MATH 532: Linear Algebra

Chapter 4: Vector Spaces

Greg Fasshauer

Department of Applied Mathematics
Illinois Institute of Technology

Spring 2015

Outline

Spaces and Subspaces

While the discussion of vector spaces can be rather dry and abstract, they are an essential tool for describing the world we work in, and to understand many practically relevant consequences.

After all, linear algebra is pretty much the workhorse of modern applied mathematics.

Moreover, many concepts we discuss now for traditional “vectors” apply also to vector spaces of functions, which form the foundation of functional analysis.

Vector Space

Definition

A set $\mathcal{V}$ of elements (vectors) is called a vector space (or linear space) over the scalar field $\mathcal{F}$ if

(A1) $\boldsymbol{x} + \boldsymbol{y} \in \mathcal{V}$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$ (closed under addition),

(A2) $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$,

(A3) $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$,

(A4) There exists a zero vector $\mathbf{0} \in \mathcal{V}$ such that $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$ for every $\boldsymbol{x} \in \mathcal{V}$,

(A5) For every $\boldsymbol{x} \in \mathcal{V}$ there is a negative $(-\boldsymbol{x}) \in \mathcal{V}$ such that $\boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}$,

(M1) $\alpha\boldsymbol{x} \in \mathcal{V}$ for every $\alpha \in \mathcal{F}$ and $\boldsymbol{x} \in \mathcal{V}$ (closed under scalar multiplication),

(M2) $(\alpha\beta)\boldsymbol{x} = \alpha(\beta\boldsymbol{x})$ for all $\alpha\beta \in \mathcal{F}$, $\boldsymbol{x} \in \mathcal{V}$,

(M3) $\alpha(\boldsymbol{x} + \boldsymbol{y}) = \alpha\boldsymbol{x} + \alpha\boldsymbol{y}$ for all $\alpha \in \mathcal{F}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$,

(M4) $(\alpha + \beta)\boldsymbol{x} = \alpha\boldsymbol{x} + \beta\boldsymbol{x}$ for all $\alpha, \beta \in \mathcal{F}$, $\boldsymbol{x} \in \mathcal{V}$,

(M5) $1\boldsymbol{x} = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{V}$.

Examples of vector spaces

- $\mathcal{V} = \mathbb{R}^m$ and $\mathcal{F} = \mathbb{R}$ (traditional real vectors)
- $\mathcal{V} = \mathbb{C}^m$ and $\mathcal{F} = \mathbb{C}$ (traditional complex vectors)
- $\mathcal{V} = \mathbb{R}^{m \times n}$ and $\mathcal{F} = \mathbb{R}$ (real matrices)
- $\mathcal{V} = \mathbb{C}^{m \times n}$ and $\mathcal{F} = \mathbb{C}$ (complex matrices)

But also

- $\mathcal{V}$ is polynomials of a certain degree with real coefficients, $\mathcal{F} = \mathbb{R}$
- $\mathcal{V}$ is continuous functions on an interval $[a, b]$, $\mathcal{F} = \mathbb{R}$

Subspaces

Definition

Let $\mathcal{S}$ be a nonempty subset of $\mathcal{V}$. If $\mathcal{S}$ is a vector space, then $\mathcal{S}$ is called a subspace of $\mathcal{V}$.

Q: What is the difference between a subset and a subspace?
A: The structure provided by the axioms (A1)–(A5), (M1)–(M5)

Theorem

The subset $\mathcal{S} \subseteq \mathcal{V}$ is a subspace of $\mathcal{V}$ if and only if

$$\alpha \boldsymbol{x} + \beta \boldsymbol{y} \in \mathcal{S} \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}, \ \alpha, \beta \in \mathcal{F}. \tag{1}$$

Remark

$\mathcal{Z} = \{\mathbf{0}\}$ *is called the trivial subspace.*

Proof.

"$\Longrightarrow$": Clear, since we actually have

$$(1) \quad \Longleftrightarrow \quad \text{(A1) and (M1)}$$

"$\Longleftarrow$": Only (A1), (A4), (A5) and (M1) need to be checked (why?).

In fact, we see that (A1) and (M1) imply (A4) and (A5):

If $\boldsymbol{x} \in \mathcal{S}$, then — using (M1) — $-1\boldsymbol{x} = -\boldsymbol{x} \in \mathcal{S}$, i.e., (A5) holds.

Using (A1), $\boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0} \in \mathcal{S}$, so that (A4) holds. $\square$

Definition

Let $\mathcal{S} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\} \subseteq \mathcal{V}$. The span of $\mathcal{S}$ is

$$\operatorname{span}(\mathcal{S}) = \left\{ \sum_{i=1}^{r} \alpha_i \boldsymbol{v}_i : \ \alpha_i \in \mathcal{F} \right\}.$$

Remark

- $\operatorname{span}(\mathcal{S})$ *contains all possible linear combinations of vectors in* $\mathcal{S}$.
- *One can easily show that* $\operatorname{span}(\mathcal{S})$ *is a subspace of* $\mathcal{V}$.

Example (Geometric interpretation)

1. If $\mathcal{S} = \{\boldsymbol{v}_1\} \subseteq \mathbb{R}^3$, then $\operatorname{span}(\mathcal{S})$ is the line through the origin with direction $\boldsymbol{v}_1$.
2. If $\mathcal{S} = \{\boldsymbol{v}_1, \boldsymbol{v}_2 : \ \boldsymbol{v}_1 \neq \alpha \boldsymbol{v}_2, \ \alpha \neq 0\} \subseteq \mathbb{R}^3$, then $\operatorname{span}(\mathcal{S})$ is the plane through the origin “spanned by” $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

Definition

Let $\mathcal{S} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\} \subseteq \mathcal{V}$. If $\operatorname{span} \mathcal{S} = \mathcal{V}$ then $\mathcal{S}$ is called a spanning set for $\mathcal{V}$.

Remark

- *A spanning set is sometimes referred to as a (finite) frame.*
- *A spanning set is not the same as a basis since the spanning set may include redundancies.*

Example

- $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$ is a spanning set for $\mathbb{R}^3$.
- $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \right\}$ is also a spanning set for $\mathbb{R}^3$.

Connection to linear systems

Theorem

Let $\mathcal{S} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$ be the set of columns of an $m \times n$ matrix A*.* span$(\mathcal{S}) = \mathbb{R}^m$ *if and only if for every $\boldsymbol{b} \in \mathbb{R}^m$ there exists an $\boldsymbol{x} \in \mathbb{R}^n$ such that* $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ *(i.e., if and only if $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ is consistent for every $\boldsymbol{b} \in \mathbb{R}^m$).*

Proof.

By definition, $\mathcal{S}$ is a spanning set for $\mathbb{R}^m$ if and only if for every $\boldsymbol{b} \in \mathbb{R}^m$ there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$\boldsymbol{b} = \alpha_1 \boldsymbol{a}_1 + \ldots + \alpha_n \boldsymbol{a}_n = \mathsf{A}\boldsymbol{x},$$

where $\mathsf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \end{pmatrix}_{m \times n}$ and $\boldsymbol{x} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$. $\square$

Remark

The sum

$$\mathcal{X} + \mathcal{Y} = \{\boldsymbol{x} + \boldsymbol{y} : \ \boldsymbol{x} \in \mathcal{X}, \ \boldsymbol{y} \in \mathcal{Y}\}$$

is a subspace of $\mathcal{V}$ *provided* $\mathcal{X}$ *and* $\mathcal{Y}$ *are subspaces.*

If $\mathcal{S}_{\mathcal{X}}$ *and* $\mathcal{S}_{\mathcal{Y}}$ *are spanning sets for* $\mathcal{X}$ *and* $\mathcal{Y}$*, respectively, then* $\mathcal{S}_{\mathcal{X}} \cup \mathcal{S}_{\mathcal{Y}}$ *is a spanning set for* $\mathcal{X} + \mathcal{Y}$*.*

Four Fundamental Subspaces

Recall that a linear function $f : \mathbb{R}^n \to \mathbb{R}^m$ satisfies

$$f(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha f(\boldsymbol{x}) + \beta f(\boldsymbol{y}) \qquad \forall \alpha, \beta \in \mathbb{R},\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Example

Let A be a real $m \times n$ matrix and

$$f(\boldsymbol{x}) = \mathsf{A}\boldsymbol{x}, \qquad \boldsymbol{x} \in \mathbb{R}^n.$$

The function f is linear since $\mathsf{A}(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha\mathsf{A}\boldsymbol{x} + \beta\mathsf{A}\boldsymbol{y}$.
Moreover, the range of f,

$$\mathcal{R}(f) = \{\mathsf{A}\boldsymbol{x} : \ \boldsymbol{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m,$$

is a subspace of $\mathbb{R}^m$ since for all $\alpha, \beta \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\alpha(\underbrace{\mathsf{A}\boldsymbol{x}}_{\in\mathcal{R}(f)}) + \beta(\underbrace{\mathsf{A}\boldsymbol{y}}_{\in\mathcal{R}(f)}) = \mathsf{A}(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) \in \mathcal{R}(f).$$

Remark

For the situation in this example we can also use the terminology range of A (or image of A), i.e.,

$$R(\mathsf{A}) = \{\mathsf{A}\boldsymbol{x} : \ \boldsymbol{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$$

Similarly,

$$R(\mathsf{A}^T) = \left\{\mathsf{A}^T\boldsymbol{y} : \ \boldsymbol{y} \in \mathbb{R}^m\right\} \subseteq \mathbb{R}^n$$

is called the range of A^T.

Column space and row space

Since

$$\mathsf{A}\boldsymbol{x} = \alpha_1 \boldsymbol{a}_1 + \ldots + \alpha_n \boldsymbol{a}_n,$$

we have $R(\mathsf{A}) = \operatorname{span}\{\boldsymbol{a}_1, \ldots \boldsymbol{a}_n\}$, i.e.,

$R(\mathsf{A})$ is the column space of A.

Similarly,

$R(\mathsf{A}^T)$ is the row space of A.

Example

Consider

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

By definition

- the columns of A span $R(A)$, i.e., they form a spanning set of $R(A)$,
- the rows of A span $R(A^T)$, i.e., they form a spanning set of $R(A^T)$,

However, since

$$(A)_{*3} = 2(A)_{*2} - (A)_{*1} \quad \text{and} \quad (A)_{3*} = 2(A)_{2*} - (A)_{1*}$$

we also have

- $R(A) = \text{span}\{(A)_{*1}, (A)_{*2}\}$
- $R(A^T) = \text{span}\{(A)_{1*}, (A)_{2*}\}$

In general, how do we find such minimal spanning sets as in the previous example?

