The Singular Value Decomposition and Least Squares Problems

Tom Lyche

University of Oslo

Norway

Applications of SVD

1. solving over-determined equations
2. statistics, principal component analysis
3. numerical determination of the rank of a matrix
4. search engines (Google,...)
5. theory of matrices
6. and lots of other applications...

Singular Value Decomposition

1. Works for any matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$
2. $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ with $\boldsymbol{U}$, $\boldsymbol{V}$ unitary and $\boldsymbol{\Sigma} = \left[\begin{smallmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{smallmatrix}\right] \in \mathbb{C}^{m,n}$
3. $\boldsymbol{\Sigma}_1 = \text{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$,
4. r is the rank of $\boldsymbol{A}$.
5. We define $\sigma_{r+1} = \cdots \sigma_n = 0$ if $r < n$ and call $\sigma_1, \ldots, \sigma_n$ the singular values of $\boldsymbol{A}$.
6. The columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of $\boldsymbol{U}$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ of $\boldsymbol{V}$ are called left- and right singular vectors respectively.

Relation to eigenpairs for $A^T A$ and A

1. $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}_i = \sigma_i^2\boldsymbol{v}_i$ for $i = 1, \ldots, n$.
2. The columns of $\boldsymbol{V}$ are orthonormal eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$
3. The columns of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$

Three forms of SVD

Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is the SVD of $\boldsymbol{A} \in \mathbb{R}^{m,n}$ and let $r := \#\Sigma_1$. We partition $\boldsymbol{U}$ and $\boldsymbol{V}$ as follows

- $\boldsymbol{U} = [\boldsymbol{U}_1, \boldsymbol{U}_2]$, $\boldsymbol{U}_1 \in \mathbb{R}^{m,r}$, $\boldsymbol{U}_2 \in \mathbb{R}^{m,m-r}$
- $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2]$, $\boldsymbol{V}_1 \in \mathbb{R}^{n,r}$, $\boldsymbol{V}_2 \in \mathbb{R}^{n,n-r}$.
- $\boldsymbol{A} = [\boldsymbol{U}_1, \boldsymbol{U}_2] \begin{bmatrix} \Sigma_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{bmatrix} = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^T$

The three forms

1. $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ — full form
2. $\boldsymbol{A} = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^T$ — compact form
3. $\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sum_{i=1}^{\min(m,n)} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ — outer product form

Subspaces of A

column space and the null space of a matrix

$$\operatorname{span}(\boldsymbol{A}) := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \text{ for some } \boldsymbol{x} \in \mathbb{R}^n\},$$
$$\ker(\boldsymbol{A}) := \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

- $\operatorname{span}(\boldsymbol{A})$ is a subspace of $\mathbb{R}^m$.
- $\ker(\boldsymbol{A})$ is a subspace of $\mathbb{R}^n$.

We say that A is a basis for a subspace $\mathcal{S}$ of $\mathbb{R}^m$ if

1. $\mathcal{S} = \operatorname{span}(\boldsymbol{A})$,
2. A has linearly independent columns, i. e., $\ker(\boldsymbol{A}) = \{\boldsymbol{0}\}$.

Recall the four fundamental subspaces
$\operatorname{span}(\boldsymbol{A})$, $\operatorname{span}(\boldsymbol{A}^T)$, $\ker(\boldsymbol{A})$, $\ker(\boldsymbol{A}^T)$.

The 4 fundamental Subspaces

Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the SVD of $A \in \mathbb{R}^{m,n}$. Then $\boldsymbol{A}^T = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T$ and $\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma}$, $\boldsymbol{A}^T\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^T$ or

$$\boldsymbol{A}\left[\boldsymbol{V}_1, \boldsymbol{V}_2\right] = \left[\boldsymbol{U}_1, \boldsymbol{U}_2\right]\begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{A}^T\left[\boldsymbol{U}_1, \boldsymbol{U}_2\right] = \left[\boldsymbol{V}_1, \boldsymbol{V}_2\right]\begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

$\boldsymbol{A}\boldsymbol{V}_1 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1$, $\quad\boldsymbol{U}_1$ is an orthonormal basis for $\operatorname{span}(\boldsymbol{A})$

$\boldsymbol{A}^T\boldsymbol{U}_2 = \mathbf{0}$, $\quad\boldsymbol{U}_2$ is an orthonormal basis for $\ker(\boldsymbol{A}^T)$

$\boldsymbol{A}^T\boldsymbol{U}_1 = \boldsymbol{V}_1\boldsymbol{\Sigma}_1$, $\quad\boldsymbol{V}_1$ is an orthonormal basis for $\operatorname{span}(\boldsymbol{A}^T)$

$\boldsymbol{A}\boldsymbol{V}_2 = \mathbf{0}$, $\quad\boldsymbol{V}_2$ is an orthonormal basis for $\ker(\boldsymbol{A})$.

