Least Squares Problems

Tom Lyche

University of Oslo

Norway

Overdetermined Equations

- Given $A^{m,n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.
- The system $A\boldsymbol{x} = \boldsymbol{b}$ is over-determined if $m > n$.
- This system has a solution if $\boldsymbol{b} \in \text{Span}(A)$, the column space of 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 $\|A\boldsymbol{x} - \boldsymbol{b}\|$.
- We will only consider the Euclidian norm here.

The Least Squares Problem

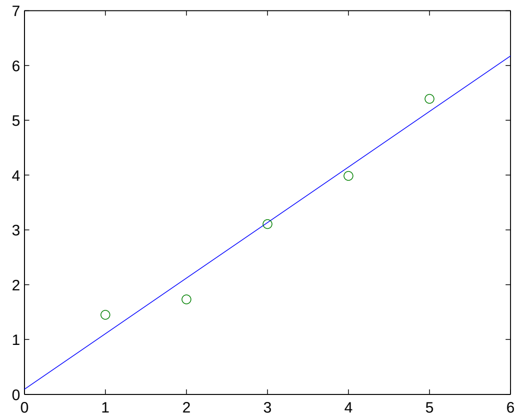
- Given $A^{m,n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ with $m \geq n \geq 1$. The problem to find $\boldsymbol{x} \in \mathbb{R}^n$ that minimizes $\|A\boldsymbol{x} - \boldsymbol{b}\|_2$ is called the least squares problem.
- A minimizing vector $\boldsymbol{x}$ is called a least squares solution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

Example 1

- $$x_1 = 1 \qquad\qquad\qquad \begin{bmatrix}1\\1\\1\end{bmatrix} \qquad\qquad\qquad \begin{bmatrix}1\\1\\2\end{bmatrix}$$
  $$\begin{aligned} x_1 &= 1 \\ x_1 &= 1, \quad A = \begin{bmatrix}1\\1\\1\end{bmatrix}, \quad \boldsymbol{x} = [x_1], \quad \boldsymbol{b} = \begin{bmatrix}1\\1\\2\end{bmatrix}, \\ x_1 &= 2 \end{aligned}$$
- Our approach is to minimize the least squares sum
  $$\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 = (x_1 - 1)^2 + (x_1 - 1)^2 + (x_1 - 2)^2.$$
- 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.

Linear regression

- Given m points $(t_i, y_i)_{i=1}^m$ in the (t, y) plane.
- Example:
  $(t_i, y_i)_{i=1}^5 = [(1, 1.4501), (2, 1.7311), (3, 3.1068), (4, 3.9860), (5, 5.3913)]$



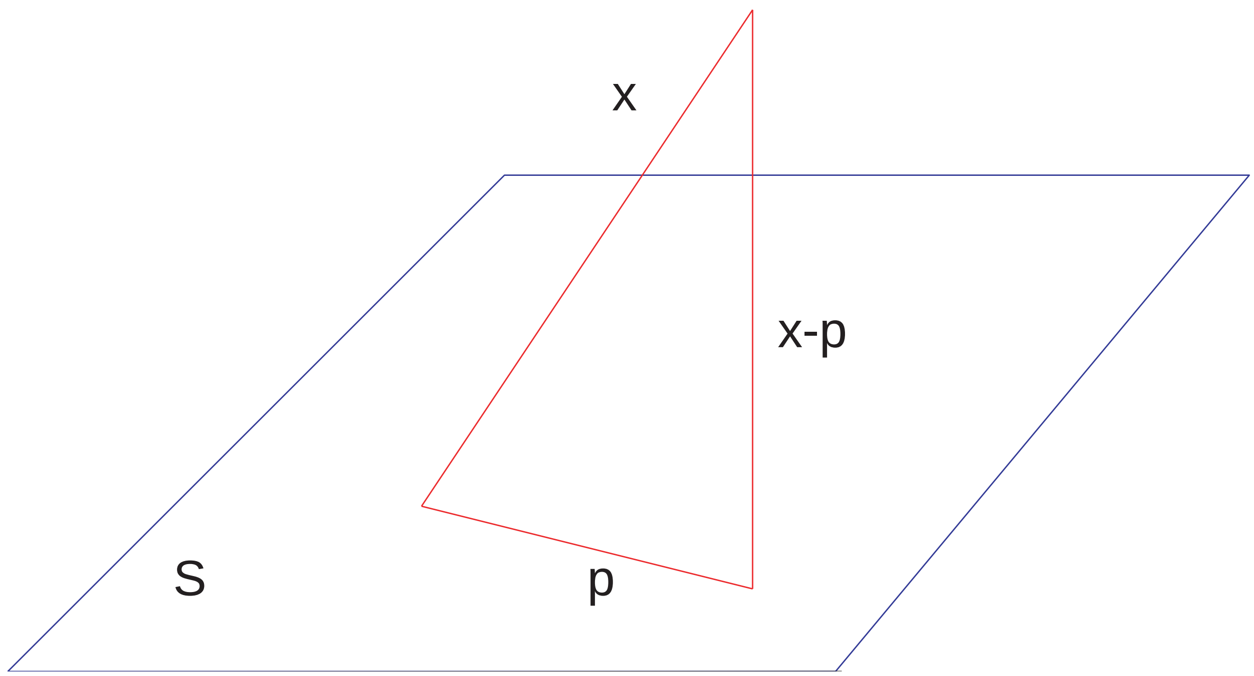
- Find a straight line $y(t) = x_1 + tx_2$ such that $y_i = x_1 + t_i x_2$ all i.
- We obtain a least squares problem with