An important tool is

Lemma

Let A, B *be* $m \times n$ *matrices. Then*

(1) $R(A^T) = R(B^T) \iff A \overset{row}{\sim} B \quad (\iff E_A = E_B)$.

(2) $R(A) = R(B) \iff A \overset{col}{\sim} B \quad (\iff E_{A^T} = E_{B^T})$.

Proof

1. "$\Longleftarrow$": Assume $A \overset{\text{row}}{\sim} B$, i.e., there exists a nonsingular matrix P such that

$$PA = B \quad \Longleftrightarrow \quad A^T P^T = B^T.$$

Now $\boldsymbol{a} \in R(A^T) \Longleftrightarrow \boldsymbol{a} = A^T \boldsymbol{y}$ for some $\boldsymbol{y}$.
We rewrite this as

$$\begin{aligned} & \boldsymbol{a} = \underbrace{A^T P^T}_{=B^T} P^{-T} \boldsymbol{y} \\ \Longleftrightarrow \quad & \boldsymbol{a} = B^T \boldsymbol{x} \quad \text{for } \boldsymbol{x} = P^{-T} \boldsymbol{y} \\ \Longleftrightarrow \quad & \boldsymbol{a} \in R(B^T). \end{aligned}$$

(cont.)

"$\Longrightarrow$": Assume $R(\mathsf{A}^T) = R(\mathsf{B}^T)$, i.e.,

$$\operatorname{span}\{(\mathsf{A})_{1*}, \ldots, (\mathsf{A})_{m*}\} = \operatorname{span}\{(\mathsf{B})_{1*}, \ldots, (\mathsf{B})_{m*}\},$$

i.e., the rows of A are linear combinations of rows of B and vice versa.

Now apply row operations to A (all collected in P) to obtain

$$\mathsf{PA} = \mathsf{B}, \quad \text{i.e., } \mathsf{A} \overset{\text{row}}{\sim} \mathsf{B}.$$

2. Let $\mathsf{A} = \mathsf{A}^T$ and $\mathsf{B} = \mathsf{B}^T$ in (1). $\square$

Theorem

Let A *be an* $m \times n$ *matrix and* U *any row echelon form obtained from* A*. Then*

1. $R(\mathsf{A}^T) =$ *span of nonzero rows of* U*.*
2. $R(\mathsf{A}) =$ *span of basic columns of* A*.*

Remark

Later we will see that any minimal span of the columns of A *forms a basis for* $R(\mathsf{A})$*.*

Proof

1. This follows from (1) in the Lemma since $A \overset{\text{row}}{\sim} U$.

2. Assume the columns of A are permuted (with a matrix Q_1) such that

$$AQ_1 = \begin{pmatrix} B & N \end{pmatrix},$$

where B contains the basic columns, and N the nonbasic columns.

By definition, the nonbasic columns are linear combinations of the basic columns, i.e., there exists a nonsingular Q_2 such that

$$\begin{pmatrix} B & N \end{pmatrix} Q_2 = \begin{pmatrix} B & O \end{pmatrix},$$

where O is a zero matrix.

(cont.)

Putting this together, we have

$$\mathsf{A}\underbrace{\mathsf{Q}_1\mathsf{Q}_2}_{=\mathsf{Q}} = \begin{pmatrix} \mathsf{B} & \mathsf{O} \end{pmatrix},$$

so that $\mathsf{A} \overset{\text{col}}{\sim} \begin{pmatrix} \mathsf{B} & \mathsf{O} \end{pmatrix}$.

(2) in the Lemma says that

$$R(\mathsf{A}) = \operatorname{span}\{\mathsf{B}_{*1}, \ldots, \mathsf{B}_{*r}\},$$

where $r = \operatorname{rank}(\mathsf{A})$. $\square$

So far, we have two of the four fundamental subspaces:

$$R(\mathsf{A}) \quad \text{and} \quad R(\mathsf{A}^T).$$

Third fundamental subspace: $N(\mathsf{A}) = \{\boldsymbol{x} : \ \mathsf{A}\boldsymbol{x} = \mathbf{0}\} \subseteq \mathbb{R}^n$,

$$\boxed{N(\mathsf{A}) \text{ is the nullspace of A}}$$

(also called the kernel of A)

Fourth fundamental subspace: $N(\mathsf{A}^T) = \{\boldsymbol{y} : \ \mathsf{A}^T\boldsymbol{y} = \mathbf{0}\} \subseteq \mathbb{R}^m$,

$$\boxed{N(\mathsf{A}^T) \text{ is the left nullspace of A}}$$

Remark

$N(\mathsf{A})$ is a linear space, i.e., a subspace of $\mathbb{R}^n$.

To see this, assume $\boldsymbol{x}, \boldsymbol{y} \in N(\mathsf{A})$, i.e., $\mathsf{A}\boldsymbol{x} = \mathsf{A}\boldsymbol{y} = \mathbf{0}$. Then

$$\mathsf{A}(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha\mathsf{A}\boldsymbol{x} + \beta\mathsf{A}\boldsymbol{y} = \mathbf{0},$$

so that $\alpha\boldsymbol{x} + \beta\boldsymbol{y} \in N(\mathsf{A})$.

How to find a (minimal) spanning set for $N(\mathsf{A})$

Find a row echelon form U of A and solve $\mathsf{U}\boldsymbol{x} = \mathbf{0}$.

Example

We can compute $\mathsf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \longrightarrow \mathsf{U} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{pmatrix}$.

So that $\mathsf{U}\boldsymbol{x} = \mathbf{0} \implies \begin{cases} x_2 &= -2x_3 \\ x_1 &= -2x_2 - 3x_3 = x_3 \end{cases}$, or

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_3 \\ -2x_3 \\ x_3 \end{pmatrix} = x_3 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Therefore

$$N(\mathsf{A}) = \operatorname{span}\left\{ \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \right\}.$$

Remark

We will see later that — as in the example — if $\text{rank}(A) = r$*, then* $N(A)$ *is spanned by* $n - r$ *vectors.*

Theorem

Let A *be an* $m \times n$ *matrix. Then*

1. $N(A) = \{\mathbf{0}\} \iff \text{rank}(A) = n.$
2. $N(A^T) = \{\mathbf{0}\} \iff \text{rank}(A) = m.$

Proof.

1. We know $\text{rank}(A) = n \iff A\boldsymbol{x} = \mathbf{0}$, but that implies $\boldsymbol{x} = \mathbf{0}$.
2. Repeat (1) with $A = A^T$ and use $\text{rank}(A^T) = \text{rank}(A)$.

$\square$

How to find a spanning set of $N(A^T)$

Theorem

Let A *be an* $m \times n$ *matrix with* $\text{rank}(A) = r$*, and let* P *be a nonsingular matrix so that* $PA = U$ *(row echelon form). Then the last* $m - r$ *rows of* P *span* $N(A^T)$.

Remark

We will later see that this spanning set is also a basis for $N(A^T)$.

Proof

Partition P as $\mathsf{P} = \begin{pmatrix} \mathsf{P}_1 \\ \mathsf{P}_2 \end{pmatrix}$, where P_1 is $r \times m$ and P_2 is $m - r \times m$.

The claim of the theorem implies that we should show that $R(\mathsf{P}_2^T) = N(\mathsf{A}^T)$.

We do this in two parts:

1. Show that $R(\mathsf{P}_2^T) \subseteq N(\mathsf{A}^T)$.
2. Show that $N(\mathsf{A}^T) \subseteq R(\mathsf{P}_2^T)$.

(cont.)

1. Partition $U_{m\times n} = \begin{pmatrix} C \\ O \end{pmatrix}$ with $C \in \mathbb{R}^{r\times n}$ and $O \in \mathbb{R}^{m-r\times n}$ (a zero matrix).
   Then
   $$PA = U \quad \Longleftrightarrow \quad \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} A = \begin{pmatrix} C \\ O \end{pmatrix} \quad \Longrightarrow \quad P_2 A = O.$$
   This also means that
   $$A^T P_2^T = O^T,$$
   i.e., every column of P_2^T is in $N(A^T)$ so that $R(P_2^T) \subseteq N(A^T)$.

(cont.)

2. Now, show $N(\mathsf{A}^T) \subseteq R(\mathsf{P}_2^T)$.
   We assume $\boldsymbol{y} \in N(\mathsf{A}^T)$ and show that then $\boldsymbol{y} \in R(\mathsf{P}_2^T)$.
   By definition,

$$\boldsymbol{y} \in N(\mathsf{A}^T) \quad \Longrightarrow \quad \mathsf{A}^T\boldsymbol{y} = \mathbf{0} \quad \Longleftrightarrow \quad \boldsymbol{y}^T\mathsf{A} = \mathbf{0}^T.$$

   Since $\mathsf{PA} = \mathsf{U} \implies \mathsf{A} = \mathsf{P}^{-1}\mathsf{U}$, and so

$$\mathbf{0}^T = \boldsymbol{y}^T\mathsf{P}^{-1}\mathsf{U} = \boldsymbol{y}^T\mathsf{P}^{-1}\begin{pmatrix} \mathsf{C} \\ \mathsf{O} \end{pmatrix}$$

   or

$$\mathbf{0}^T = \boldsymbol{y}^T\mathsf{Q}_1\mathsf{C}, \quad \text{where} \quad \mathsf{P}^{-1} = \begin{pmatrix} \underbrace{\mathsf{Q}_1}_{m\times r} & \underbrace{\mathsf{Q}_2}_{m\times m-r} \end{pmatrix}.$$

(cont.)

However, since $\operatorname{rank}(\mathsf{C}) = r$ and C is $m \times n$ we get (using $m = r$ in our earlier theorem)

$$N(\mathsf{C}^T) = \{\mathbf{0}\}$$

and therefore $\boldsymbol{y}^T \mathsf{Q}_1 = \mathbf{0}^T$.

Obviously, this implies that we also have

$$\boldsymbol{y}^T \mathsf{Q}_1 \mathsf{P}_1 = \mathbf{0}^T \tag{2}$$

(cont.)

Now $P = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$ and $P^{-1} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ so that

$$I = P^{-1}P = Q_1P_1 + Q_2P_2$$

or

$$Q_1P_1 = I - Q_2P_2. \tag{3}$$

Now we insert (3) into (2) and get

Therefore $\boldsymbol{y} \in R(P_2^T)$. $\square$

Finally,

Theorem

Let A, B *be* $m \times n$ *matrices.*

1. $N(\mathsf{A}) = N(\mathsf{B}) \quad \Longleftrightarrow \quad \mathsf{A} \overset{row}{\sim} \mathsf{B}$.
2. $N(\mathsf{A}^T) = N(\mathsf{B}^T) \quad \Longleftrightarrow \quad \mathsf{A} \overset{col}{\sim} \mathsf{B}$.

Proof.

