Week 8: aim to cover

- Data fitting, linear least squares (Lecture 15)
- Linear least squares (Lab 8)
- QR factorization, SVD (Lecture 16)

Data fitting

A very common task:
given a set of data $\{x_i, y_i\}, i = 1 \cdots m$ with observational error
find a line $y = a + bx$ that 'fits' the data
$\implies$ we want $y_i = a + bx_i + \epsilon_i$
If $m = 2$, can interpolate; **what if** $m > 2$**?**

Two possible approaches

1 from linear algebra, 1 from calculus (optimization)

Write $y_i = f(x_i)$

$$\begin{aligned} a + bx_1 &= y_1 \\ a + bx_2 &= y_2 \\ \vdots &= \vdots \\ a + bx_m &= y_m \end{aligned}$$

Overdetermined linear system

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix}$$

a linear system $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A}$ an $m \times 2$ rectangular matrix.
Since more rows than columns $\rightarrow$ **overdetermined system** — unless $\mathbf{b}$ is exceptional, there is no solution to such a system

Minimum residual solution

there is no solution to such a system $\rightarrow$ what is the best we can do?

exact solution $\implies \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{0}$

so let's find $\mathbf{x}$ that minimizes residual $\mathbf{r}$ — but in what norm?

can do more statistics (inference) if we choose the 2-norm

$\rightarrow$ minimize $\| \mathbf{r} \|_2$

[other choices of norm give other forms of fitting]

Least squares problem

Definition

the **least squares solution** to $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $m > n = \text{rank}(\mathbf{A})$ is the solution $\mathbf{x}$ that minimizes $\| \mathbf{r} \|_2$

In our case, $r_i = y_i - a - bx_i, \quad i = 1 \cdots m$
so to minimize $\| \mathbf{r} \|_2$ means to minimize
$\| \mathbf{r} \|_2^2 = \mathbf{r}^T\mathbf{r} = \sum_i (y_i - a - bx_i)^2 \equiv S$ the sum of squared residuals
hence the name — method of least squares
The necessary conditions for a stationary point (min/max) are:

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0$$

Finding minimum of S

$$\frac{\partial S}{\partial a} = 0 \implies -2\sum_i (y_i - a - bx_i) = 0 \implies \sum y_i = ma + b\sum x_i$$

$$\frac{\partial S}{\partial b} = 0 \implies -2\sum_i x_i(y_i - a - bx_i) = 0 \Rightarrow \sum x_i y_i = a\sum x_i + b\sum x_i y_i$$

which gives a 2×2 matrix equation for a, b

$$\begin{bmatrix} m & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

The normal equations

In terms of the overdetermined system, $\mathbf{Ax} = \mathbf{b}$, this square system is

the **normal equations**

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$$

$\rightarrow$ the simplest way to solve any overdetermined system is to solve the normal equations

Another derivation

In $\mathbf{Ax} = \mathbf{b}$, $\mathbf{b}$ is a vector in $\mathbb{R}^m$; the columns of $\mathbf{A} = [a_1 a_2]$ are each vectors in $\mathbb{R}^m$.
Any linear combination of those columns can be written
$x_1 a_1 + x_2 a_2 = \mathbf{Ax}$: these form a 2D subspace of $\mathbb{R}^m$.
We want to find a vector in this subspace (i.e. all vectors of the form $\mathbf{Ax}$) as close to $\mathbf{b}$ as possible. How?
To minimize $\| \mathbf{r} \|_2$, we make $\mathbf{r} \perp$ the subspace i.e. $\mathbf{r} \perp$ the columns of $\mathbf{A}$.
$\implies a_1^T \mathbf{r} = 0$; $a_2^T \mathbf{r} = 0$ or $\mathbf{A}^T \mathbf{r} = 0$

$$\mathbf{A}^T \mathbf{r} = \mathbf{A}^T(\mathbf{b} - \mathbf{Ax}) = \mathbf{A}^T \mathbf{b} - \mathbf{A}^T \mathbf{Ax} = \mathbf{0}$$

i.e. the normal equations
$\implies$ $\mathbf{Ax}$ is the projection of $\mathbf{b}$ onto the subspace formed by columns of $\mathbf{A}$

Are the normal equations invertible?

