between subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$ is defined to be

$$\operatorname{gap}(\mathcal{M}, \mathcal{N}) = \max\left\{\delta(\mathcal{M}, \mathcal{N}),\ \delta(\mathcal{N}, \mathcal{M})\right\}, \tag{76}$$

where $\delta(\mathcal{M}, \mathcal{N}) = \max\limits_{\substack{\boldsymbol{m} \in \mathcal{M} \\ \|\boldsymbol{m}\|_2 = 1}} \operatorname{dist}(\boldsymbol{m}, \mathcal{N})$.

Gap Between Subspaces

The **gap** between subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$ is defined to be

$$\operatorname{gap}(\mathcal{M}, \mathcal{N}) = \max \left\{ \delta(\mathcal{M}, \mathcal{N}),\ \delta(\mathcal{N}, \mathcal{M}) \right\}, \tag{76}$$

where $\delta(\mathcal{M}, \mathcal{N}) = \max\limits_{\substack{\boldsymbol{m} \in \mathcal{M} \\ \|\boldsymbol{m}\|_2 = 1}} \operatorname{dist}(\boldsymbol{m}, \mathcal{N})$.

Properties:

- $\operatorname{gap}(\mathcal{M}, \mathcal{N}) = \|\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}\|_2$.
- $\operatorname{gap}(\mathcal{M}, \mathcal{N}) = \max \left\{ \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{N}})\boldsymbol{P}_{\mathcal{M}}\|_2,\ \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{M}})\boldsymbol{P}_{\mathcal{N}}\|_2 \right\}$.
- $\operatorname{gap}(\mathcal{M}, \mathcal{N}) = 1$ whenever $\dim \mathcal{M} \neq \dim \mathcal{N}$ (e.g., Remark 2).
- If $\dim \mathcal{M} = \dim \mathcal{N}$, then $\delta(\mathcal{M}, \mathcal{N}) = \delta(\mathcal{N}, \mathcal{M})$, and
  - $\operatorname{gap}(\mathcal{M}, \mathcal{N}) = 1$ when $\mathcal{M}^{\perp} \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^{\perp}$) $\neq \{\boldsymbol{0}\}$,
  - $\operatorname{gap}(\mathcal{M}, \mathcal{N}) < 1$ when $\mathcal{M}^{\perp} \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^{\perp}$) $= \{\boldsymbol{0}\}$.

Gap Between Subspaces

The **gap** between subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$ is defined to be

$$\mathrm{gap}(\mathcal{M}, \mathcal{N}) = \max\left\{\delta(\mathcal{M}, \mathcal{N}),\ \delta(\mathcal{N}, \mathcal{M})\right\}, \tag{76}$$

where $\delta(\mathcal{M}, \mathcal{N}) = \max\limits_{\substack{\boldsymbol{m} \in \mathcal{M} \\ \|\boldsymbol{m}\|_2 = 1}} \mathrm{dist}(\boldsymbol{m}, \mathcal{N})$.

Properties:

- $\mathrm{gap}(\mathcal{M}, \mathcal{N}) = \|\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}\|_2$.
- $\mathrm{gap}(\mathcal{M}, \mathcal{N}) = \max\left\{\|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{N}})\boldsymbol{P}_{\mathcal{M}}\|_2,\ \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{M}})\boldsymbol{P}_{\mathcal{N}}\|_2\right\}$.
- $\mathrm{gap}(\mathcal{M}, \mathcal{N}) = 1$ whenever $\dim\mathcal{M} \neq \dim\mathcal{N}$ (e.g., Remark 2).
- If $\dim\mathcal{M} = \dim\mathcal{N}$, then $\delta(\mathcal{M}, \mathcal{N}) = \delta(\mathcal{N}, \mathcal{M})$, and
  - $\mathrm{gap}(\mathcal{M}, \mathcal{N}) = 1$ when $\mathcal{M}^\perp \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^\perp$) $\neq \{\mathbf{0}\}$,
  - $\mathrm{gap}(\mathcal{M}, \mathcal{N}) < 1$ when $\mathcal{M}^\perp \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^\perp$) $= \{\mathbf{0}\}$.