See [Mey00, Section 4.2]. □

Linear Independence

Definition

A set of vectors $\mathcal{S} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is called linearly independent if

$$\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \ldots + \alpha_n \boldsymbol{v}_n = 0 \quad \Longrightarrow \quad \alpha_1 = \alpha_2 = \ldots = \alpha_n = 0.$$

Otherwise $\mathcal{S}$ is linearly dependent.

Remark

Linear independence is a property of a set, not of vectors.

Example

Is $\mathcal{S} = \left\{ \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix}, \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix}, \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix} \right\}$ linearly independent?

Consider

$$\alpha_1 \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\iff \quad \mathsf{A}\boldsymbol{x} = \mathbf{0}, \quad \text{where } \mathsf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \ \boldsymbol{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

Example ((cont.))

Since

$$A \overset{\text{row}}{\sim} E_A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

we know that $N(A)$ is nontrivial,
i.e., the system $A\boldsymbol{x} = \boldsymbol{0}$ has a nonzero solution, and therefore $\mathcal{S}$ is linearly dependent.

More generally,

Theorem

Let A *be an* $m \times n$ *matrix.*

1. *The columns of* A *are linearly independent if and only if* $N(\mathsf{A}) = \{\mathbf{0}\} \iff \operatorname{rank}(\mathsf{A}) = n$.
2. *The rows of* A *are linearly independent if and only if* $N(\mathsf{A}^T) = \{\mathbf{0}\} \iff \operatorname{rank}(\mathsf{A}) = m$.

Proof.

See [Mey00, Section 4.3]. $\square$

Definition

A square matrix A is called diagonally dominant if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \qquad i = 1, \ldots, n.$$

Remark

- *Aside from being nonsingular (see next slide), diagonally dominant matrices are important since they ensure that Gaussian elimination will succeed without pivoting.*
- *Also, diagonally dominance ensures convergence of certain iterative solvers (more later).*

Theorem

Let A *be an* $n \times n$ *matrix. If* A *is diagonally dominant then* A *is nonsingular.*

Proof

We will show that $N(\mathsf{A}) = \{\mathbf{0}\}$ since then we know that $\operatorname{rank}(\mathsf{A}) = n$ and A is nonsingular.

We will do this with a proof by contradiction.

We assume that there exists an $\boldsymbol{x}(\neq \mathbf{0}) \in N(\mathsf{A})$ and we will conclude that A cannot be diagonally dominant.

(cont.)

If $\boldsymbol{x} \in N(\mathsf{A})$ then $\mathsf{A}\boldsymbol{x} = \mathbf{0}$.

Now we take k so that x_k is the maximum (in absolute value) component of $\boldsymbol{x}$ and consider

$$\mathsf{A}_{k*}\boldsymbol{x} = 0.$$

We can rewrite this as

$$\sum_{j=1}^{n} a_{kj}x_j = 0 \quad \Longleftrightarrow \quad a_{kk}x_k = -\sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}x_j.$$

(cont.)

Now we take absolute values:

$$|a_{kk}x_k| = \left|\sum_{\substack{j=1\\j\neq k}}^{n} a_{kj}x_j\right| \leq \sum_{\substack{j=1\\j\neq k}}^{n} |a_{kj}|\,|x_j|$$

$$\leq \underbrace{|x_k|}_{\text{max. component}} \sum_{\substack{j=1\\j\neq k}}^{n} |a_{kj}|$$

Finally, dividing both sides by $|x_k|$ yields

$$|a_{kk}| \leq \sum_{\substack{j=1\\j\neq k}}^{n} |a_{kj}|,$$

which shows that A cannot be diagonally dominant (which is a contradiction since A was assumed to be diagonally dominant). $\square$

Example

Consider m real numbers $x_1, \ldots, x_m$ such that $x_i \neq x_j$, $i \neq j$.
Show that the columns of the Vandermonde matrix

$$\mathsf{V} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ & & & \vdots & \\ 1 & x_m & x_m^2 & \cdots & x_m^{n-1} \end{pmatrix}$$

form a linearly independent set provided $n \leq m$.
From above, the columns of V are linearly independent if and only if $N(\mathsf{V}) = \{\mathbf{0}\}$

$$\iff \quad \mathsf{V}\boldsymbol{z} = \mathbf{0} \quad \Longrightarrow \boldsymbol{z} = \mathbf{0}, \quad \boldsymbol{z} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}.$$

Example

(cont.)
Now $\mathsf{V}\boldsymbol{z} = \mathbf{0}$ if and only if

$$\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \ldots + \alpha_{n-1} x_i^{n-1} = 0, \quad i = 1, \ldots, m.$$

In other words, $x_1, x_2, \ldots, x_m$ are all (distinct) roots of

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_{n-1} x^{n-1}.$$

This is a polynomial of degree at most $n-1$.

It can have m distinct roots only if $m \le n-1$.

Otherwise, p is the zero polynomial, i.e., $\alpha_0 = \alpha_1 = \ldots = \alpha_{n-1} = 0$, so that the columns of V are linearly dependent.

The example implies that in the special case $m = n$ there is a unique polynomial of degree (at most) $m-1$ that interpolates the data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \subset \mathbb{R}^2$.
We see this by writing the polynomial in the form

$$\ell(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_{m-1} t^{m-1}.$$

Then, interpolation of the data implies

$$\ell(x_i) = y_i, \quad i = 1, \ldots, m$$

or

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{m-1} \\ & & & \vdots & \\ 1 & x_m & x_m^2 & \cdots & x_m^{m-1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{m-1} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

Since the columns of V are linearly independent it is nonsingular, and the coefficients $\alpha_0, \ldots, \alpha_{m-1}$ are uniquely determined.

In fact,

$$\ell(t) = \sum_{i=1}^{m} y_i L_i(t) \qquad \text{(Lagrange interpolation polynomial)}$$

$$\text{with } L_i(t) = \prod_{\substack{k=1\\k\neq i}}^{m} (t - x_k) \Big/ \prod_{\substack{k=1\\k\neq i}}^{m} (x_i - x_k) \qquad \text{(Lagrange functions).}$$

To verify (4) we note that the degree of ℓ is $m-1$ (since each L_i is of degree $m-1$) and

$$L_i(x_j) = \delta_{ij}, \quad i,j = 1,\ldots,m,$$

so that

$$\ell(x_j) = \sum_{i=1}^{m} y_i \underbrace{L_i(x_j)}_{=\delta_{ij}} = y_j, \quad j = 1,\ldots,m.$$

Theorem

Let $\mathcal{S} = \{\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots, \boldsymbol{u}_n\} \subseteq \mathcal{V}$ be nonempty. Then

1. *If $\mathcal{S}$ contains a linearly dependent subset, then $\mathcal{S}$ is linearly dependent.*
2. *If $\mathcal{S}$ is linearly independent, then every subset of $\mathcal{S}$ is also linearly independent.*
3. *If $\mathcal{S}$ is linearly independent and if $\boldsymbol{v} \in \mathcal{V}$, then $\mathcal{S}_{ext} = \mathcal{S} \cup \{\boldsymbol{v}\}$ is linearly independent if and only if $\boldsymbol{v} \notin \mathrm{span}(\mathcal{S})$.*
4. *If $\mathcal{S} \subseteq \mathbb{R}^m$ and $n > m$, then $\mathcal{S}$ must be linearly dependent.*

Proof

1. If $\mathcal{S}$ contains a linearly dependent subset, $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$ say, then there exist nontrivial coefficients $\alpha_1, \ldots, \alpha_k$ such that
$$\alpha_1 \boldsymbol{u}_1 + \ldots + \alpha_k \boldsymbol{u}_k = \mathbf{0}.$$
Clearly, then
$$\alpha_1 \boldsymbol{u}_1 + \ldots + \alpha_k \boldsymbol{u}_k + 0\boldsymbol{u}_{k+1} + \ldots + 0\boldsymbol{u}_n = \mathbf{0}$$
and $\mathcal{S}$ is also linearly dependent.
2. Follows from (1) by contraposition.

(cont.)

3. "$\Longrightarrow$": Assume $\mathcal{S}_{\text{ext}}$ is linearly independent. Then $\boldsymbol{v}$ can't be a linear combination of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$.

   "$\Longleftarrow$": Assume $\boldsymbol{v} \notin \text{span}(\mathcal{S})$ and consider

   $$\alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2 + \ldots + \alpha_n \boldsymbol{u}_n + \alpha_{n+1} \boldsymbol{v} = 0.$$

   First, $\alpha_{n+1} = 0$ since otherwise $\boldsymbol{v} \in \text{span}(\mathcal{S})$.

   That leaves

   $$\alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2 + \ldots + \alpha_n \boldsymbol{u}_n = 0.$$

   However, the linear independence of $\mathcal{S}$ implies $\alpha_i = 0$, $i = 1, \ldots, n$, and therefore $\mathcal{S}_{\text{ext}}$ is linearly independent.

(cont.)

4. We know that the columns of an $m \times n$ matrix A are linearly independent if and only if rank(A) $= n$.

   Here $A = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_n \end{pmatrix}$ with $\boldsymbol{u}_i \in \mathbb{R}^m$.

   If $n > m$, then rank(A) $\leq m$ and $\mathcal{S}$ must be linearly dependent. $\square$

Bases and Dimension

Earlier we introduced the concept of a spanning set of a vector space $\mathcal{V}$, i.e.,

$$\mathcal{V} = \text{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$$

Now

Definition

Consider a vector space $\mathcal{V}$ with spanning set $\mathcal{S}$. If $\mathcal{S}$ is also linearly independent then we call it a basis of $\mathcal{V}$.

Example

1. $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ is the standard basis for $\mathbb{R}^n$.
2. The columns/rows of an $n \times n$ matrix A with $\text{rank}(A) = n$ form a basis for $\mathbb{R}^n$.

Remark

Linear algebra deals with finite-dimensional linear spaces.

Functional analysis can be considered as infinite-dimensional linear algebra, where the linear spaces are usually function spaces such as

- *infinitely differentiable functions with Taylor (polynomial) basis*

$$\{1, x, x^2, x^3, \ldots\}$$

- *square integrable functions with Fourier basis*

$$\{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots\}$$

Earlier we mentioned the idea of minimal spanning sets.

Theorem

Let $\mathcal{V}$ be a subspace of $\mathbb{R}^m$ and let

$$\mathcal{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n\} \subseteq \mathcal{V}.$$

The following are equivalent:

1. *$\mathcal{B}$ is a basis for $\mathcal{V}$.*
2. *$\mathcal{B}$ is a minimal spanning set for $\mathcal{V}$.*
3. *$\mathcal{B}$ is a maximal linearly independent subset of $\mathcal{V}$.*

Remark

We say "a basis" here since $\mathcal{V}$ can have many different bases.