Theorem

If $\mathbf{A}$ *has rank 2, then* $\mathbf{A}^T\mathbf{A}$ *is square and nonsingular*

Proof:

The pseudoinverse

So if $\mathbf{A}$ has rank 2, we can write

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

which is also written

$$\mathbf{x} = \mathbf{A}^{\dagger}\mathbf{b}$$

where $\mathbf{A}^{\dagger} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the **pseudoinverse** of $\mathbf{A}$
Of course, don't compute the pseudoinverse;
just solve the normal equations

An example

Example

Linear models

This can all be generalized to **any linear model**
i.e. fit to $y = x_1\phi_1(t) + x_2\phi_2(t) + \cdots + x_n\phi_n(t)$ from data points $\{X_i, Y_i\}, i = 1 \cdots m, m > n$

- form the overdetermined system $\mathbf{Ax} = \mathbf{b}$

$$\mathbf{A} = \begin{bmatrix} \phi_1(X_1) & \phi_2(X_1) & \cdots & \phi_n(X_1) \\ \phi_1(X_2) & \phi_2(X_2) & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ \phi_1(X_m) & \phi_2(X_m) & \cdots & \phi_n(X_m) \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}$$

- form the normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ as before

Since $\mathbf{A}_{ij} = \phi_j(X_i)$, $(\mathbf{A}^T\mathbf{A})_{ij} = \sum_k \phi_i(X_k)\phi_j(X_k)$
then, if $rank(\mathbf{A}) = \min(m, n) = n$ ($\mathbf{A}$ is **full rank**), normal equations have a unique solution

◁ **Example:** polynomial curve fitting

Properties of normal equations

But $\mathbf{A}^T\mathbf{A}$ **is symmetric and positive definite**

Proof:

$\implies$ solve using Cholesky factorization $\rightarrow$ takes $\approx n^3/6$ ops.
also must form $\mathbf{A}^T\mathbf{A}$, an $n \times n$ (symmetric) matrix with $n^2/2$ different entries, each one $\sum_k \phi_i(X_k)\phi_j(X_k)$ i.e. m multiplies
$\rightarrow \frac{1}{2}n^2(m + \frac{1}{3}n)$ operations

Problems with the normal equations

BUT solving the normal equations by Cholesky is NOT the recommended way to find the least squares solution - **WHY NOT?**

1. if $\mathbf{A}$ is 'close to singular', can get $\mathbf{A}^T\mathbf{A}$ singular

2. forming normal equations CAN worsen the conditioning (sensitivity) of least squares problem

3. if $\mathbf{A}$ is rank-deficient then $\mathbf{A}^T\mathbf{A}$ is singular $\implies$ can't solve normal equations
(Cholesky factors are singular so can't solve by back-substitution)

The matrix 2-norm

◁ **Example:**

The 2-norm is the natural norm for LSQ problems (minimizing $\| \mathbf{r} \|_2$) $\Longrightarrow$ can no longer avoid the matrix 2-norm :

for a square matrix $\mathbf{A}$ (see 'MatrixNorms' for proof)

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$$

$\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ is the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$
(all eigenvalues are positive since $\mathbf{A}^T\mathbf{A}$ is positive definite).

Singular value decomposition SVD

It is easier to characterize the condition number in the 2-norm in terms of the singular values of $\mathbf{A}$. To do that we need the

Definition

A $m \times n$ real matrix $\mathbf{A}$ has the **singular value decomposition**

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where

1. $\mathbf{U}$ is $m \times m$ orthogonal matrix
2. Σ is a diagonal $m \times n$ real matrix
3. $\mathbf{V}$ is $n \times n$ orthogonal matrix

The non-negative diagonal entries $\{\sigma_k \geq 0\}$ in Σ are called the singular values of $\mathbf{A}$.
In our case, where $m > n$, there are n positive singular values $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_n > 0$, if $\mathbf{A}$ is of full rank.

End of Lecture 15