Remark 3 (Gap and maximal angle)

Because $0 \leq \mathrm{gap}(\mathcal{M}, \mathcal{N}) \leq 1$, the gap measure defines another angle between subspaces $\mathcal{M}$ and $\mathcal{N}$, as shown below.

Maximal Angle

The **maximal angle** between subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$ is defined to be the number $0 < \theta \leq \pi/2$ for which

$$\sin\theta_{\max} = \operatorname{gap}(\mathcal{M}, \mathcal{N}) = \|\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}\|_2. \tag{77}$$

Maximal Angle

The **maximal angle** between subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$ is defined to be the number $0 < \theta \leq \pi/2$ for which

$$\sin\theta_{\max} = \operatorname{gap}(\mathcal{M}, \mathcal{N}) = \|\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}\|_2. \tag{77}$$

Remark 4 (The minimal, maximal and intermediate angles)

For applications requiring knowledge of the degree of separation between a pair of nontrivial complementary subspaces, the minimal angle does the job. Similarly, the maximal angle adequately handles the task for subspaces of equal dimension. However, neither the minimal nor maximal angle may be of much help for more general subspaces. For example, if $\mathcal{M}$ and $\mathcal{N}$ are subspaces of unequal dimension that have a nontrivial intersection, then $\theta_{\min} = 0$ and $\theta_{\max} = \pi/2$, but neither of these numbers might convey the desired information. Hence, it seems natural to try to formulate definitions of 'intermediate' angles between $\theta_{\min}$ and $\theta_{\max}$.

Principal Angles

For nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$ with $k = \min\{\dim\mathcal{M}, \dim\mathcal{N}\}$, the **principal angles** between $\mathcal{M} = \mathcal{M}_1$ and $\mathcal{N} = \mathcal{N}_1$, $0 \leq \theta_i \leq \pi/2$, are recursively defined as

$$\cos\theta_i = \max_{\substack{\boldsymbol{u}\in\mathcal{M}_i, \boldsymbol{v}\in\mathcal{N}_i \\ \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1}} \boldsymbol{v}^T\boldsymbol{u} = \boldsymbol{v}_i^T\boldsymbol{u}_i, \ i = 1, 2, \cdots, k, \tag{78}$$

where $\|\boldsymbol{u}_i\|_2 = \|\boldsymbol{v}_i\|_2 = 1$, $\mathcal{M}_i = \boldsymbol{u}_{i-1}^{\perp} \cap \mathcal{M}_{i-1}$, and $\mathcal{N}_i = \boldsymbol{v}_{i-1}^{\perp} \cap \mathcal{N}_{i-1}$.

Principal Angles

For nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathcal{R}^n$ with $k = \min\{\dim\mathcal{M}, \dim\mathcal{N}\}$, the **principal angles** between $\mathcal{M} = \mathcal{M}_1$ and $\mathcal{N} = \mathcal{N}_1$, $0 \leq \theta_i \leq \pi/2$, are recursively defined as

$$\cos\theta_i = \max_{\substack{\boldsymbol{u}\in\mathcal{M}_i, \boldsymbol{v}\in\mathcal{N}_i \\ \|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^T\boldsymbol{u} = \boldsymbol{v}_i^T\boldsymbol{u}_i, \ i = 1, 2, \cdots, k, \tag{78}$$

where $\|\boldsymbol{u}_i\|_2 = \|\boldsymbol{v}_i\|_2 = 1$, $\mathcal{M}_i = \boldsymbol{u}_{i-1}^{\perp} \cap \mathcal{M}_{i-1}$, and $\mathcal{N}_i = \boldsymbol{v}_{i-1}^{\perp} \cap \mathcal{N}_{i-1}$.