Proof

Since it is difficult to directly relate (2) and (3), our strategy will be to prove

- Show (1) $\Longrightarrow$ (2) and (2) $\Longrightarrow$ (1), so that (1) $\Longleftrightarrow$ (2).
- Show (1) $\Longrightarrow$ (3) and (3) $\Longrightarrow$ (1), so that (1) $\Longleftrightarrow$ (3).

Then — by transitivity — we will also have (2) $\Longleftrightarrow$ (3).

Proof (cont.)

(1) $\implies$ (2): Assume $\mathcal{B}$ is a basis (i.e., a linearly independent spanning set) of $\mathcal{V}$ and show that it is minimal.
Assume $\mathcal{B}$ is not minimal, i.e., we can find a smaller spanning set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ for $\mathcal{V}$ with $k \leq n$ elements.
But then

$$\boldsymbol{b}_j = \sum_{i=1}^{k} \alpha_{ij} \boldsymbol{x}_i, \quad j = 1, \ldots, n,$$

or

$$\mathsf{B} = \mathsf{X}\mathsf{A},$$

where

$$\begin{aligned}
\mathsf{B} &= \begin{pmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_n \end{pmatrix} \in \mathbb{R}^{m \times n}, \\
\mathsf{X} &= \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_k \end{pmatrix} \in \mathbb{R}^{m \times k}, \\
[\mathsf{A}]_{ij} &= \alpha_{ij}, \quad \mathsf{A} \in \mathbb{R}^{k \times n}.
\end{aligned}$$

Proof (cont.)

Now, $\operatorname{rank}(A) \le k < n$, which implies $N(A)$ is nontrivial, i.e., there exists a $\boldsymbol{z} \neq \mathbf{0}$ such that

$$A\boldsymbol{z} = \mathbf{0}.$$

But then

$$B\boldsymbol{z} = XA\boldsymbol{z} = \mathbf{0},$$

and therefore $N(B)$ is nontrivial.

However, since $\mathcal{B}$ is a basis, the columns of B are linearly independent (i.e., $N(B) = \{\mathbf{0}\}$) — and that is a contradiction.

Therefore, $\mathcal{B}$ has to be minimal.

Proof (cont.)

(2) $\Longrightarrow$ (1): Assume $\mathcal{B}$ is a minimal spanning set and show that it must also be linearly independent.

This is clear since

- if $\mathcal{B}$ were linearly dependent,
- then we would be able to remove at least one vector from $\mathcal{B}$ and still have a spanning set
- but then it would not have been minimal.

Proof (cont.)

(3) $\Longrightarrow$ (1): Assume $\mathcal{B}$ is a maximal linearly independent subset of $\mathcal{V}$ and show that $\mathcal{B}$ is a basis of $\mathcal{V}$.

Assume that $\mathcal{B}$ is not a basis, i.e., there exists a $\boldsymbol{v} \in \mathcal{V}$ such that $\boldsymbol{v} \notin \operatorname{span}\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$.

Then — by an earlier theorem — the extension set $\mathcal{B} \cup \{\boldsymbol{v}\}$ is linearly independent.

But this contradicts the maximality of $\mathcal{B}$, so that $\mathcal{B}$ has to be a basis.

Proof (cont.)

(1) $\Longrightarrow$ (3): Assume $\mathcal{B}$ is a basis, but not a maximal linearly independent subset of $\mathcal{V}$, and show that this leads to a contradiction.

Let

$$\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\} \subseteq \mathcal{V}, \quad \text{with } k > n$$

be a maximal linearly independent subset of $\mathcal{V}$ (note that such a set always exists).

But then $\mathcal{Y}$ must be a basis for $\mathcal{V}$ by our "(1) $\Longrightarrow$ (3)" argument.

On the other hand, $\mathcal{Y}$ has more vectors than $\mathcal{B}$ and a basis has to be a minimal spanning set.

Therefore $\mathcal{B}$ has to already be a maximal linearly independent subset of $\mathcal{V}$. $\square$

Remark

Above we remarked that $\mathcal{B}$ is not unique, i.e., a vector space $\mathcal{V}$ can have many different bases.

However, the number of elements in all of these bases is unique.

Definition

The dimension of the vector space $\mathcal{V}$ is given by

$$\dim \mathcal{V} = \text{the number of elements in any basis of } \mathcal{V}.$$

Special case: by convention

$$\dim\{\mathbf{0}\} = 0.$$

Example

Consider

$$\mathcal{P} = \operatorname{span}\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\} \subset \mathbb{R}^3.$$

Geometrically, $\mathcal{P}$ corresponds to the plane $z = 0$, i.e., the xy-plane.

Note that $\dim \mathcal{P} = 2$.

Moreover, any subspace of $\mathbb{R}^3$ has dimension at most 3.

In general,

Theorem

Let $\mathcal{M}$ and $\mathcal{N}$ be vector spaces such that $\mathcal{M} \subseteq \mathcal{N}$. Then

1. $\dim \mathcal{M} \leq \dim \mathcal{N}$,
2. $\dim \mathcal{M} = \dim \mathcal{N} \quad \Longrightarrow \quad \mathcal{M} = \mathcal{N}$.

Proof.

See [Mey00]. □

Back to the 4 fundamental subspaces

Consider an $m \times n$ matrix A with $\text{rank}(A) = r$.

$R(A)$ We know that

$$R(A) = \text{span}\{\text{columns of } A\}.$$

If $\text{rank}(A) = r$, then only r columns of A are linearly independent, i.e.,

$$\dim R(A) = r.$$

A basis of $R(A)$ is given by the basic columns of A (determined via a row echelon form U).

$R(A^T)$ We know that

$$R(A^T) = \text{span}\{\text{rows of A}\}.$$

Again, $\text{rank}(A) = r$ implies that only r rows of A are linearly independent, i.e.,

$$\dim R(A^T) = r.$$

A basis of $R(A^T)$ is given by the nonzero rows of U (from the LU factorization of A).

$N(A^T)$ One of our earlier theorems states that the last $m - r$ rows of P span $N(A^T)$ (where P is nonsingular such that $PA = U$ is in row echelon form).

Since P is nonsingular these rows are linearly independent and so

$$\dim N(A^T) = m - r.$$

A basis of $N(A^T)$ is given by the last $m - r$ rows of P.

$N(\mathsf{A})$ Replace A by A^T above so that

$$\dim N\left((\mathsf{A}^T)^T\right) = n - \operatorname{rank}(\mathsf{A}^T) = n - r$$

so that

$$\dim N(\mathsf{A}) = n - r.$$

A basis of $N(\mathsf{A})$ is given by the $n - r$ linearly independent solutions of $\mathsf{A}\boldsymbol{x} = \mathbf{0}$.

Theorem

For any $m \times n$ matrix A *we have*

$$\dim R(\mathsf{A}) + \dim N(\mathsf{A}) = n.$$

This follows directly from the above discussion of $R(\mathsf{A})$ and $N(\mathsf{A})$.

The theorem shows that there is always a balance between the rank of A and the dimension of its nullspace.

Example

Find the dimension and a basis for

$$\mathcal{S} = \text{span}\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 \\ 6 \\ 9 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 6 \\ 3 \end{pmatrix} \right\}.$$

Before we even do any calculations we know that

$$\mathcal{S} \subseteq \mathbb{R}^4, \quad \text{so that } \dim \mathcal{S} \leq 4.$$

We will now answer this question in two different ways using

$$\mathsf{A} = \begin{pmatrix} 1 & 2 & 2 & 3 & 1 \\ 2 & 4 & 4 & 6 & 2 \\ 3 & 6 & 6 & 9 & 6 \\ 1 & 2 & 4 & 5 & 3 \end{pmatrix}.$$

Example (cont.)

Via $R(A)$, i.e., by finding the basic columns of A:

$$A = \begin{pmatrix} 1 & 2 & 2 & 3 & 1 \\ 2 & 4 & 4 & 6 & 2 \\ 3 & 6 & 6 & 9 & 6 \\ 1 & 2 & 4 & 5 & 3 \end{pmatrix} \xrightarrow{\text{G.–J.}} E_A = \begin{pmatrix} 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Therefore, $\dim \mathcal{S} = 3$ and

$$\mathcal{S} = \operatorname{span}\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 6 \\ 3 \end{pmatrix} \right\}$$

since the basic columns of E_A are the first, third and fifth columns.

Example (cont.)

Via $R(A^T)$, i.e., $R(A) = \text{span}\{\text{rows of } A^T\}$, i.e., we need the nonzero rows of U (from the LU factorization of A^T:

$$A^T = \begin{pmatrix} 1 & 2 & 3 & 1 \\ 2 & 4 & 6 & 2 \\ 2 & 4 & 6 & 4 \\ 3 & 6 & 9 & 4 \\ 1 & 2 & 6 & 3 \end{pmatrix} \xrightarrow{\text{zero out } [A^T]_{*,1}} \begin{pmatrix} 1 & 2 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 2 \end{pmatrix} \xrightarrow{\text{permute}} \underbrace{\begin{pmatrix} 1 & 2 & 3 & 1 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{=U}$$

Therefore, $\dim \mathcal{S} = 3$ and

$$\mathcal{S} = \text{span}\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \end{pmatrix} \right\}$$

since the nonzero rows of U are the first, second and third rows.

Example

Extend

$$\mathcal{S} = \text{span}\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 6 \\ 3 \end{pmatrix} \right\}$$

to a basis for $\mathbb{R}^4$.
The procedure will be to augment the columns of $\mathcal{S}$ by an identity matrix , i.e., to form

$$\mathsf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 2 & 0 & 1 & 0 & 0 \\ 3 & 6 & 0 & 0 & 1 & 0 \\ 1 & 3 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and then to get a basis via the basic columns of U.

Example (cont.)

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 2 & 0 & 1 & 0 & 0 \\ 3 & 6 & 0 & 0 & 1 & 0 \\ 1 & 3 & 0 & 0 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 3 & -3 & 0 & 1 & 0 \\ 0 & 2 & -1 & 0 & 0 & 1 \end{pmatrix}$$

$$\longrightarrow \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & 0 & 1 \\ 0 & 0 & -\frac{3}{2} & 0 & 1 & -\frac{3}{2} \\ 0 & 0 & -2 & 1 & 0 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & 0 & 1 \\ 0 & 0 & -\frac{3}{2} & 0 & 1 & -\frac{3}{2} \\ 0 & 0 & 0 & 1 & -\frac{4}{3} & 2 \end{pmatrix}$$

so that the basic columns are $[A]_{*1}, [A]_{*2}, [A]_{*3}, [A]_{*4}$ and

$$\mathbb{R}^4 = \text{span}\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 6 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \right\}.$$

Earlier we defined the sum of subspaces $\mathcal{X}$ and $\mathcal{Y}$ as

$$\mathcal{X} + \mathcal{Y} = \{\boldsymbol{x} + \boldsymbol{y} : \ \boldsymbol{x} \in \mathcal{X}, \ \boldsymbol{y} \in \mathcal{Y}\}$$

Theorem

If $\mathcal{X}, \mathcal{Y}$ are subspaces of $\mathcal{V}$, then

$$\dim(\mathcal{X} + \mathcal{Y}) = \dim \mathcal{X} + \dim \mathcal{Y} - \dim(\mathcal{X} \cap \mathcal{Y}).$$

Proof.

