Lecture 12

Today

Methods for solving least squares problems

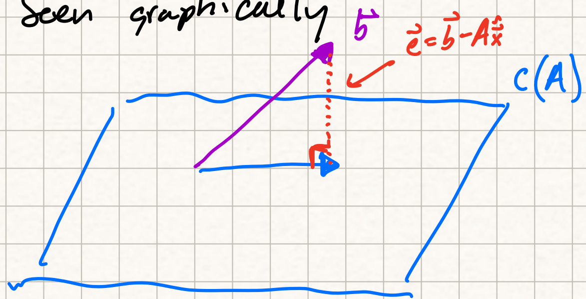
- Solving the normal equations
  - Cholesky
  - QR
- Pseudoinverse
- Penalized least squares

Solving the normal equations

$A\vec{x} = b$ If A is n by p, we can try to minimize $\|A\hat{\vec{x}} - b\|_2^2$.

It turns out minimizing $\|A\hat{\vec{x}} - b\|_2^2$ is equivalent to finding a solution to the normal equation $A^T A\hat{\vec{x}} = A^T\vec{b}$ when $A^T A$ is nonsingular.

1. Show using calculus.
2. Seen graphically



The projection of $\vec{b}$ onto $C(A)$, $A\hat{\vec{x}}$ is orthogonal to the error $\vec{b} - A\hat{\vec{x}}$.

This means that $A^T\vec{b} - A^T A\vec{x} = \vec{0}$ and thus $\boxed{A^T\vec{b} = A^T A\vec{x}}$

To solve normal equations:

when $A^T A$ is invertible, $\hat{\vec{x}} = (A^T A)^{-1} A^T\vec{b}$

def The projection matrix $P = A(A^T A)^{-1}A^T$ maps $\vec{b}$ onto the column space of A.

def The projection $A\vec{x} = P\vec{b} = A(A^T A)^{-1}A^T\vec{b}$

In practice, we generally do not want to invert $(A^T A)$

Cholesky decomposition

theorem if $S \in \mathbb{R}^{n\times n}$ is symmetric positive definite, there exists a unique lower triangular matrix $L \in \mathbb{R}^{n\times n}$ with positive diagonal entries such that $S = LL^T$

L is called the Cholesky factor and LL^T is the Cholesky factorization

Q. When is $A^T A$ symmetric? Always

When is $A^T A$ symmetric positive definite?

$\vec{x}A^T A\vec{x} > 0$

$\|A\vec{x}\| > 0$

when A is full rank (has linearly independent cols)

$A^TA = LL^T$

Solve $A^TA\hat{\vec{x}} = A^T\vec{b}$ by letting $A^T\vec{b} = \vec{c}$ and $A^TA = LL^T$

$= LL^T\hat{\vec{x}} = \vec{c}$

in R

chol(A)

Solve $L\vec{y} = \vec{c}$ (forward)

and $L^T\hat{\vec{x}} = \vec{y}$ (backward)

$x \sim N(0,1)$

Note if $\vec{x} \sim N(\vec{0}, I_n)$, $L\vec{x} \sim N(\vec{0}, LL^T)$ $\quad \alpha x \sim N(0, \alpha^2)$

Solving normal equations via $A = QR$

The condition number of A^TA is $\|A^TA\|_2 \|(A^TA)^{-1}\|_2 = \frac{\sigma_1^2}{\sigma_n^2}$

Instead of solving $\hat{\vec{x}} = A(A^TA)^{-1}A^T\vec{b}$, we can use the QR decomposition:

Write $A = QR$ where Q is an orthogonal matrix $\quad K_2(Q) = 1$

R is an upper triangular matrix

We can compute QR using Gram-Schmidt.

Then $\hat{X} = (A^TA)^{-1}A^T\vec{b} = (R^TQ^TQR)^{-1}R^TQ^T\vec{b} = (R^TR)^{-1}R^TQ^T\vec{b} = R^{-1}Q^T\vec{b}$

The benefit of QR is not speed, but accuracy.

It turns out Gram-Schmidt is not the best way to compute QR

– instead you can use Householder Rotations

What if A^TA is not invertible?

Pseudoinverse.

If A is invertible, the solution to $A\vec{x} = \vec{b}$ is $A^{-1}\vec{b}$.

If A is not square, we can still compute the pseudoinverse

Desired properties

- If A is invertible, we want the pseudoinverse $A^+ = A^{-1}$
- If A is m by n, A^+ is n by m.
- $A^+A\vec{x} = \vec{x}$ when $\vec{x}$ is in row space of A $\qquad (A^+A)$ is $n \times n$
- $AA^+\vec{b} = \vec{b}$ when $\vec{b}$ is in column space of A

def The Moore-Penrose pseudo inverse of $A \in \mathbb{R}^{m \times n}$ satisfies

- $AA^+A = A$
- $A^+AA^+ = A^+$
- $(AA^+)^T = AA^+$
- $(A^+A)^T = A^+A$

theorem A^+ always exists and is unique.

How do we compute A^+?

The pseudoinverse of $A = U\Sigma V^T$ is $A^+ = V\Sigma^+ U^T$

How do we take pseudoinverse of Σ?

ex
$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \Sigma^+ = \begin{bmatrix} 1/\sigma_1 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Check the four conditions.

It turns out the pseudoinverse allows us to compute the minimum norm least squares solution to $A\vec{x} = \vec{b}$

let $\hat{\vec{x}}^+ = A^+\vec{b}$, Then

- $\hat{\vec{x}}^+$ minimizes $\|\vec{b} - A\vec{x}\|_2^2$ (least squares)
- if another $\hat{\vec{x}}$ minimizes $\|\vec{b} - A\vec{x}\|_2^2$, then $\|\hat{\vec{x}}^+\| \le \|\hat{\vec{x}}\|$ (minimum norm)

So we can use SVD to compute A^+ and solve least squares problems, even if A^TA is not invertible.

Penalized least squares.

If there is no unique solution to $A^T A\vec{x} = A^T\vec{b}$, there will be a unique solution to $(A^T A + \delta^2 I)\vec{x} = A^T\vec{b}$

This is equivalent to minimizing

$$\|A\vec{x}-\vec{b}\|_2^2 + \delta^2\|\vec{x}\|_2^2 \leftarrow \text{penalty term}$$

This approach is often called ridge regression

It turns out $A^TA + \delta^2 I$ is invertible for $\delta > 0$

ex Consider the 1 by 1 matrix $A = [\sigma]$

$$(A^TA + \delta^2 I)^{-1}A^T = \frac{\sigma}{\sigma^2+\delta^2}$$

Then the limit as $\delta \to 0$ is 0 if $\sigma = 0$ and $\frac{1}{\sigma}$ otherwise.

ex Consider a diagonal matrix Σ

$(\Sigma^T\Sigma + \delta^2 I)^{-1}\Sigma^T$ has diagonal entries $\dfrac{\sigma_i}{\sigma_i^2+\delta^2}$

So it can be shone that the limit of $(A^TA + \delta^2 I)^{-1}A^T$ is A^+

To see this, consider that as $\delta \to 0$, $(\Sigma^T\Sigma + \delta^2 I)^{-1}\Sigma^T \to \Sigma^+$