We obtain the fundamental relations

1. $\dim(\operatorname{span}(\boldsymbol{A})) + \dim(\ker(\boldsymbol{A})) = \#\boldsymbol{A}$:=number of columns of $\boldsymbol{A}$,
2. $\dim(\operatorname{span}(\boldsymbol{A}^T)) = \dim(\operatorname{span}(\boldsymbol{A})) =:$ rank$(\boldsymbol{A}) = \#\boldsymbol{\Sigma}_1$.

Existence of SVD

Theorem 1. *Every matrix has an SVD.*

Uniqueness

- If the SVD of $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ then $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T$.
- Thus $\sigma_1^2, \ldots \sigma_n^2$ are uniquely given as the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ arranged in descending order.
- Taking the positive square root uniquely determines the singular values.
- From the proof of the existence theorem it follows that the orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are in general not uniquely given.

Application I, rank

- Gauss-Jordan cannot be used to determine rank numerically
- Use singular value decomposition
- numerically will normally find $\sigma_n > 0$.
- Determine minimal r so that $\sigma_{r+1}, \ldots, \sigma_n$ are "close" to round off unit.

Application II, overdetermined Equati

- Given $\boldsymbol{A}^{m,n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.
- The system $\boldsymbol{Ax} = \boldsymbol{b}$ is over-determined if $m > n$.
- This system has a solution if $\boldsymbol{b} \in \operatorname{span}(\boldsymbol{A})$, the column space of $\boldsymbol{A}$, but normally this is not the case and we can only find an approximate solution.
- A general approach is to choose a vector norm $\|\cdot\|$ and find $\boldsymbol{x}$ which minimizes $\|\boldsymbol{Ax} - \boldsymbol{b}\|$.
- We will only consider the Euclidian norm here.

The Least Squares Problem

- Given $\boldsymbol{A}^{m,n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ with $m \geq n \geq 1$. The problem to find $\boldsymbol{x} \in \mathbb{R}^n$ that minimizes $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ is called the least squares problem.
- A minimizing vector $\boldsymbol{x}$ is called a least squares solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.
- Several ways to analyze:
- Quadratic minimization
- Orthogonal Projections
- SVD

Quadratic minimization

- Define function $E : \mathbb{R}^n \to \mathbb{R}$ by $E(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$
- $E(\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} - 2\boldsymbol{c}^T\boldsymbol{x} + \alpha$, where
- $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A}$, $\boldsymbol{c} := \boldsymbol{A}^T\boldsymbol{b}$ and $\alpha := \boldsymbol{b}^T\boldsymbol{b}$.
- $\boldsymbol{B}$ is positive semidefinite and positive definite if $\boldsymbol{A}$ has rank n.
- Since the Hessian $\boldsymbol{H}E(\boldsymbol{x}) := \left(\frac{\partial^2 E(\boldsymbol{x})}{\partial x_i \partial x_j}\right) = 2\boldsymbol{B}$ we can find minimum by setting partial derivatives equal zero.
- $\nabla E(\boldsymbol{x}) := \left(\frac{\partial E(\boldsymbol{x})}{\partial x_i}\right) = 2(\boldsymbol{B}\boldsymbol{x} - \boldsymbol{c}) = \boldsymbol{0}$
- Normal equations $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$.

A simple example

- $$x_1 = 1 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$$
$$\begin{aligned} x_1 &= 1 \\ x_1 &= 1, \\ x_1 &= 2 \end{aligned} \quad \boldsymbol{A} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{x} = [x_1], \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix},$$
- Quadratic minimization problem:
$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = (x_1 - 1)^2 + (x_1 - 1)^2 + (x_1 - 2)^2.$$
- Setting the first derivative with respect to x_1 equal to zero we obtain $2(x_1 - 1) + 2(x_1 - 1) + 2(x_1 - 2) = 0$ or $6x_1 - 8 = 0$ or $x_1 = 4/3$
- The second derivative is positive (it is equal to 6) and $x = 4/3$ is a global minimum.