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \sum_{i=1}^m (x_1 + t_i x_2 - y_i)^2.$$

Recall result on orthogonal projection

Theorem(Best Approximation) Let $\mathcal{S}$ be a subspace in a real or complex inner product space $(\mathcal{V}, \mathbb{F}, \langle \cdot, \cdot, \rangle)$. Let $x \in \mathcal{V}$, and $\boldsymbol{p} \in \mathcal{S}$. The following statements are equivalent

1. $\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle = 0, \quad$ for all $\boldsymbol{s} \in \boldsymbol{S}$.
2. $\|\boldsymbol{x} - \boldsymbol{s}\| > \|\boldsymbol{x} - \boldsymbol{p}\|$ for all $\boldsymbol{s} \in \mathcal{S}$ with $\boldsymbol{s} \neq \boldsymbol{p}$.



Existence and Uniqueness

Theorem 1. *The least squares problem always has a solution. The solution is unique if and only if A has linearly independent columns.*

Proof.

- We apply the inner product setup with $\mathcal{V} = \mathbb{R}^n$, the usual inner product in $\mathbb{R}^n$, $\mathcal{S}$ equals Span$(A) := \{A\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\}$, the column space of A, and $\boldsymbol{x} = \boldsymbol{b}$.
- The inner product norm is the Euclidian norm $\|\cdot\|_2$.
- Let $\boldsymbol{p}$ be the orthogonal projection of $\boldsymbol{b}$ into $\boldsymbol{C}(A)$. Since $\boldsymbol{p} \in$ Span(A) there is an $\boldsymbol{x}^* \in \mathbb{R}^n$ such that $A\boldsymbol{x}^* = \boldsymbol{p}$.
- If $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{x} \neq \boldsymbol{x}^*$ then $\boldsymbol{s} := A\boldsymbol{x} \in$ Span(A) and by the best approximation theorem $\|\boldsymbol{b} - A\boldsymbol{x}^*\|_2 = \|\boldsymbol{b} - \boldsymbol{p}\|_2 < \|\boldsymbol{b} - \boldsymbol{s}\|_2$.

Since $\boldsymbol{p}$ is unique and $A\boldsymbol{x}^* = \boldsymbol{p}$ the least squares problem has a unique solution if and only if A has linearly independent columns. □

Characterization of any least squares solution

By the best approximation theorem $\langle \boldsymbol{b} - \boldsymbol{p}, \boldsymbol{s} \rangle = 0$ for all $\boldsymbol{s} \in$ Span(A) and we see that $\boldsymbol{b} - \boldsymbol{p}$ belongs to the orthogonal complement Span$(A)^\perp$ of A. We now recall:

Theorem 2. *The orthogonal complement of the column space of a matrix $A \in \mathbb{R}^{m,n}$ is the null space of A^T. In symbols*

$$\textit{Span}(A)^\perp = \textit{Ker}(A^T) := \{\boldsymbol{y} \in \mathbb{R}^m : A^T\boldsymbol{y} = \boldsymbol{0}\}. \tag{1}$$

It follows that $\boldsymbol{x}^*$ minimizes $\|A\boldsymbol{x} - \boldsymbol{b}\|_2$ if and only if $\boldsymbol{b} - A\boldsymbol{x}^*$ belongs to Span$(A)^\perp$ or $A^T(\boldsymbol{b} - A\boldsymbol{x}^*) = \boldsymbol{0}$.

We obtain the linear system $A^T A\boldsymbol{x}^* = A^T\boldsymbol{b}$

The Normal Equations

The least squares solution can be found by solving a linear system.