See [Mey00], but the basic idea is pretty clear.
We want to avoid double counting. □

Corollary

Let A *and* B *be* $m \times n$ *matrices. Then*

$$\operatorname{rank}(\mathsf{A}+\mathsf{B}) \le \operatorname{rank}(\mathsf{A}) + \operatorname{rank}(\mathsf{B}).$$

Proof

First we note that

$$R(\mathsf{A}+\mathsf{B}) \subseteq R(\mathsf{A}) + R(\mathsf{B}) \tag{4}$$

since for any $\boldsymbol{b} \in R(\mathsf{A}+\mathsf{B})$ we have

$$\boldsymbol{b} = (\mathsf{A}+\mathsf{B})\boldsymbol{x} = \mathsf{A}\boldsymbol{x} + \mathsf{B}\boldsymbol{x} \in R(\mathsf{A}) + R(\mathsf{B}).$$

(cont.)

Now,

$$
\begin{aligned}
\operatorname{rank}(A+B) &= \dim R(A+B) \\
&\overset{(4)}{\leq} \dim(R(A)+R(B)) \\
&\overset{\text{Thm}}{=} \dim R(A) + \dim R(B) - \dim\left(R(A) \cap R(B)\right) \\
&\leq \dim R(A) + \dim R(B) \\
&= \mathit{rank}(A) + \mathit{rank}(B)
\end{aligned}
$$

$\square$

More About Rank

We know that $\mathsf{A} \sim \mathsf{B}$ if and only if $\operatorname{rank}(\mathsf{A}) = \operatorname{rank}(\mathsf{B})$.

Thus (for invertible P, Q), $\mathsf{PAQ} = \mathsf{B}$ implies $\operatorname{rank}(\mathsf{A}) = \operatorname{rank}(\mathsf{PAQ})$.

As we now show, it is a general fact that multiplication by a nonsingular matrix does not change the rank of a given matrix.

Moreover, multiplication by an arbitrary matrix can only lower the rank.

Theorem

Let A be an $m \times n$ matrix, and let B by $n \times p$. Then

$$\operatorname{rank}(\mathsf{AB}) = \operatorname{rank}(\mathsf{B}) - \dim\left(N(\mathsf{A}) \cap R(\mathsf{B})\right).$$

Remark

Note that if A is nonsingular, then $N(\mathsf{A}) = \{\mathbf{0}\}$ so that $\dim\left(N(\mathsf{A}) \cap R(\mathsf{B})\right) = 0$ and $\operatorname{rank}(\mathsf{AB}) = \operatorname{rank}(\mathsf{B})$.

Proof

Let $\mathcal{S} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_s\}$ be a basis for $N(\mathsf{A}) \cap R(\mathsf{B})$.

Since $N(\mathsf{A}) \cap R(\mathsf{B}) \subseteq R(\mathsf{B})$ we know that

$$\dim(R(\mathsf{B})) = s + t, \quad \text{for some } t \geq 0.$$

We can construct an extension set such that

$$\mathcal{B} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_s, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_t\}$$

is a basis for $R(\mathsf{B})$.

(cont.)

If we can show that $\dim(R(AB)) = t$ then

$$\operatorname{rank}(B) = \dim(R(B)) = s + t = \dim\left(N(A) \cap R(B)\right) + \dim(R(AB)),$$

and we are done.

Therefore, we now show that $\dim(R(AB)) = t$.
In particular, we show that

$$\mathcal{T} = \{A\boldsymbol{z}_1, A\boldsymbol{z}_2, \ldots, A\boldsymbol{z}_t\}$$

is a basis for $R(AB)$.

We do this by showing that

1. $\mathcal{T}$ is a spanning set for $R(AB)$,
2. $\mathcal{T}$ is linearly independent.

(cont.)

Spanning set: Consider an arbitrary $\boldsymbol{b} \in R(\mathsf{AB})$. It can be written as

$$\boldsymbol{b} = \mathsf{AB}\boldsymbol{y} \quad \text{for some } \boldsymbol{y}.$$

But then $\mathsf{B}\boldsymbol{y} \in R(\mathsf{B})$, so that

$$\mathsf{B}\boldsymbol{y} = \sum_{i=1}^{s} \xi_i \boldsymbol{x}_i + \sum_{j=1}^{t} \eta_j \boldsymbol{z}_j$$

and

$$\boldsymbol{b} = \mathsf{AB}\boldsymbol{y} = \sum_{i=1}^{s} \xi_i \mathsf{A}\boldsymbol{x}_i + \sum_{j=1}^{t} \eta_j \mathsf{A}\boldsymbol{z}_j = \sum_{j=1}^{t} \eta_j \mathsf{A}\boldsymbol{z}_j$$

since $\boldsymbol{x}_i \in N(\mathsf{A})$.

(cont.)

Linear independence: Let's use the definition of linear independence and look at

$$\sum_{i=1}^{t} \alpha_i \mathsf{A} \boldsymbol{z}_i = \mathbf{0} \quad \Longleftrightarrow \quad \mathsf{A} \sum_{i=1}^{t} \alpha_i \boldsymbol{z}_i = \mathbf{0}.$$

The identity on the right implies that $\sum_{i=1}^{t} \alpha_i \boldsymbol{z}_i \in N(\mathsf{A})$.

But we also have $\boldsymbol{z}_i \in \mathcal{B}$, i.e., $\sum_{i=1}^{t} \alpha_i \boldsymbol{z}_i \in R(\mathsf{B})$.

And so together

$$\sum_{i=1}^{t} \alpha_i \boldsymbol{z}_i \in N(\mathsf{A}) \cap R(\mathsf{B}).$$

(cont.)

Now, since $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\}$ is a basis for $N(\mathsf{A}) \cap R(\mathsf{B})$ we have

$$\sum_{i=1}^{t} \alpha_i \boldsymbol{z}_i = \sum_{j=1}^{s} \beta_j \boldsymbol{x}_j \quad \Longleftrightarrow \quad \sum_{i=1}^{t} \alpha_i \boldsymbol{z}_i - \sum_{j=1}^{s} \beta_j \boldsymbol{x}_j = \boldsymbol{0}.$$

But $\mathcal{B} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t\}$ is linearly independent, so that $\alpha_1 = \cdots = \alpha_t = \beta_1 = \cdots = \beta_s = 0$ and therefore $\mathcal{T}$ is also linearly independent. $\square$

It turns out that $\dim(N(\mathsf{A}) \cap R(\mathsf{B}))$ is relatively difficult to determine.

Therefore, the following upper and lower bounds for $\operatorname{rank}(\mathsf{AB})$ are useful.

Theorem

Let A *be an* $m \times n$ *matrix, and let* B *by* $n \times p$. *Then*

1. $\operatorname{rank}(\mathsf{AB}) \leq \min\{\operatorname{rank}(\mathsf{A}), \operatorname{rank}(\mathsf{B})\}$,
2. $\operatorname{rank}(\mathsf{AB}) \geq \operatorname{rank}(\mathsf{A}) + \operatorname{rank}(\mathsf{B}) - n$.

Proof of (1)

We show that $\operatorname{rank}(AB) \leq \operatorname{rank}(A)$ and $\operatorname{rank}(AB) \leq \operatorname{rank}(B)$.

The previous theorem states

$$\operatorname{rank}(AB) = \operatorname{rank}(B) - \underbrace{\dim(N(A) \cap R(B))}_{\geq 0} \leq \operatorname{rank}(B).$$

Similarly,

$$\operatorname{rank}(AB) = \operatorname{rank}(AB)^T = \operatorname{rank}(B^T A^T) \overset{\text{as above}}{\leq} \operatorname{rank}(A^T) = \operatorname{rank}(A).$$

To make things as tight as possible we take the smaller of the two upper bounds.

Proof of (2)

We begin by noting that $N(\mathsf{A}) \cap R(\mathsf{B}) \subseteq N(\mathsf{A})$.

Therefore,

$$\dim(N(\mathsf{A}) \cap R(\mathsf{B})) \leq \dim(N(\mathsf{A})) = n - \operatorname{rank}(\mathsf{A}).$$

But then (using the previous theorem)

$$\begin{aligned}\operatorname{rank}(\mathsf{AB}) &= \operatorname{rank}(\mathsf{B}) - \dim(N(\mathsf{A}) \cap R(\mathsf{B})) \\ &\geq \operatorname{rank}(\mathsf{B}) - n + \operatorname{rank}(\mathsf{A}).\end{aligned}$$

$\square$

To prepare for our study of least squares solutions, where the matrices $\mathsf{A}^T\mathsf{A}$ and $\mathsf{A}\mathsf{A}^T$ are important, we prove

Lemma

Let A *be a real* $m \times n$ *matrix. Then*

1. $\operatorname{rank}(\mathsf{A}^T\mathsf{A}) = \operatorname{rank}(\mathsf{A}\mathsf{A}^T) = \operatorname{rank}(\mathsf{A})$.
2. $R(\mathsf{A}^T\mathsf{A}) = R(\mathsf{A}^T), \quad R(\mathsf{A}\mathsf{A}^T) = R(\mathsf{A})$.
3. $N(\mathsf{A}^T\mathsf{A}) = N(\mathsf{A}), \quad N(\mathsf{A}\mathsf{A}^T) = N(\mathsf{A}^T)$.

Proof

From our earlier theorem we know

$$\operatorname{rank}(\mathsf{A}^T\mathsf{A}) = \operatorname{rank}(\mathsf{A}) - \dim(N(\mathsf{A}^T) \cap R(\mathsf{A})).$$

For (1) to be true we need to show $\dim(N(\mathsf{A}^T) \cap R(\mathsf{A})) = 0$, i.e., $N(\mathsf{A}^T) \cap R(\mathsf{A}) = \{\mathbf{0}\}$.
This is true since

$$\boldsymbol{x} \in N(\mathsf{A}^T) \cap R(\mathsf{A}) \quad \Longrightarrow \quad \mathsf{A}^T\boldsymbol{x} = \mathbf{0} \text{ and } \boldsymbol{x} = \mathsf{A}\boldsymbol{y} \text{ for some } \boldsymbol{y}.$$

Therefore (using $\boldsymbol{x}^T = \boldsymbol{y}^T\mathsf{A}^T$)

$$\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\mathsf{A}^T\boldsymbol{x} = 0.$$

But

$$\boldsymbol{x}^T\boldsymbol{x} = 0 \quad \Longleftrightarrow \quad \sum_{i=1}^{m} x_i^2 = 0 \quad \Longrightarrow \quad \boldsymbol{x} = \mathbf{0}.$$

(cont.)