Theory; Direct sum and Orthogonal Su

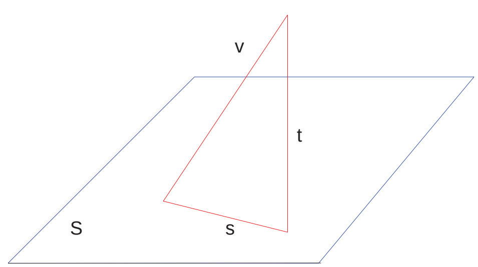
Suppose $\mathcal{S}$ and $\mathcal{T}$ are subspaces of a vector space $(\mathcal{V}, \mathbb{F})$. We define

1. Sum: $\mathcal{X} := \mathcal{S} + \mathcal{T} := \{\boldsymbol{s} + \boldsymbol{t} : \boldsymbol{s} \in \mathcal{S}$ and $\boldsymbol{t} \in \mathcal{T}\}$;
2. Direct Sum: If $\mathcal{S} \cap \mathcal{T} = \{\mathbf{0}\}$, then $\mathcal{S} \oplus \mathcal{T} := \mathcal{S} + \mathcal{T}$.
3. Orthogonal Sum: Suppose $(\mathcal{V}, \mathbb{F}, \langle \cdot, \cdot \rangle)$ is an inner product space. Then $\mathcal{S} \oplus \mathcal{T}$ is an orthogonal sum if $\langle \boldsymbol{s}, \boldsymbol{t} \rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$ and all $\boldsymbol{t} \in \mathcal{T}$.
4. orthogonal complement:
   $\mathcal{T} = \mathcal{S}^{\perp} := \{\boldsymbol{x} \in \mathcal{X} : \langle \boldsymbol{s}, \boldsymbol{x} \rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}\}$.

Basic facts

Lemma 1. *Suppose $\mathcal{S}$ and $\mathcal{T}$ are subspaces of a vector space $(\mathcal{V}, \mathbb{F})$.*

1. $\mathcal{S} + \mathcal{T} = \mathcal{T} + \mathcal{S}$ *and* $\mathcal{S} + \mathcal{T}$ *is a subspace of* $\mathcal{V}$.
2. $\dim(\mathcal{S} + \mathcal{T}) = \dim \mathcal{S} + \dim \mathcal{T} - \dim(\mathcal{S} \cap \mathcal{T})$
3. $\dim(\mathcal{S} \oplus \mathcal{T}) = \dim \mathcal{S} + \dim \mathcal{T}$. *Every* $\boldsymbol{v} \in \mathcal{S} \oplus \mathcal{T}$ *can be decomposed uniquely as* $\boldsymbol{v} = \boldsymbol{s} + \boldsymbol{t}$, *where* $\boldsymbol{s} \in \mathcal{S}$ *and* $\boldsymbol{t} \in \mathcal{T}$. $\boldsymbol{s}$ *is called the projection of* $\boldsymbol{v}$ *into* $\mathcal{S}$.
4. *Pythagoras: If* $\langle \boldsymbol{s}, \boldsymbol{t} \rangle = 0$ *then* $\|\boldsymbol{s} + \boldsymbol{t}\|^2 = \|\boldsymbol{s}\|^2 + \|\boldsymbol{t}\|^2$.
5. *Here* $\|\boldsymbol{v}\| := \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}$.



Column space of A and null space of A

- $\mathbb{R}^m = \operatorname{span}(\boldsymbol{A}) \oplus \ker(\boldsymbol{A}^T)$ and this is an orthogonal sum.
- Thus $\ker(\boldsymbol{A}^T) = \operatorname{span}(\boldsymbol{A})^\perp$ the orthogonal complement of $\operatorname{span}(\boldsymbol{A})$.
- Example

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \operatorname{span}(\boldsymbol{A}) = \operatorname{span}(\boldsymbol{e}_1, \boldsymbol{e}_2), \quad \ker(\boldsymbol{A}^T) = \boldsymbol{e}_3.$$