- It's possible to prove that $\theta_{\min} = \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k \leq \theta_{\max}$, where $\theta_k = \theta_{\max}$ when $\dim\mathcal{M} = \dim\mathcal{N}$.
- The vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are not uniquely defined, but the θ_i's are unique. In fact, it can be proven that the $\sin\theta_i$'s are singular values for $\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}$. Furthermore, if $\dim\mathcal{M} \geq \dim\mathcal{N} = k$, then the $\cos\theta_i$'s are the singular values of $\boldsymbol{V}_2^T\boldsymbol{U}_1$, and the $\sin\theta_i$'s are the singular values of $\boldsymbol{V}_2^T\boldsymbol{U}_2\boldsymbol{U}_2^T$, where $\boldsymbol{U} = [\boldsymbol{U}_1|\boldsymbol{U}_2]$ and $\boldsymbol{V} = [\boldsymbol{V}_1|\boldsymbol{V}_2]$ are orthogonal matrices in which the columns of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ constitute orthonormal bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively, and $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are orthonormal bases for $\mathcal{N}^{\perp}$ and $\mathcal{N}$, respectively (i.e., $\boldsymbol{P}_{\mathcal{M}} = \boldsymbol{U}_1\boldsymbol{U}_1^T$, $\boldsymbol{I} - \boldsymbol{P}_{\mathcal{M}} = \boldsymbol{U}_2\boldsymbol{U}_2^T$, $\boldsymbol{P}_{\mathcal{N}} = \boldsymbol{V}_2\boldsymbol{V}_2^T$, $\boldsymbol{I} - \boldsymbol{P}_{\mathcal{N}} = \boldsymbol{V}_1\boldsymbol{V}_1^T$).

Special Case: Distance Between Subspaces of Same Dimensions

Suppose $\mathcal{M}$ and $\mathcal{N}$ are subspaces of $\mathbb{R}^n$ and that $\dim(\mathcal{M}) = \dim(\mathcal{N})$. Then, the distance between $\mathcal{M}$ and $\mathcal{N}$ are defined as

$$\operatorname{dist}(\mathcal{M}, \mathcal{N}) = \|\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}\|_2. \tag{79}$$

[2] [R2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th Ed., The John Hopkins University Press, 2013.

Special Case: Distance Between Subspaces of Same Dimensions

Suppose $\mathcal{M}$ and $\mathcal{N}$ are subspaces of $\mathbb{R}^n$ and that $\dim(\mathcal{M}) = \dim(\mathcal{N})$. Then, the distance between $\mathcal{M}$ and $\mathcal{N}$ are defined as

$$\operatorname{dist}(\mathcal{M}, \mathcal{N}) = \|\boldsymbol{P}_{\mathcal{M}} - \boldsymbol{P}_{\mathcal{N}}\|_2. \tag{79}$$

In particular, the distance between a pair of subspaces can be characterized in terms of the blocks of a certain orthogonal matrix, as shown below.

Theorem 8

Suppose $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_2 \end{bmatrix}$ *and* $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{Z}_2 \end{bmatrix}$ *are* $n \times n$ *orthogonal matrices and* $\boldsymbol{W}_1, \boldsymbol{Z}_1 \in \mathbb{R}^{n\times k}$. *If* $\mathcal{M} \triangleq \mathcal{R}(\boldsymbol{W}_1)$ *and* $\mathcal{N} \triangleq \mathcal{R}(\boldsymbol{Z}_1)$, *then*

$$\operatorname{dist}(\mathcal{M}, \mathcal{N}) = \left\|\boldsymbol{W}_1^T \boldsymbol{Z}_2\right\|_2 = \left\|\boldsymbol{Z}_1^T \boldsymbol{W}_2\right\|_2. \tag{80}$$

Proof: See Section 2.5.3 of [R2].[2]

[2] [R2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th Ed., The John Hopkins University Press, 2013.

Application: Determine the Volume of n-Dimensional Parallelepiped

Parallelepiped: A solid in $\mathcal{R}^m$ with parallel opposing faces whose adjacent sides are defined by vectors from a linearly independent set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ is called an n-dimensional parallelepiped.