Theorem 3. *Suppose $A \in \mathbb{R}^{m,n}$ with $m > n$ and $\boldsymbol{b} \in \mathbb{R}^m$. The following is equivalent*

1. $\boldsymbol{x}^*$ *minimizes* $\|A\boldsymbol{x} - \boldsymbol{b}\|_2$.
2. $A^T A\boldsymbol{x}^* = A^T\boldsymbol{b}$. *(normal equations)*

The matrix $A^T A$ is nonsingular if and only if A has linearly independent columns.

Proof. The previous discussion shows that $1 \Leftrightarrow 2$. From an example in Chapter 8 we know that $A^T A$ is positive semidefinite and positive definite (and hence nonsingular) if and only if A has linearly independent columns. □

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

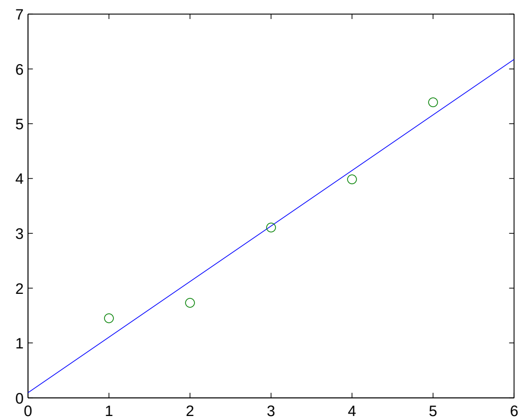
An observation

- We show that minimizing $\|A\boldsymbol{x}-\boldsymbol{b}\|_2$ and $\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2$ are equivalent. Define functions $f:\mathbb{R}^n \to [0,\infty)$, $\phi:[0,\infty)\to[0,\infty)$, and $g:\mathbb{R}^n\to[0,\infty)$ by
- $f(\boldsymbol{x}) := \|A\boldsymbol{x}-\boldsymbol{b}\|_2$
- $\phi(t) = t^2$
- $g(\boldsymbol{x}) := \phi(f(\boldsymbol{x})) = \|A\boldsymbol{x}-\boldsymbol{b}\|_2^2$.
- Then f has a global minimum at $\boldsymbol{x}^*$ if and only if g has a global minimum at $\boldsymbol{x}^*$.
- This follows since ϕ is strictly increasing.
- For if $h \neq 0$ then $g(x^*+h) - g(x^*) = \phi(f(x^*+h)) - \phi(f(x^*))$ so one difference is positive if and only the other is positive.

Linear Regression

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

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



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

$$\begin{matrix} x_1 = 0.0772 \\ x_2 = 1.0451 \end{matrix}, \quad x = A \backslash y$$

Cholesky Factorization

1. Form the normal equations $B = A^T A$ and $c = A^T b$.
2. Find the Cholesky factorization $B = LL^T$ and find x.

1. for i=1:n

```
  for j=i:n                          % Compute only one half of B
    B(i,j)=A(:,i)'*A(:,j);           % mn(n+1) flops
  end
  c(i)=A(:i)'*b;                     % mn flops
end
```

2. n^3/3 flops

Forming the normal equations requires $O(mn^2)$ flops and this represents more work than solving them.

Why not always use the normal equations?

- Forming $A^T A$ squares the condition number of A.
- If A is ill conditioned then $A^T A$ will be severely ill conditoned
- Sometimes in applications A does not have full rank.
- Consider two other methods
- QR factorization
- Singular value decomposition

Squaring the Condition Number

The difference between the computed solution y of $Bx = c$ and the exact solution x satisfies

$$\frac{1}{\text{cond}_2(B)}\frac{\|r\|_2}{\|c\|_2} \le \frac{\|x-y\|_2}{\|x\|_2} \le \text{cond}_2(B)\frac{\|r\|_2}{\|c\|_2},$$

where $r = c - By$. If $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n > 0$ are the singular values of A then $\sigma_1^2, \sigma_2^2 \ldots, \sigma_n^2$ are the singular values of $B = A^T A$ and

$$\text{cond}_2(B) = \frac{\sigma_1^2}{\sigma_n^2} = \big(\text{cond}_2(A)\big)^2$$

For this reason one should not use Cholesky factorization to solve the least squares problem when A has a large condition number.