Proof that $\mathbb{R}^m = \text{span}(\boldsymbol{A}) \oplus \ker(\boldsymbol{A}^T)$ using SVD

- $\boldsymbol{s}^T\boldsymbol{t} = 0$ for all $\boldsymbol{s} \in \text{span}(\boldsymbol{A})$ and $\boldsymbol{t} \in \ker(A^T)$.
- For if $\boldsymbol{s} \in \text{span}(\boldsymbol{A})$ and $\boldsymbol{t} \in \ker(A^T)$ then $\boldsymbol{s} = \boldsymbol{A}\boldsymbol{x}$ for some $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{A}^T\boldsymbol{t} = \boldsymbol{0}$.
- But then $\langle \boldsymbol{s}, \boldsymbol{t} \rangle = (\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{t} = \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{t}) = 0$
- Suppose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T$ is the SVD of $\boldsymbol{A}$.
- Then $\boldsymbol{I} = \boldsymbol{U}\boldsymbol{U}^T = [\boldsymbol{U}_1\ \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \end{bmatrix} = \boldsymbol{U}_1\boldsymbol{U}_1^T + \boldsymbol{U}_2\boldsymbol{U}_2^T$.
- For any $\boldsymbol{b} \in \mathbb{R}^m$ we have $\boldsymbol{b} = (\boldsymbol{U}_1\boldsymbol{U}_1^T + \boldsymbol{U}_2\boldsymbol{U}_2^T)\boldsymbol{b} = \boldsymbol{b}_1 + \boldsymbol{b}_2$, where
- $\boldsymbol{b}_1 := \boldsymbol{U}_1\boldsymbol{U}_1^T\boldsymbol{b} = \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b}$ with $\boldsymbol{A}^\dagger = \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^T$,
- and $\boldsymbol{b}_2 := \boldsymbol{U}_2\boldsymbol{U}_2^T$ belongs to $\ker(\boldsymbol{A}^T)$ since $\boldsymbol{A}^T\boldsymbol{b}_2 = (\boldsymbol{V}_1\boldsymbol{\Sigma}_1\boldsymbol{U}_1^T)\boldsymbol{U}_2\boldsymbol{U}_2^T\boldsymbol{b} = \boldsymbol{V}_1\boldsymbol{\Sigma}_1(\boldsymbol{U}_1^T\boldsymbol{U}_2)\boldsymbol{U}_2^T\boldsymbol{b} = \boldsymbol{0}$.

Projections and pseudoinverse

- $\boldsymbol{b}_1 := \boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{b}$ is the projection of $\boldsymbol{b}$ into $\operatorname{span}(\boldsymbol{A})$.
- The matrix $\boldsymbol{A}^\dagger := \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^T = \boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{U}^T \in \mathbb{R}^{n,m}$ is called the pseudoinverse of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \in \mathbb{R}^{m,n}$.
- $\boldsymbol{\Sigma}^\dagger := \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{n,m}$ is the pseudoinverse of $\boldsymbol{\Sigma}$.
- $\boldsymbol{b}_2 := (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\dagger)\boldsymbol{b}$ is the projection of $\boldsymbol{b}$ into $\ker(\boldsymbol{A}^T)$.
- Example

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{I}_3\boldsymbol{A}\boldsymbol{I}_2, \quad \boldsymbol{A}^\dagger = \boldsymbol{I}_2\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

- $\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, \quad \boldsymbol{b}_1 = \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix}$
- $\boldsymbol{b}_2 = (\boldsymbol{I}_3 - \boldsymbol{A}\boldsymbol{A}^\dagger)\boldsymbol{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}\boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ b_3 \end{bmatrix}$

LSQ; Existence and Uniqueness

Theorem 2. *The least squares problem always has a solution. The solution is unique if and only if $\boldsymbol{A}$ has linearly independent columns.*

Proof.