Application: Determine the Volume of n-Dimensional Parallelepiped

Parallelepiped: A solid in $\mathcal{R}^m$ with parallel opposing faces whose adjacent sides are defined by vectors from a linearly independent set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ is called an n-dimensional parallelepiped.

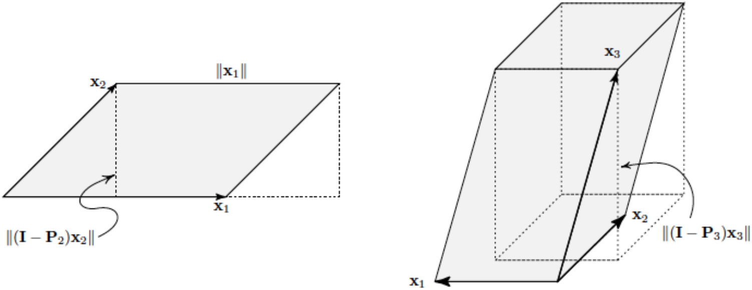


Figure 6: A parallelogram and a three-dimensional parallelepiped.

Application: Determine the Volume of n-Dimensional Parallelepiped

1) Case I ($n = 2$)
    - The **width** of the dotted rectangle: $v_1 = \|\boldsymbol{x}_1\|_2$;
    - The **projected height**: $v_2 = \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2$, where $\boldsymbol{P}_2$ is the orthogonal projector onto the space (line) spanned by $\boldsymbol{x}_1$, and $\boldsymbol{I} - \boldsymbol{P}_2$ is the orthogonal projector onto $\operatorname{span}\{\boldsymbol{x}_1\}^{\perp}$;

$$V_2 = \|\boldsymbol{x}_1\|_2 \ \left\|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\right\|_2 = v_1 v_2. \tag{81}$$

Application: Determine the Volume of n-Dimensional Parallelepiped

1) Case I ($n = 2$)
   - The **width** of the dotted rectangle: $v_1 = \|\boldsymbol{x}_1\|_2$;
   - The **projected height**: $v_2 = \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2$, where $\boldsymbol{P}_2$ is the orthogonal projector onto the space (line) spanned by $\boldsymbol{x}_1$, and $\boldsymbol{I} - \boldsymbol{P}_2$ is the orthogonal projector onto $\operatorname{span}\{\boldsymbol{x}_1\}^\perp$;

$$V_2 = \|\boldsymbol{x}_1\|_2 \; \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2 = v_1 v_2. \tag{81}$$

2) Case II ($n = 3$)
   - The area of the base: $V_2 = \|\boldsymbol{x}_1\|_2 \, \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2 = v_1 v_2$;
   - The projected height: $v_3 = \|(\boldsymbol{I} - \boldsymbol{P}_3)\boldsymbol{x}_3\|_2$, where $\boldsymbol{P}_3$ is the orthogonal projector onto $\operatorname{span}\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$;

$$V_3 = \|\boldsymbol{x}_1\|_2 \, \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2 \, \|(\boldsymbol{I} - \boldsymbol{P}_3)\boldsymbol{x}_3\|_2 = v_1 v_2 v_3. \tag{82}$$

Application: Determine the Volume of n-Dimensional Parallelepiped

1) Case I ($n = 2$)
   - The **width** of the dotted rectangle: $v_1 = \|\boldsymbol{x}_1\|_2$;
   - The **projected height**: $v_2 = \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2$, where $\boldsymbol{P}_2$ is the orthogonal projector onto the space (line) spanned by $\boldsymbol{x}_1$, and $\boldsymbol{I} - \boldsymbol{P}_2$ is the orthogonal projector onto $\mathrm{span}\{\boldsymbol{x}_1\}^\perp$;

$$V_2 = \|\boldsymbol{x}_1\|_2 \; \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2 = v_1 v_2. \tag{81}$$