$\text{rank}(AA^T) = \text{rank}(A^T)$ obtained by switching A and A^T, and then use $\text{rank}(A^T) = \text{rank}(A)$.

The first part of (2) follows from $R(A^TA) \subseteq R(A^T)$ (see HW) and

$$\dim(R(A^TA)) = \text{rank}(A^TA) \stackrel{(1)}{=} \text{rank}(A^T) = \dim(R(A^T))$$

since for $\mathcal{M} \subseteq \mathcal{N}$ with $\dim \mathcal{M} = \dim \mathcal{N}$ one has $\mathcal{M} = \mathcal{N}$ (from an earlier theorem).

The other part of (2) follows by switching A and A^T.

(cont.)

The first part of (3) follows from $N(\mathsf{A}) \subseteq N(\mathsf{A}^T\mathsf{A})$ (see HW) and

$$\dim(N(\mathsf{A})) = n - \operatorname{rank}(\mathsf{A}) = n - \operatorname{rank}(\mathsf{A}^T\mathsf{A}) = \dim(N(\mathsf{A}^T\mathsf{A}))$$

using the same reasoning as above.

The other part of (3) follows by switching A and A^T. $\square$

Connection to least squares and normal equations

Consider a — possibly inconsistent — linear system

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$$

with $m \times n$ matrix A (and $\boldsymbol{b} \notin R(\mathsf{A})$ if inconsistent).

To find a "solution" we multiply both sides by A^T to get the normal equations:

$$\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b},$$

where $\mathsf{A}^T\mathsf{A}$ is an $n \times n$ matrix.

Theorem

Let A *be an* $m \times n$ *matrix,* $\boldsymbol{b}$ *an* m*-vector, and consider the normal equations*

$$\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}$$

associated with $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$*.*

1. *The normal equations are always consistent, i.e., for every* A *and* $\boldsymbol{b}$ *there exists at least one* $\boldsymbol{x}$ *such that* $\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}$*.*
2. *If* $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ *is consistent, then* $\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}$ *has the same solution set (the least squares solution of* $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$*).*
3. $\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}$ *has a unique solution if and only if* $\operatorname{rank}(\mathsf{A}) = n$*. Then*
$$\boldsymbol{x} = (\mathsf{A}^T\mathsf{A})^{-1}\mathsf{A}^T\boldsymbol{b},$$
regardless of whether $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ *is consistent or not.*
4. *If* $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ *is consistent and has a unique solution, then the same holds for* $\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}$ *and* $\boldsymbol{x} = (\mathsf{A}^T\mathsf{A})^{-1}\mathsf{A}^T\boldsymbol{b}$*.*

Proof

(1) follows from our previous lemma, i.e.,

$$\mathsf{A}^T \boldsymbol{b} \in R(\mathsf{A}^T) = R(\mathsf{A}^T\mathsf{A}).$$

To show (2) we assume the $\boldsymbol{p}$ is some particular solution of $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$, i.e., $\mathsf{A}\boldsymbol{p} = \boldsymbol{b}$.

If we multiply by A^T, then

$$\mathsf{A}^T\mathsf{A}\boldsymbol{p} = \mathsf{A}^T\boldsymbol{p},$$

so that $\boldsymbol{p}$ is also a solution of the normal equations.

(cont.)

Now, the general solution of $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ is from the set (see Problem 2 on HW#4)

$$\mathcal{S} = \boldsymbol{p} + N(\mathsf{A}).$$

Moreover, the general solution of $\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}$ is of the form

$$\boldsymbol{p} + N(\mathsf{A}^T\mathsf{A}) \stackrel{\text{lemma}}{=} \boldsymbol{p} + N(\mathsf{A}) = \mathcal{S}.$$

(cont.)

For (3) we want to show that $A^T A\boldsymbol{x} = A^T\boldsymbol{b}$ has a unique solution if and only if $\operatorname{rank}(A) = n$.

What we know immediately is that $A^T A\boldsymbol{x} = A^T\boldsymbol{b}$ has a unique solution if and only if $\operatorname{rank}(A^T A) = n$.
Since we showed earlier that $\operatorname{rank}(A^T A) = \operatorname{rank}(A)$ this part is done.

Now, if $\operatorname{rank}(A^T A) = n$ we know that $A^T A$ is invertible (even though A^T and A may not be) and therefore

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b} \quad \Longleftrightarrow \quad \boldsymbol{x} = (A^T A)^{-1} A^T\boldsymbol{b}.$$

To show (4) we note that $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution if and only if $\operatorname{rank}(A) = n$. But $\operatorname{rank}(A^T A) = \operatorname{rank}(A)$ and the rest follows from (3). $\square$

Remark

The normal equations are not recommended for serious computations since they are often rather ill-conditioned since one can show that

$$cond(\mathsf{A}^T\mathsf{A}) = cond(\mathsf{A})^2.$$

There's an example in [Mey00] that illustrates this fact.

Historical definition of rank

Let A be an $m \times n$ matrix. Then A has rank r if there exists at least one nonsingular $r \times r$ submatrix of A (and none larger).

Example

The matrix

$$\mathsf{A} = \begin{pmatrix} 1 & 2 & 2 & 3 & 1 \\ 2 & 4 & 4 & 6 & 2 \\ 3 & 6 & 6 & 9 & 6 \\ 1 & 2 & 4 & 5 & 3 \end{pmatrix}$$

cannot have rank 4 since rows one and two are linearly dependent.

But $\operatorname{rank}(\mathsf{A}) \geq 2$ since $\begin{pmatrix} 9 & 6 \\ 5 & 3 \end{pmatrix}$ is nonsingular.

Example (cont.)

In fact, $\operatorname{rank}(A) = 3$ since

$$\begin{pmatrix} 4 & 6 & 2 \\ 6 & 9 & 6 \\ 4 & 5 & 3 \end{pmatrix}$$

is nonsingular.

Note that other singular 3×3 submatrices are allowed, such as

$$\begin{pmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 3 & 6 & 6 \end{pmatrix}.$$

Earlier we showed that

$$\text{rank}(AB) \leq \text{rank}(A),$$

i.e., multiplication by another matrix does not increase the rank of a given matrix, i.e., we can't "fix" a singular system by multiplication.

Now

Theorem

Let A *and* E *be* $m \times n$ *matrices. Then*

$$\text{rank}(A + E) \geq \text{rank}(A),$$

provided the entries of E *are "sufficiently small".*

This theorem has at least two fundamental consequences of practical importance:

- Beware!! A theoretically singular system may become nonsingular, i.e., have a “solution” — just due to round-off error.
- We may want to intentionally “fix” a singular system, so that it has a “solution”. One such strategy is known as Tikhonov regularization, i.e.,

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b} \;\longrightarrow\; (\mathsf{A} + \mu \mathsf{I})\boldsymbol{x} = \boldsymbol{b},$$

  where μ is a (small) regularization parameter.

Proof

We assume that $\operatorname{rank}(A) = r$ and that we have nonsingular P and Q such that we can convert A to rank normal form, i.e.,

$$PAQ = \begin{pmatrix} I_r & O \\ O & O \end{pmatrix}.$$

Then — formally — $PEQ = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}$ with appropriate blocks E_{ij}.

This allows us to write

$$P(A+E)Q = \begin{pmatrix} I_r + E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}.$$

(cont.)

Now, we note that

$$(\mathsf{I}-\mathsf{B})(\mathsf{I}+\mathsf{B}+\mathsf{B}^2+\ldots+\mathsf{B}^{k-1}) = \mathsf{I}-\mathsf{B}^k \\ \to \mathsf{I},$$

provided the entries of B are "sufficiently small" (i.e., so that $\mathsf{B}^k \to \mathsf{O}$ for $k \to \infty$).
Therefore $(\mathsf{I}-\mathsf{B})^{-1}$ exists.

This technique is known as the Neumann series expansion of the inverse of $\mathsf{I}-\mathsf{B}$.

(cont.)

Now, letting $B = -E_{11}$, we know that $(I_r + E_{11})^{-1}$ exists and we can write

$$\begin{pmatrix} I_r & O \\ -E_{21}(I_r + E_{11})^{-1} & I_{m-r} \end{pmatrix} \begin{pmatrix} I_r + E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix} \begin{pmatrix} I_r & -(I_r + E_{11})^{-1}E_{12} \\ O & I_{n-r} \end{pmatrix}$$
$$= \begin{pmatrix} I_r + E_{11} & O \\ O & S \end{pmatrix},$$

where $S = E_{22} - E_{21}(I_r + E_{11})^{-1}E_{12}$ is the Schur complement of $I + E_{11}$ in PAQ.

(cont.)

The Schur complement calculation shows that

$$A + E \sim \begin{pmatrix} I_r + E_{11} & O \\ O & S \end{pmatrix}.$$

But then this rank normal form with invertible diagonal blocks tells us

$$\begin{aligned} \operatorname{rank}(A + E) &= \operatorname{rank}(I_r + E_{11}) + \operatorname{rank}(S) \\ &= \operatorname{rank}(A) + \operatorname{rank}(S) \\ &\geq \operatorname{rank}(A). \end{aligned}$$

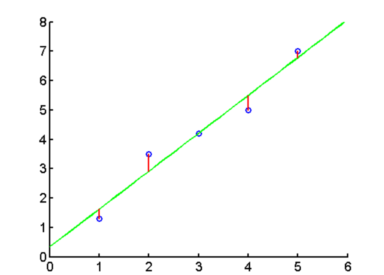
□

Linear least squares (linear regression)

Given: data $\{(t_1, b_1), (t_2, b_2), \ldots, (t_m, b_m)\}$

Find: "best fit" by a line

t	1	2	3	4	5
b	1.3	3.5	4.2	5.0	7.0



Idea for best fit

Minimize the sum of the squares of the vertical distances of line from the data points.