- Let $\boldsymbol{b} = \boldsymbol{b}_1 + \boldsymbol{b}_2$, where $\boldsymbol{b}_1 \in \operatorname{span}(\boldsymbol{A})$ is the (orthogonal) projection of $\boldsymbol{b}$ into $\operatorname{span}(\boldsymbol{A})$ and $\boldsymbol{b}_2 \in \ker(\boldsymbol{A}^T)$.
- Since $\boldsymbol{b}_1 \in \operatorname{span}(\boldsymbol{A})$ there is an $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$. Thus $\boldsymbol{b}_2 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$.
- By Pythagoras, for any $\boldsymbol{s} \in \operatorname{span}(\boldsymbol{A})$ with $\boldsymbol{s} \neq \boldsymbol{b}_1$

$$\|\boldsymbol{b} - \boldsymbol{s}\|^2 = \|\boldsymbol{b}_1 - \boldsymbol{s}\|^2 + \|\boldsymbol{b}_2\|^2 = \|\boldsymbol{b}_1 - \boldsymbol{s}\|^2 + \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 > \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2.$$

- Since the projection $\boldsymbol{b}_1$ is unique, the least squares solution $\boldsymbol{x}$ is unique if and only if $\boldsymbol{A}$ has linearly independent columns.

$\square$

The Normal Equations

Theorem 3. *Any solution $\boldsymbol{x}$ of the least squares problem is a solution of the linear system*

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

The system is nonsingular if and only if $\boldsymbol{A}$ has linearly independent columns.

Proof.

- Since $\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \in \ker(\boldsymbol{A}^T)$, we have $\boldsymbol{A}^T(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{0}$ or $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$.
- $\boldsymbol{A}^T\boldsymbol{A}$ is nonsingular. Suppose $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ for some $\boldsymbol{x} \in \mathbb{R}^n$. Then $0 = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|_2^2$. Hence $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ which implies that $\boldsymbol{x} = \boldsymbol{0}$ if and only if $\boldsymbol{A}$ has linearly independent columns.

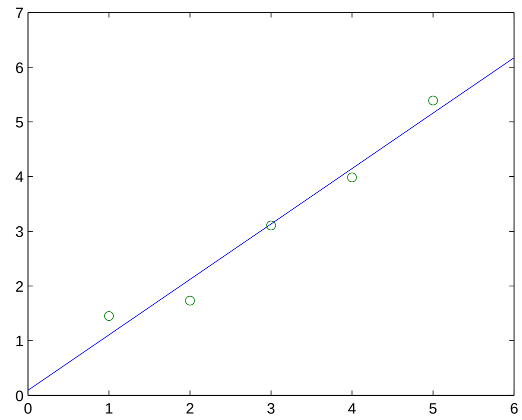
$\square$

The linear system $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$ is called the **normal equations**.

Linear Regression

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \min_{x_1 x_2} \sum_{i=1}^{m} (x_1 + t_i x_2 - y_i)^2.$$

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \quad b = \begin{bmatrix} 1.4501 \\ 1.7311 \\ 3.1068 \\ 3.9860 \\ 5.3913 \end{bmatrix}$$



$$A^T A = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix} \quad c = A^T b = \begin{bmatrix} 16.0620 \\ 58.6367 \end{bmatrix} \quad \begin{aligned} 5x_1 + 15x_2 &= 16.0620 \\ 15x_1 + 55x_2 &= 58.6367 \end{aligned}$$

$$\begin{aligned} x_1 &= 0.0772 \\ x_2 &= 1.0451 \end{aligned}, \quad x = A \backslash b$$

Analysis of LSQ using $A = U\Sigma V^T$

Define $\boldsymbol{y} := \boldsymbol{V}^T\boldsymbol{x} = \begin{bmatrix} \boldsymbol{V}_1^T\boldsymbol{x} \\ \boldsymbol{V}_2^T\boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}$. Recall $\|\boldsymbol{U}\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2$ for any $\boldsymbol{U} \in \mathbb{R}^{n,n}$ with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and any $\boldsymbol{v} \in \mathbb{R}^n$.

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = \|\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{b} - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{y}\|_2^2 = \|\boldsymbol{U}^T\boldsymbol{b} - \boldsymbol{\Sigma}\boldsymbol{y}\|_2^2 = \| \begin{bmatrix} \boldsymbol{U}_1^T\boldsymbol{b} \\ \boldsymbol{U}_2^T\boldsymbol{b} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}$$

$$= \| \begin{bmatrix} \boldsymbol{U}_1^T\boldsymbol{b} - \boldsymbol{\Sigma}_1\boldsymbol{y}_1 \\ \boldsymbol{U}_2^T\boldsymbol{b} \end{bmatrix} \|_2^2 = \|\boldsymbol{U}_1^T\boldsymbol{b} - \boldsymbol{\Sigma}_1\boldsymbol{y}_1\|_2^2 + \|\boldsymbol{U}_2^T\boldsymbol{b}\|_2^2.$$