2) Case II ($n = 3$)
   - The area of the base: $V_2 = \|\boldsymbol{x}_1\|_2 \, \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2 = v_1 v_2$;
   - The projected height: $v_3 = \|(\boldsymbol{I} - \boldsymbol{P}_3)\boldsymbol{x}_3\|_2$, where $\boldsymbol{P}_3$ is the orthogonal projector onto $\mathrm{span}\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$;

$$V_3 = \|\boldsymbol{x}_1\|_2 \, \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2 \, \|(\boldsymbol{I} - \boldsymbol{P}_3)\boldsymbol{x}_3\|_2 = v_1 v_2 v_3. \tag{82}$$

3) Case III ($n > 3$)

$$V_n = \|\boldsymbol{x}_1\|_2 \, \|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{x}_2\|_2 \cdots \|(\boldsymbol{I} - \boldsymbol{P}_n)\boldsymbol{x}_n\|_2 = v_1 v_2 \cdots v_n, \tag{83}$$

where $\boldsymbol{P}_k$ is the orthogonal projector onto $\mathrm{span}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{k-1}\}$. Clearly, if $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ is an orthogonal set, we have $V_n = \|\boldsymbol{x}_1\|_2 \, \|\boldsymbol{x}_2\|_2 \cdots \|\boldsymbol{x}_n\|_2$.

Table of Contents

1. Illustrative Applications (System Identification, Linear Prediction, and Data Fitting)
2. Least Squares, the Minimum 2-Norm Solution, and the Matrix Pseudo-Inverse
   - Overdetermined Case: the LS Problem
   - Underdetermined Case: The Minimum 2-Norm Problem
   - General Case: Pseudo-Inverse
   - Why LS?
3. Subspace and Linear Independence
4. Orthogonal Projections
   - Orthogonal Projector
   - Gap Between Subspaces
5. Variants of Least Squares (2-Norm Regularized LS and Total Least Squares)

Variant: 2-Norm Regularized LS

Assume a full column rank $\boldsymbol{A}$.

Intuition: replace $\boldsymbol{x}_{\mathrm{LS}} = (\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H\boldsymbol{y}$ by $\boldsymbol{x}_{\mathrm{RLS}} = (\boldsymbol{A}^H\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^H\boldsymbol{y}$, for some $\lambda > 0$, where the term $\lambda\boldsymbol{I}$ is added to improve the system conditioning (cf. Eq. (42)), thereby attempting to reduce noise sensitivity.

2-**norm regularized LS formulation**: find an $\boldsymbol{x}$ that solves

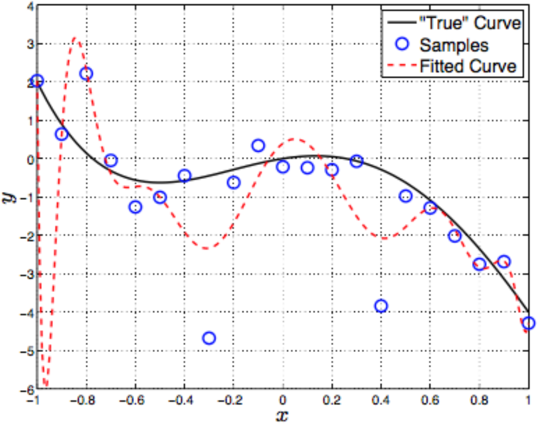
$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2+\lambda\|\boldsymbol{x}\|_2^2. \tag{84}$$

for some pre-determined constant $\lambda > 0$.

- $\lambda\|\boldsymbol{x}\|_2^2$ is a penalty term for suppressing the magnitude of $\boldsymbol{x}$;
- the solution to the above problem is