More precisely, let

$$f(t) = \alpha + \beta t$$

with α, β such that

$$\begin{aligned}\sum_{i=1}^{m} \varepsilon_i^2 &= \sum_{i=1}^{m} (f(t_i) - b_i)^2 \\ &= \sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 = G(\alpha, \beta) \quad \longrightarrow \quad \text{min}\end{aligned}$$

From calculus, necessary (and sufficient) condition for minimum

$$\frac{\partial G(\alpha,\beta)}{\partial \alpha} = 0, \qquad \frac{\partial G(\alpha,\beta)}{\partial \beta} = 0.$$

where

$$\frac{\partial G(\alpha,\beta)}{\partial \alpha} = 2\sum_{i=1}^{m} (\alpha + \beta t_i - b_i)\,, \qquad \frac{\partial G(\alpha,\beta)}{\partial \beta} = 2\sum_{i=1}^{m} (\alpha + \beta t_i - b_i)\, t_i$$

Equivalently,

$$\left(\sum_{i=1}^m 1\right)\alpha + \left(\sum_{i=1}^m t_i\right)\beta = \sum_{i=1}^m b_i$$
$$\left(\sum_{i=1}^m t_i\right)\alpha + \left(\sum_{i=1}^m t_i^2\right)\beta = \sum_{i=1}^m b_i t_i$$

which can be written as

$$\mathsf{Q}\boldsymbol{x} = \boldsymbol{y}$$

with

$$\mathsf{Q} = \begin{pmatrix} \sum_{i=1}^m 1 & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} \sum_{i=1}^m b_i \\ \sum_{i=1}^m b_i t_i \end{pmatrix}$$

We can write each of these sums as inner products:

$$\sum_{i=1}^{m} 1 = \mathbf{1}^T \mathbf{1}, \quad \sum_{i=1}^{m} t_i = \mathbf{1}^T \boldsymbol{t} = \boldsymbol{t}^T \mathbf{1}, \quad \sum_{i=1}^{m} t_i^2 = \boldsymbol{t}^T \boldsymbol{t}$$

$$\sum_{i=1}^{m} b_i = \mathbf{1}^T \boldsymbol{b} = \boldsymbol{b}^T \mathbf{1}, \quad \sum_{i=1}^{m} b_i t_i = \boldsymbol{b}^T \boldsymbol{t} = \boldsymbol{t}^T \boldsymbol{b},$$

where

$$\mathbf{1}^T = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}, \quad \boldsymbol{t}^T = \begin{pmatrix} t_1 & \cdots & t_m \end{pmatrix}, \quad \boldsymbol{b}^T = \begin{pmatrix} b_1 & \cdots & b_m \end{pmatrix}$$

With this notation we have

$$\begin{aligned} \mathsf{Q}\boldsymbol{x} = \boldsymbol{y} \quad &\Longleftrightarrow \quad \begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\boldsymbol{t} \\ \boldsymbol{t}^T\mathbf{1} & \boldsymbol{t}^T\boldsymbol{t} \end{pmatrix} \boldsymbol{x} = \begin{pmatrix} \mathbf{1}^T\boldsymbol{b} \\ \boldsymbol{t}^T\boldsymbol{b} \end{pmatrix} \\ &\Longleftrightarrow \quad \mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}, \qquad \mathsf{A}^T = \begin{pmatrix} \mathbf{1}^T \\ \boldsymbol{t}^T \end{pmatrix}, \quad \mathsf{A} = \begin{pmatrix} \mathbf{1} & \boldsymbol{t} \end{pmatrix} \end{aligned}$$

Therefore we can find the parameters of the line, $\boldsymbol{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, by solving the square linear system

$$\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}.$$

Also note that since $\varepsilon_i = \alpha + \beta t_i - b_i$ we have

$$\begin{aligned} \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix} &= \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \alpha + \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix} \beta - \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \\ &= \mathbf{1}\alpha + \boldsymbol{t}\beta - \boldsymbol{b} = \mathsf{A}\boldsymbol{x} - \boldsymbol{b}. \end{aligned}$$

This implies that

$$G(\alpha, \beta) = \sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathsf{A}\boldsymbol{x} - \boldsymbol{b})^T(\mathsf{A}\boldsymbol{x} - \boldsymbol{b}).$$

Example

Data:

t	-1	0	1	2	3	4	5	6
b	10	9	7	5	4	3	0	-1

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b} \iff \begin{pmatrix} \sum_{i=1}^8 1 & \sum_{i=1}^8 t_i \\ \sum_{i=1}^8 t_i & \sum_{i=1}^8 t_i^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^8 b_i \\ \sum_{i=1}^8 b_i t_i \end{pmatrix}$$

$$\iff \begin{pmatrix} 8 & 20 \\ 20 & 92 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 37 \\ 25 \end{pmatrix}$$

$$\implies \alpha \approx 8.643, \ \beta \approx -1.607$$

So that the best fit line to the given data is

$$f(t) \approx 8.643 - 1.607t.$$

General Least Squares

The general least squares problem behaves analogously to the linear example.

Theorem

Let A *be a real* $m \times n$ *matrix and* $\boldsymbol{b}$ *an m-vector. Any vector* $\boldsymbol{x}$ *that minimizes the square of the residual* $\mathsf{A}\boldsymbol{x} - \boldsymbol{b}$*, i.e.,*

$$G(\boldsymbol{x}) = (\mathsf{A}\boldsymbol{x} - \boldsymbol{b})^T(\mathsf{A}\boldsymbol{x} - \boldsymbol{b})$$

is called a least squares solution of $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$*.*
The set of all least squares solutions is obtained by solving the normal equations

$$\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}.$$

Moreover, a unique solution exists if and only if $\operatorname{rank}(\mathsf{A}) = n$ *so that*

$$\boldsymbol{x} = (\mathsf{A}^T\mathsf{A})^{-1}\mathsf{A}^T\boldsymbol{b}.$$

Proof

The statement about uniqueness follows directly from our earlier theorem on p. 92 on the normal equations.

To characterize the least squares solutions we first show that if $\boldsymbol{x}$ minimizes $G(\boldsymbol{x})$ then $\boldsymbol{x}$ satisfies $\mathsf{A}^T\mathsf{A}\boldsymbol{x} = \mathsf{A}^T\boldsymbol{b}$.

As in our earlier example, a necessary condition for the minimum is: $\frac{\partial G(\boldsymbol{x})}{\partial x_i} = 0$, $i = 1, \ldots, n$.

Let's first work out what $G(\boldsymbol{x})$ looks like:

$$\begin{aligned} G(\boldsymbol{x}) &= (\mathsf{A}\boldsymbol{x} - \boldsymbol{b})^T(\mathsf{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= \boldsymbol{x}^T\mathsf{A}^T\mathsf{A}\boldsymbol{x} - \boldsymbol{x}^T\mathsf{A}^T\boldsymbol{b} - \boldsymbol{b}^T\mathsf{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{b} \\ &= \boldsymbol{x}^T\mathsf{A}^T\mathsf{A}\boldsymbol{x} - 2\boldsymbol{x}^T\mathsf{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} \end{aligned}$$

since $\boldsymbol{b}^T\mathsf{A}\boldsymbol{x} = (\boldsymbol{b}^T\mathsf{A}\boldsymbol{x})^T = \boldsymbol{x}^T\mathsf{A}^T\boldsymbol{b}$ is a scalar.

(cont.)

Therefore

$$\frac{\partial G(\boldsymbol{x})}{\partial x_i} = \frac{\partial \boldsymbol{x}^T}{\partial x_i} \mathsf{A}^T \mathsf{A} \boldsymbol{x} + \boldsymbol{x}^T \mathsf{A}^T \mathsf{A} \frac{\partial \boldsymbol{x}}{\partial x_i} - 2 \frac{\partial \boldsymbol{x}^T}{\partial x_i} \mathsf{A}^T \boldsymbol{b}$$
$$= \boldsymbol{e}_i^T \mathsf{A}^T \mathsf{A} \boldsymbol{x} + \boldsymbol{x}^T \mathsf{A}^T \mathsf{A} \boldsymbol{e}_i - 2 \boldsymbol{e}_i^T \mathsf{A}^T \boldsymbol{b}$$
$$= 2 \boldsymbol{e}_i^T \mathsf{A}^T \mathsf{A} \boldsymbol{x} - 2 \boldsymbol{e}_i^T \mathsf{A}^T \boldsymbol{b}$$

since $\boldsymbol{x}^T \mathsf{A}^T \mathsf{A} \boldsymbol{e}_i = (\boldsymbol{x}^T \mathsf{A}^T \mathsf{A} \boldsymbol{e}_i)^T = \boldsymbol{e}_i^T \mathsf{A}^T \mathsf{A} \boldsymbol{x}$ is a scalar.

This means that

$$\frac{\partial G(\boldsymbol{x})}{\partial x_i} = 0 \quad \Longleftrightarrow \quad (\mathsf{A}^T)_{i*} \mathsf{A} \boldsymbol{x} = (\mathsf{A}^T)_{i*} \boldsymbol{b}.$$

If we collect all such conditions (for $i = 1, \ldots, n$) in one linear system we get

$$\mathsf{A}^T \mathsf{A} \boldsymbol{x} = \mathsf{A}^T \boldsymbol{b}.$$

(cont.)

To verify that we indeed have a minimum we show that if $\boldsymbol{z}$ is a solution of the normal equations then $G(\boldsymbol{z})$ is minimal.

$$\begin{aligned} G(\boldsymbol{z}) &= (\mathsf{A}\boldsymbol{z}-\boldsymbol{b})^T(\mathsf{A}\boldsymbol{z}-\boldsymbol{b}) \\ &= \boldsymbol{z}^T\mathsf{A}^T\mathsf{A}\boldsymbol{z} - 2\boldsymbol{z}^T\mathsf{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} \\ &= \boldsymbol{z}^T(\underbrace{\mathsf{A}^T\mathsf{A}\boldsymbol{z}-\mathsf{A}^T\boldsymbol{b}}_{=0}) - \boldsymbol{z}^T\mathsf{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} = -\boldsymbol{z}^T\mathsf{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b}. \end{aligned}$$

Now, for any other $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{u}$ we have

$$\begin{aligned} G(\boldsymbol{y}) &= (\boldsymbol{z}+\boldsymbol{u})^T\mathsf{A}^T\mathsf{A}(\boldsymbol{z}+\boldsymbol{u}) - 2(\boldsymbol{z}+\boldsymbol{u})^T\mathsf{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} \\ &= G(\boldsymbol{z}) + \boldsymbol{u}^T\mathsf{A}^T\mathsf{A}\boldsymbol{u} + \underbrace{\boldsymbol{z}^T\mathsf{A}^T\mathsf{A}\boldsymbol{u}}_{=\boldsymbol{u}^T\mathsf{A}^T\mathsf{A}\boldsymbol{z}} + \boldsymbol{u}^T\mathsf{A}^T\mathsf{A}\boldsymbol{z} - 2\boldsymbol{u}^T\underbrace{\mathsf{A}^T\boldsymbol{b}}_{\mathsf{A}^T\mathsf{A}\boldsymbol{z}} \\ &= G(\boldsymbol{z}) + \boldsymbol{u}^T\mathsf{A}^T\mathsf{A}\boldsymbol{u} \geq G(\boldsymbol{z}) \end{aligned}$$

since $\boldsymbol{u}^T\mathsf{A}^T\mathsf{A}\boldsymbol{u} = \sum_{i=1}^m (\mathsf{A}\boldsymbol{u})_i^2 \geq 0$. $\square$

Remark

Using this framework we can compute least squares fits from any linear function space.