We have $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 \geq \|\boldsymbol{U}_2^T\boldsymbol{b}\|_2$ for all $\boldsymbol{x} \in \mathbb{R}^n$ with equality if and only if

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{y} = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^T\boldsymbol{b} \\ \boldsymbol{y}_2 \end{bmatrix} = \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^T\boldsymbol{b} + \boldsymbol{V}_2\boldsymbol{y}_2, \text{ for all } \boldsymbol{y}_2 \in \mathbb{R}^{n-r}. \quad (1)$$

The general solution of $\min \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$

- The columns of $\boldsymbol{V}_2$ is a basis for $\ker(\boldsymbol{A})$ so that $\ker(\boldsymbol{A}) = \{\boldsymbol{z} = \boldsymbol{V}_2\boldsymbol{y}_2 : \boldsymbol{y}_2 \in \mathbb{R}^{n-r}\}$.
- Therefore the solution set is

$$\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 \text{ is minimized } \} = \boldsymbol{A}^\dagger \boldsymbol{b} + \ker(\boldsymbol{A}).$$

- If $r = n$ then $\boldsymbol{A}$ has linearly independent columns and $\boldsymbol{A}^T\boldsymbol{A}$ is nonsingular.
- Since $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$ we obtain $\boldsymbol{A}^\dagger = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ in this case.

The Minimal Norm Solution

- Suppose $\boldsymbol{A}$ is rank deficient ($r < n$).
- Let $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{V}_2 \boldsymbol{y}_2$ be a solution of min$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$.
- $\boldsymbol{A}^\dagger \boldsymbol{b}$ and $\boldsymbol{V}_2 \boldsymbol{y}_2$ are orthogonal
- By Pythagoras $\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{A}^\dagger \boldsymbol{b}\|_2^2 + \|\boldsymbol{V}_2 \boldsymbol{y}_2\|_2^2 \geq \|\boldsymbol{A}^\dagger \boldsymbol{b}\|_2^2$.
- The solution $\boldsymbol{x}^* = \boldsymbol{A}^\dagger \boldsymbol{b}$ is called the minimal norm solution to the LSQ problem.
- Orthogonal. Since $\boldsymbol{V}_2^T \boldsymbol{A}^\dagger = (\boldsymbol{V}_2^T \boldsymbol{V}_1)\Sigma_1^{-1} \boldsymbol{U}_1^T = \boldsymbol{0}$ we have $(\boldsymbol{V}_2 \boldsymbol{y}_2)^T \boldsymbol{A}^\dagger \boldsymbol{b} = 0$ for any $\boldsymbol{y}_2$.

More on the pseudoinverse

- If $\boldsymbol{A}$ is square and nonsingular then $\boldsymbol{A}^\dagger = \boldsymbol{A}^{-1}$.
- $\boldsymbol{A}^\dagger$ is always defined.
- Thus $\boldsymbol{A}^\dagger$ is a generalization of usual inverse.
- If $\boldsymbol{B} \in \mathbb{R}^{n,m}$ satisfies
  1. $\boldsymbol{ABA} = \boldsymbol{A}$
  2. $\boldsymbol{BAB} = \boldsymbol{B}$
  3. $(\boldsymbol{BA})^T = \boldsymbol{BA}$
  4. $(\boldsymbol{AB})^T = \boldsymbol{AB}$

  then $\boldsymbol{B} = \boldsymbol{A}^\dagger$.
- Thus $\boldsymbol{A}^\dagger$ is uniquely defined by these axioms.

Example

Show that the pseudoinverse of $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$ is $\boldsymbol{B} = \frac{1}{4}\left[\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{smallmatrix}\right]$.

We have $\boldsymbol{BA} = \frac{1}{2}\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ and $\boldsymbol{AB} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Thus

1. $\boldsymbol{ABA} = \boldsymbol{A}$
2. $\boldsymbol{BAB} = \boldsymbol{B}$
3. $(\boldsymbol{BA})^T = \boldsymbol{BA}$
4. $(\boldsymbol{AB})^T = \boldsymbol{AB}$

and hence $\boldsymbol{A}^\dagger = \boldsymbol{B}$.