QR-factorization

Suppose A has full rank and that $A = Q_1 R_1$ is a reduced QR-factorization of A, that is $Q_1 \in \mathbb{R}^{m,n}$ has orthonormal columns and $R_1 \in \mathbb{R}^{n,n}$ is nonsingular and upper triangular. Then

$$A^T A x = A^T b \Rightarrow R_1^T Q_1^T Q_1 R_1 x = R_1^T Q_1^T b \Rightarrow R_1^T R_1 x = R_1^T Q_1^T b.$$

Since R_1^T is nonsingular we find the least squares solution by solving the triangular system

$$R_1 x = Q_1^T b.$$

Example

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} -0.4472 & -0.6325 \\ -0.4472 & -0.3162 \\ -0.4472 & 0.0000 \\ -0.4472 & 0.3162 \\ -0.4472 & 0.6325 \end{bmatrix} * \begin{bmatrix} -2.2361 & -6.7082 \\ 0 & 3.1623 \end{bmatrix} = Q_1 R_1,$$

$$b = \begin{bmatrix} 1.4501 \\ 1.7311 \\ 3.1068 \\ 3.9860 \\ 5.3913 \end{bmatrix}, \quad c := Q_1^T b = \begin{bmatrix} -7.1831 \\ 3.3048 \end{bmatrix}$$

$$R_1 x = c \Leftrightarrow \begin{matrix} -2.2361 x_1 - 6.7082 x_2 = -7.1831 \\ 3.1623 x_2 = 3.3048 \end{matrix} \Leftrightarrow \begin{matrix} x_1 = 0.0772 \\ x_2 = 1.0451 \end{matrix}.$$

Cholesky and Householder

Cholesky

- $O(mn^2)$ flops
- works only for full rank problems
- squares the condition number of A

Householder

- $O(2mn^2) - n^3/6$ flops.
- Can be used for rank deficient problems

Discussion:

- Cholesky more econimical
- QR factorization should be used for problems with large condition numbers
- Cholesky cannot be used for rank deficient problems

The Singular Value Decomposition (SVD)

Suppose $A = U\Sigma V^T = U_1\Sigma_1 V_1^T$ is the SVD of $A \in \mathbb{R}^{m,n}$. Recall that if rank$(A) = r$ then

- $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}$
- $\Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r),\ \sigma_1 \geq \cdots \geq \sigma_r > 0$,
- $$U \in \mathbb{R}^{m,m},\ U^T U = I,\ U = (U_1, U_2),\ U_1 \in \mathbb{R}^{m,r},\ U_2 \in \mathbb{R}^{m,m-r}$$
- $V \in \mathbb{R}^{n,n},\ V^T V = I,\ V = (V_1, V_2),\ V_1 \in \mathbb{R}^{n,r},\ V_2 \in \mathbb{R}^{n,n-r}$.
- $AV_2 = 0$ and V_2 is an orthonormal basis for $N(A)$

The Pseudo-inverse of $A = U\Sigma V^T$

The pseudo-inverse of $A \in \mathbb{R}^{m,n}$ is a matrix $A^\dagger \in \mathbb{R}^{n,m}$ given by

$$A^\dagger := V\Sigma^\dagger U^T = V_1\Sigma_1^{-1}U_1^T, \text{ where } \Sigma^\dagger := \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \tag{2}$$

Example:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} = U\Sigma V^T = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}$$

$$\Sigma^\dagger = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A^\dagger = V\Sigma^\dagger U^T = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

$A^\dagger := V\Sigma^\dagger U^T$

- Since Σ_1 is nonsingular $A^\dagger$ is always defined.
- generalization of usual inverse. To see this note that
- $AA^\dagger = U_1\Sigma_1 V_1^T V_1 \Sigma_1^{-1} U_1^T = U_1 U_1^T$.
- $A^\dagger A = V_1 \Sigma_1^{-1} U_1^T U_1 \Sigma_1 V_1^T = V_1 V_1^T$.
- If $A \in \mathbb{R}^{n,n}$ is square and nonsingular then $U_1 = U$ and $V_1 = V$ and $AA^\dagger = A^\dagger A = I$. Thus $A^\dagger$ is the usual inverse.

