Finishing up least squares

Take $A \in \mathbb{R}^{m \times n}$, $m \geq n$. If $m > n$, the problem is overconstrained, i.e. we have more equations than unknowns.

Least squares.

$$\min_{x \in \mathbb{R}^n} \| A\underline{x} - \underline{b} \|_2^2 .$$

There are multiple ways to derive the normal equations. One approach is to define the function:

$$f(x) = \| A\underline{x} - \underline{b} \|_2^2 .$$

At a minimum x^* of f,

$$\nabla f(\underline{x}^*) = 0 .$$

What is f?

$$f(x) = \underline{x}^T A^T A \underline{x} - 2\underline{x}^T A^T \underline{b} + \underline{b}^T \underline{b}$$

$$= \underline{x}^T C \underline{x} - 2\underline{x}^T \underline{d} + \underline{b}^T \underline{b} .$$

$$C = A^T A , \quad \underline{d} = A^T \underline{b} .$$

let's write out $\underline{x}^T C \underline{x}$ in component form:

First, $(C\underline{x})_k = \sum_{i=1}^{m} C_{ki}\, x_i$.

Then $\underline{x}^T C \underline{x} = \sum_{k=1}^{m} x_k (C\underline{x})_k$

$$= \sum_{k=1}^{m} \sum_{i=1}^{m} x_k C_{ki} x_i .$$

Now, we compute the contribution of this term to the jth component of ∇f.

$$\frac{\partial}{\partial x_j}\left(\underline{x}^T C \underline{x}\right) = \frac{\partial}{\partial x_j}\left(\sum_{k=1}^{m} \sum_{i=1}^{m} x_k C_{ki} x_i\right).$$

$$= \sum_{k=1}^{m} \sum_{i=1}^{m} \left(\frac{\partial x_k}{\partial x_j} C_{ki} x_i + x_k C_{ki} \frac{\partial x_i}{\partial x_j} \right)$$

$$= \sum_{k=1}^{m} \sum_{i=1}^{m} \left(\delta_{kj} C_{ki} x_i + \delta_{ij} C_{ki} x_k \right).$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{m} \delta_{kj} C_{ki} x_i + \sum_{k=1}^{m} \sum_{i=1}^{m} \delta_{ij} C_{ki} x_k$$

$$= \sum_{i=1}^{m} C_{ji} x_i + \sum_{k=1}^{m} C_{kj} x_k$$

Note that $C = A^T A$, so C is symmetric $\Rightarrow C_{ij} = C_{ji}$. Then we can write:

$$\frac{\partial}{\partial x_j}\left(\underline{x}^T C \underline{x}\right) = 2 \sum_{i=1}^{n} C_{ji} x_i$$

$$= 2\left(A^T A \underline{x}\right)_j$$

Similarly, you can show:

$$\frac{\partial}{\partial x_j}\left(2\underline{x}^T \underline{d}\right) = 2 d_j$$

$$= 2\left(A^T \underline{b}\right)_j$$

and trivially $\frac{\partial}{\partial x_j}\left(b^T b\right) = 0$.

So: $$\frac{\partial f}{\partial x_j} = 2\left(A^T A \underline{x}\right)_j - 2\left(A^T \underline{b}\right)_j$$

implying that $\nabla f = A^T A \underline{x} - A^T \underline{b}$.

So then $\nabla f(x^*) = 0$ is equivalent to

$$A^T A \underline{x}^* = A^T \underline{b}.$$

— These are called the normal equations.

Here is another way to derive the normal equations via the fundamental theorem of linear Alg.

Thm: (Fundamental Theorem of Linear Algebra) (FTLA)

Let $A \in \mathbb{R}^{n \times m}$. Then.

$$\text{Null}(A) \oplus \text{Ran}(A^T) = \mathbb{R}^m$$
$$\text{Null}(A^T) \oplus \text{Ran}(A) = \mathbb{R}^n.$$

Note the notation $V \oplus W = Z$, means

That for each $z \in Z$,

$\exists\ v \in V,\ w \in W,\ v^T w = 0$ and

$$z = v + w.$$

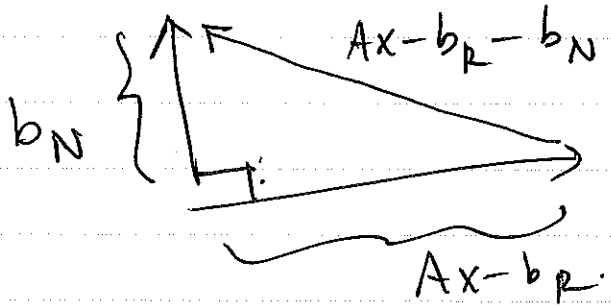
↝

Idea: The least squares problem is $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$.

Since $b \in \mathbb{R}^m$, by FTLA, $\exists\, b_R \in \text{Ran}(A)$ and $b_N \in \text{Null}(A^T)$ so that $b = b_R + b_N$.

So, $\|Ax - b\|_2^2 = \|Ax - b_R - b_N\|_2^2$

Note that $Ax - b_R \in \text{Ran}(A)$ and $b_N \in \text{Null}(A^T)$, so $Ax - b_R \perp b_N$.