Example

1. Let $f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$, i.e., we can use quadratic polynomials (or any other degree).
2. Let $f(t) = \alpha_0 + \alpha_1 \sin t + \alpha_2 \cos t$, i.e., we can use trigonometric polynomials.
3. Let $f(t) = \alpha e^t + \beta\sqrt{t}$, i.e., we can use just about anything we want.

Regression in Statistics (BLUE)

One assumes that there is a random process that generates data as a random variable Y of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n,$$

where $X_1, \ldots, X_n$ are (input) random variables and $\beta_1, \ldots, \beta_n$ are unknown parameters.
Now the actually observed data may be affected by noise, i.e.,

$$y = Y + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ (normally distributed with mean zero and variance σ^2) is another random variable denoting the noise.
To determine the model parameters $\beta_1, \ldots, \beta_n$ we now look at measurements, i.e.,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_n x_{i,n} + \varepsilon, \quad i = 1, \ldots, m.$$

In matrix-vector form this gives us

$$\boldsymbol{y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Now, the least squares solution of $\mathsf{X}\boldsymbol{\beta} = \boldsymbol{y}$, i.e., $\hat{\boldsymbol{\beta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y}$ is in fact the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$.
To show this one needs an assumption that the error is unbiased, i.e., $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$.
Then

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\mathsf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbb{E}[\mathsf{X}\boldsymbol{\beta}] + \mathbb{E}[\boldsymbol{\varepsilon}] = \mathsf{X}\boldsymbol{\beta}$$

and therefore

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y}] = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\mathbb{E}[\boldsymbol{y}] \\ &= (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\mathsf{X}\boldsymbol{\beta} = \boldsymbol{\beta},\end{aligned}$$

so that the estimator is indeed unbiased.

Remark

One can also show (maybe later) that $\hat{\beta}$ has minimal variance among all unbiased linear estimators, so it is the best linear unbiased estimator of the model parameters.

In fact, the theorem ensuring this is the so-called Gauss-Markov theorem.

Kriging: a regression approach

Assume: the approximate value of a realization of a zero-mean (Gaussian) random field is given by a linear predictor of the form

$$\hat{Y}_{\boldsymbol{x}} = \sum_{j=1}^{N} Y_{\boldsymbol{x}_j} w_j(\boldsymbol{x}) = \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y},$$

where $\hat{Y}_{\boldsymbol{x}}$ and $Y_{\boldsymbol{x}_j}$ are random variables, $\boldsymbol{Y} = \begin{pmatrix} Y_{\boldsymbol{x}_1} & \cdots & Y_{\boldsymbol{x}_N} \end{pmatrix}^T$, and $\boldsymbol{w}(\boldsymbol{x}) = \begin{pmatrix} w_1(\boldsymbol{x}) & \cdots & w_N(\boldsymbol{x}) \end{pmatrix}^T$ is a vector of weight functions at $\boldsymbol{x}$.

Since all of the $Y_{\boldsymbol{x}_j}$ have zero mean the predictor $\hat{Y}_{\boldsymbol{x}}$ is automatically unbiased.

Goal: to compute "optimal" weights $\overset{\star}{w}_j(\cdot)$, $j = 1, \ldots, N$. To this end, consider the mean-squared error (MSE) of the predictor, i.e.,

$$\mathrm{MSE}(\hat{Y}_{\boldsymbol{x}}) = \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\right)^2\right].$$

We now present some details (see [FM15]).

Covariance Kernel

We need the covariance kernel K of a random field Y with mean $\mu(\boldsymbol{x})$. It is defined via

$$\begin{aligned}
\sigma^2 K(\boldsymbol{x},\boldsymbol{z}) &= \operatorname{Cov}(Y_{\boldsymbol{x}}, Y_{\boldsymbol{z}}) = \mathbb{E}\left[(Y_{\boldsymbol{x}} - \mu(\boldsymbol{x}))(Y_{\boldsymbol{z}} - \mu(\boldsymbol{z}))\right] \\
&= \mathbb{E}\left[(Y_{\boldsymbol{x}} - \mathbb{E}[Y_{\boldsymbol{x}}])(Y_{\boldsymbol{z}} - \mathbb{E}[Y_{\boldsymbol{z}}])\right] \\
&= \mathbb{E}\left[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}} - Y_{\boldsymbol{x}}\mathbb{E}[Y_{\boldsymbol{z}}] - \mathbb{E}[Y_{\boldsymbol{x}}]Y_{\boldsymbol{z}} + \mathbb{E}[Y_{\boldsymbol{x}}]\mathbb{E}[Y_{\boldsymbol{z}}]\right] \\
&= \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}}] - \mathbb{E}[Y_{\boldsymbol{x}}]\mathbb{E}[Y_{\boldsymbol{z}}] - \mathbb{E}[Y_{\boldsymbol{x}}]\mathbb{E}[Y_{\boldsymbol{z}}] + \mathbb{E}[Y_{\boldsymbol{x}}]\mathbb{E}[Y_{\boldsymbol{z}}] \\
&= \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}}] - \mathbb{E}[Y_{\boldsymbol{x}}]\mathbb{E}[Y_{\boldsymbol{z}}] = \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}}] - \mu(\boldsymbol{x})\mu(\boldsymbol{z}).
\end{aligned}$$

Therefore, the variance of the random field,

$$\operatorname{Var}(Y_{\boldsymbol{x}}) = \mathbb{E}[Y_{\boldsymbol{x}}^2] - \mathbb{E}[Y_{\boldsymbol{x}}]^2 = \mathbb{E}[Y_{\boldsymbol{x}}^2] - \mu^2(\boldsymbol{x}),$$

corresponds to the "diagonal" of the covariance, i.e.,

$$\operatorname{Var}(Y_{\boldsymbol{x}}) = \sigma^2 K(\boldsymbol{x},\boldsymbol{x}).$$

Let's now work out the MSE:

$$\begin{aligned}\operatorname{MSE}(\hat{Y}_{\boldsymbol{x}}) &= \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\right)^2\right] \\ &= \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{x}}] - 2\mathbb{E}[Y_{\boldsymbol{x}} \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}] + \mathbb{E}[\boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y} \boldsymbol{Y}^T \boldsymbol{w}(\boldsymbol{x})]\end{aligned}$$

Now use $\mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}}] = K(\boldsymbol{x}, \boldsymbol{z})$ (the covariance, since Y is centered):

$$\operatorname{MSE}(\hat{Y}_{\boldsymbol{x}}) = \sigma^2 K(\boldsymbol{x}, \boldsymbol{x}) - 2\boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \boldsymbol{k}(\boldsymbol{x})) + \boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \mathsf{K})\boldsymbol{w}(\boldsymbol{x}),$$

where

$\sigma^2 \boldsymbol{k}(\boldsymbol{x}) = \sigma^2 \begin{pmatrix} k_1(\boldsymbol{x}) & \cdots & k_N(\boldsymbol{x}) \end{pmatrix}^T$: with

$$\sigma^2 k_j(\boldsymbol{x}) = \sigma^2 K(\boldsymbol{x}, \boldsymbol{x}_j) = \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{x}_j}]$$

K: the covariance matrix has entries $\sigma^2 K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{E}[Y_{\boldsymbol{x}_i} Y_{\boldsymbol{x}_j}]$

Finding the minimum MSE is straightforward. Differentiation and equating to zero yields

$$-2\boldsymbol{k}(\boldsymbol{x}) + 2\mathsf{K}\boldsymbol{w}(\boldsymbol{x}) = 0,$$

and so the optimum weight vector is

$$\overset{\star}{\boldsymbol{w}}(\boldsymbol{x}) = \mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}).$$

We have shown that the (simple) kriging predictor

$$\hat{Y}_{\boldsymbol{x}} = \boldsymbol{k}(\boldsymbol{x})^T \mathsf{K}^{-1} \boldsymbol{Y}$$

is the best (in the MSE sense) linear unbiased predictor (BLUP).

Since we are given the observations $\boldsymbol{y}$ as realizations of $\boldsymbol{Y}$ we can compute the prediction

$$\hat{y}_{\boldsymbol{x}} = \boldsymbol{k}(\boldsymbol{x})^T \mathsf{K}^{-1} \boldsymbol{y}.$$

The MSE of the kriging predictor with optimal weights $\overset{\star}{\boldsymbol{w}}(\cdot)$,

$$\mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \hat{Y}_{\boldsymbol{x}}\right)^2\right] = \sigma^2 \left(K(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T \mathsf{K}^{-1} \boldsymbol{k}(\boldsymbol{x})\right),$$

is known as the kriging variance.

It allows us to give confidence intervals for our prediction. It also gives rise to a criterion for choosing an optimal parametrization of the family of covariance kernels used for prediction.

Remark

For Gaussian random fields the BLUP is also the best nonlinear unbiased predictor (see, e.g., [BTA04, Chapter 2]).

Remark

1. *The simple kriging approach just described is precisely how Krige [Kri51] introduced the method:*
   - *The unknown value to be predicted is given by a weighted average of the observed values, where the weights depend on the prediction location.*
   - *Usually one assigns a smaller weight to observations further away from* $\boldsymbol{x}$.

   The latter statement implies that one should be using kernels whose associated weights decay away from $\boldsymbol{x}$. *Positive definite translation invariant kernels have this property.*

2. *More advanced kriging variants are discussed in papers such as [SWMW89, SSS13], or books such as [Cre93, Ste99, BTA04].*

References I

[BTA04] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer, Dordrecht, 2004.

[Cre93] N. Cressie, *Statistics for Spatial Data*, revised ed., Wiley–Interscience, New York, 1993.

[FM15] G. E. Fasshauer and M. J. McCourt, *Kernel-based Approximation Methods using* MATLAB, Interdisciplinary Mathematical Sciences, vol. 19, World Scientific Publishing, Singapore, 2015.

[Kri51] D. G. Krige, *A statistical approach to some basic mine valuation problems on the Witwatersrand*, J. Chem. Met. & Mining Soc., S. Africa **52** (1951), no. 6, 119–139.

[Mey00] Carl D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, PA, 2000.

[SSS13] M. Scheuerer, R. Schaback, and M. Schlather, *Interpolation of spatial data — a stochastic or a deterministic problem?*, Eur. J. Appl. Math. **24** (2013), no. 4, 601–629.

References II

[Ste99] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, Berlin; New York, 1999.

[SWMW89] Jerome Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Stat. Sci. **4** (1989), no. 4, 409–423.