Analysis of LSQ using SVD

Define

$$\boldsymbol{c} := U^T\boldsymbol{b} = \begin{bmatrix} U_1^T\boldsymbol{b} \\ U_2^T\boldsymbol{b} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \quad \boldsymbol{y} := V^T\boldsymbol{x} = \begin{bmatrix} V_1^T\boldsymbol{x} \\ V_2^T\boldsymbol{x} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\|\boldsymbol{b} - A\boldsymbol{x}\|_2^2 = \|UU^T\boldsymbol{b} - U\Sigma V^T\boldsymbol{x}\|_2^2 = \|\boldsymbol{c} - \Sigma\boldsymbol{y}\|_2^2 = \left\|\begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix} - \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}\right\|$$

$$= \left\|\begin{bmatrix} \boldsymbol{c}_1 - \Sigma_1\boldsymbol{y}_1 \\ \boldsymbol{c}_2 \end{bmatrix}\right\|_2^2 = \|\boldsymbol{c}_1 - \Sigma_1\boldsymbol{y}_1\|_2^2 + \|\boldsymbol{c}_2\|_2^2.$$

We have $\|\boldsymbol{b} - A\boldsymbol{x}\|_2 \geq \|\boldsymbol{c}_2\|_2$ for all $\boldsymbol{x} \in \mathbb{R}^n$ with equality iff

$$\boldsymbol{x} = V\boldsymbol{y} = \begin{bmatrix} V_1 & V_2 \end{bmatrix}\begin{bmatrix} \Sigma_1^{-1}\boldsymbol{c}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = A^\dagger\boldsymbol{b} + V_2\boldsymbol{y}_2, \text{ for all } \boldsymbol{y}_2 \in \mathbb{R}^{n-r}. \quad (3)$$

Projections

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

$$\boldsymbol{x} = Vy = [\, V_1 \;\; V_2 \,] \begin{bmatrix} \Sigma_1^{-1} U_1^T \boldsymbol{b} \\ \boldsymbol{y}_2 \end{bmatrix} = A^\dagger \boldsymbol{b} + V_2 \boldsymbol{y}_2, \text{ for all } \boldsymbol{y}_2 \in \mathbb{R}^{n-r}. \tag{4}$$

- The solution is unique if and only if $V_1 = V$ that is if and only if $r =$ rank$(A) = n$.
- Since $AV_2 = 0$ we have $A\boldsymbol{x} = AA^\dagger \boldsymbol{b} + AV_2 \boldsymbol{y}_2 = AA^\dagger \boldsymbol{b}$
- $\boldsymbol{b}_1 := AA^\dagger \boldsymbol{b} = U_1 U_1^T \boldsymbol{b}$ is the vector projection of $\boldsymbol{b}$ onto the column space Span(A) of A
- $\boldsymbol{b}_2 := \boldsymbol{b} - \boldsymbol{b}_1 = (I - U_1 U_1^T)\boldsymbol{b} = U_2 U_2^T \boldsymbol{b}$ is the vector projection of $\boldsymbol{b}$ onto $N(A^T)$

The Minimal Norm Solution

Suppose A is rank deficient ($r < n$) and let $\boldsymbol{z} := V\boldsymbol{y} = [\, V_1 \ V_2 \,] \begin{bmatrix} \Sigma_1^{-1} U_1^T \boldsymbol{b} \\ \boldsymbol{y}_2 \end{bmatrix} = A^\dagger \boldsymbol{b} + V_2 \boldsymbol{y}_2$ with $\boldsymbol{y}_2 \in \mathbb{R}^{n-r}$ and nonzero. Let $\boldsymbol{x} = A^\dagger \boldsymbol{b}$ be the solution corresponding to $\boldsymbol{y}_2 = 0$. Then $\|\boldsymbol{z}\|_2^2 = \|A^\dagger \boldsymbol{b}\|_2^2 + \|\boldsymbol{y}_2\|_2^2 > \|\boldsymbol{x}\|_2^2$ The vector $\boldsymbol{x} = A^\dagger \boldsymbol{b}$ is the minimal norm solution to the LSQ problem.

Three Algorithms for solving LSQ

- Solving the normal equations by Cholesky factorization
- Using the QR-factorization of A
- Using the SVD of A

QR with column pivoting or SVD is used for rank deficient problems. (Non unique solutions)

Minimal norm solution by SVD

The minimimal norm solution is $\boldsymbol{x} = A^{\dagger}\boldsymbol{b} = V_1\Sigma_1^{-1}U_1^T\boldsymbol{b}$.

1. Compute $A = U\Sigma V^T = U_1\Sigma_1 V_1^T$, the SVD of A.
2. $\boldsymbol{c} = U_1^T\boldsymbol{b}$ — rm flops
3. $\boldsymbol{d} = \Sigma_1^{-1}\boldsymbol{c}$ — r flops
4. $\boldsymbol{x} = V_1\boldsymbol{d}$ — rn flops

This is a great method if the SVD of A is known.

Computing the SVD of $A \in \mathbb{R}^{m,n}$

1. Transform A to bidiagonal form using Householder reflections
2. Use the QR-method to find the singular values.

1. Transform A to bidiagonal form using Householder reflections
2. Use the QR-method to find the singular values.