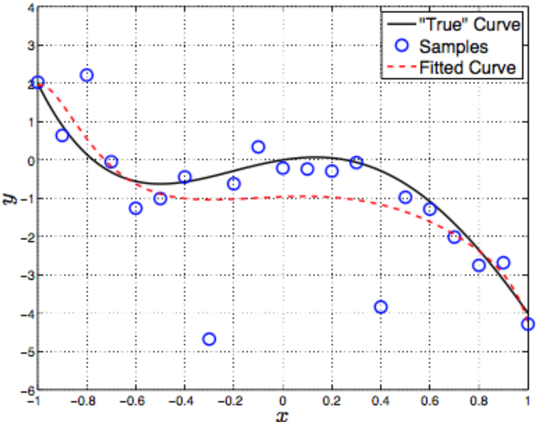
$$\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}^H\boldsymbol{y}+\lambda\boldsymbol{x}=0 \Longrightarrow \boldsymbol{x}_{\mathrm{RLS}} = \left(\boldsymbol{A}\boldsymbol{A}^H+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{A}^H\boldsymbol{y}. \tag{85}$$

Toy Demonstration: Data Fitting



- ‘True’ curve – the true $f(\boldsymbol{x})$; the true model order is $n = 4$.
- Fitted curve – the estimated $f(\boldsymbol{x})$ via LS, with a guessed model order $n = 18$.

Toy Demonstration: Data Fitting (cont'd)



- "True" curve – the true $f(\boldsymbol{x})$; the true model order is $n = 4$.
- Fitted curve – the estimated $f(\boldsymbol{x})$ via 2-norm regularized LS, with a guessed model order $n = 18$ and with $\lambda = 0.1$.

Variant: Total Least Squares

In practice, $\boldsymbol{A}$ may not be exactly known or there are measurement errors with $\boldsymbol{A}$.

Total least squares (TLS) problem:

$$\min_{\boldsymbol{E}\in\mathbb{C}^{m\times n},\,\boldsymbol{r}\in\mathbb{C}^{m},\,\boldsymbol{x}\in\mathbb{C}^{n}} \left\| \begin{bmatrix} \boldsymbol{E} & \boldsymbol{r} \end{bmatrix} \right\|_F^2, \text{ s.t. } (\boldsymbol{A}+\boldsymbol{E})\boldsymbol{x} = \boldsymbol{y}+\boldsymbol{r}. \tag{86}$$

- idea: compensate the errors in $\boldsymbol{A}$ by seeking a least squares residual.
- a hard problem at first sight.
- turns out to have a simple solution under some mild conditions.

Variant: Total Least Squares

In practice, $\boldsymbol{A}$ may not be exactly known or there are measurement errors with $\boldsymbol{A}$.

Total least squares (TLS) problem:

$$\min_{\boldsymbol{E}\in\mathbb{C}^{m\times n},\,\boldsymbol{r}\in\mathbb{C}^{m},\,\boldsymbol{x}\in\mathbb{C}^{n}} \left\| \begin{bmatrix} \boldsymbol{E} & \boldsymbol{r} \end{bmatrix} \right\|_F^2, \text{ s.t. } (\boldsymbol{A}+\boldsymbol{E})\boldsymbol{x} = \boldsymbol{y}+\boldsymbol{r}. \tag{86}$$

- idea: compensate the errors in $\boldsymbol{A}$ by seeking a least squares residual.
- a hard problem at first sight.
- turns out to have a simple solution under some mild conditions.

Theorem 9

Let $\boldsymbol{v}_i$ be the i^{th} right singular vector of matrix $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{y} \end{bmatrix}$. If $\operatorname{rank}\left(\begin{bmatrix} \boldsymbol{A} & \boldsymbol{y} \end{bmatrix}\right) = n+1$ and $\boldsymbol{v}_{n+1}(n+1) \neq 0$, then the optimal TLS solution w.r.t. $\boldsymbol{x}$ is given by

$$\boldsymbol{x}_{\text{TLS}} = -\boldsymbol{v}_{n+1}(1:n)/\boldsymbol{v}_{n+1}(n+1). \tag{87}$$

If $\boldsymbol{v}_{n+1}(n+1) = 0$, on the other hand, there is no solution.

Proof of Theorem 9

The constraint $(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{x} = \boldsymbol{y}+\boldsymbol{r}$ shown in (86) can be reformulated as

$$\begin{bmatrix} \boldsymbol{A}+\boldsymbol{E} & \boldsymbol{y}+\boldsymbol{r} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ -1 \end{bmatrix} = 0 \text{ or } \left(\begin{bmatrix} \boldsymbol{A} & \boldsymbol{y} \end{bmatrix} + \begin{bmatrix} \boldsymbol{E} & \boldsymbol{r} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{x} \\ -1 \end{bmatrix} = 0. \tag{88}$$

The SVD can be applied to find the TLS solution. Specifically,

$$\boldsymbol{C} \triangleq \begin{bmatrix} \boldsymbol{A} & \boldsymbol{y} \end{bmatrix} = \boldsymbol{U}\mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_{n+1})\boldsymbol{V}^H = \sum_{k=1}^{n+1} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^H. \tag{89}$$

Assume that the singular values are ordered with a unique smallest singular values, i.e.,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > \sigma_{n+1} > 0,$$

by recalling that the low-rank approximation to $\boldsymbol{C}$ is the matrix (Ch.5, P31)

$$\boldsymbol{C} + \boldsymbol{\Delta} = \sum_{i=1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^H. \tag{90}$$

Thus, comparing (89) with (90) yields that the perturbation is of the form

$$\boldsymbol{\Delta} = -\sigma_{n+1} \boldsymbol{u}_{n+1} \boldsymbol{v}_{n+1}^H. \tag{91}$$

Since $\mathrm{span}(\boldsymbol{C}+\boldsymbol{\Delta})$ does not contain the vector $\boldsymbol{v}_{n+1}$, the solution (88) must be a multiple of $\boldsymbol{v}_{n+1}$:

$$\begin{bmatrix} \boldsymbol{x} \\ -1 \end{bmatrix} = \alpha \boldsymbol{v}_{n+1}. \tag{92}$$

If the last component of $\boldsymbol{v}_{n+1}$ is not equal to zero, then we reach (87). Otherwise, there is no solution.

Least Squares vs. Total Least Squares

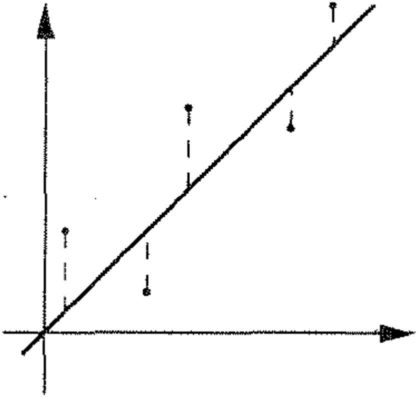


Figure 7: LS: Min. vertical distance to a line.

Least Squares vs. Total Least Squares

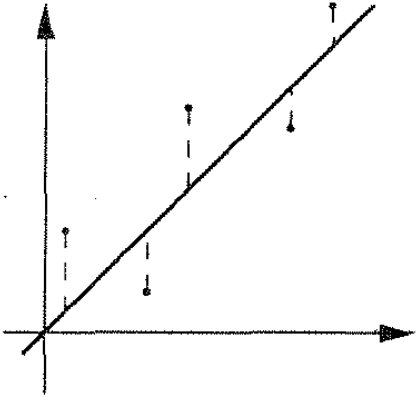


Figure 7: LS: Min. vertical distance to a line.

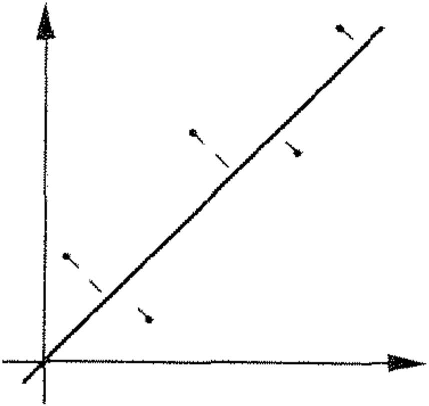


Figure 8: TLS: Min. total distance to a line.

Thank you
for your attention!

Kai Lu

E-mail: lukai86@mail.sysu.edu.cn

Web: http://seit.sysu.edu.cn/teacher/1801