By Pythagorean theorem:

$$\|Ax - b_R - b_N\|_2^2 = \|Ax - b_R\|_2^2 + \|b_N\|_2^2.$$

In other words, the minimization can be rewritten:

$$\min_{x \in \mathbb{R}^n} \| Ax - b \|_2^2 = \min_{x \in \mathbb{R}^n} \| Ax - b_R \|_2^2,$$

since the term $\| b_N \|_2^2$ doesn't depend on $x \in \mathbb{R}^n$. Note that by construction, $\exists x^*$ so that $b_R = Ax^*$. This will be a solution to the minimization problem since:

$$Ax^* - b_R = 0.$$

The normal equations are derived by multiplying by A^T:

$$A^T A x^* - A^T b_R = 0,$$

and also using $A^T b_N = 0$ by construction:

$$A^T A x^* - A^T b_R - A^T b_N = 0$$

$$\Rightarrow A^T A x^* - A^T b = 0.$$

2 >

Geometric picture from the normal equations:

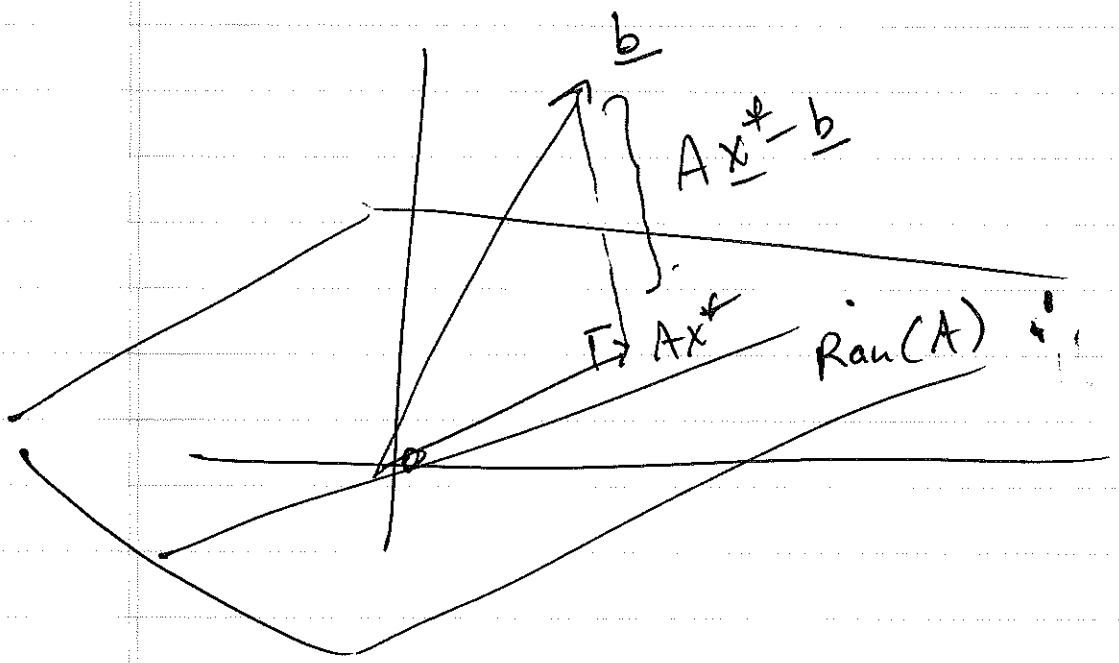
Take $y \in \mathbb{R}^n$, and look at its inner product with the normal equations:

$$y^T (A^T A x^* - A^T b) = 0$$

$$\Leftrightarrow \quad y^T A^T (A x^* - b) = 0$$

$$\Leftrightarrow \quad (Ay)^T (A x^* - b) = 0$$

Since Ay is an arbitrary vector in the $\mathrm{Ran}(A)$, this says that the residual vector $Ax^* - b$ is orthogonal to the $\mathrm{Ran}(A)$



Example: Setting up normal equations for least squares problem.

Let's say we are given a set of data

$$(1, 1) = (x_1, y_1)$$
$$(4, 1) = (x_2, y_2)$$
$$(9, 2) = (x_3, y_3).$$

And we want to fit a function to this data of the form

$$y = \alpha x^{1/2} + \beta.$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \approx \begin{bmatrix} x_1^{1/2} & 1 \\ x_2^{1/2} & 1 \\ x_3^{1/2} & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

$$= \underline{b} \qquad = \underline{A} \qquad = \underline{x}.$$

$$\underline{b} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}.$$

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2\sqrt{2} & 1 \end{bmatrix}.$$

Normal equations are

$$A^T A \underline{x} = A^T b$$

$$A^T A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2\sqrt{2} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3+2\sqrt{2} & 3 \\ 13 & 3+2\sqrt{2} \end{bmatrix}.$$

$$A^T b = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 3+2\sqrt{2} \end{bmatrix}$$

Then $\underline{x}^*$ satisfying $\nabla f(x^*) = 0$ solves the system:

$$\begin{bmatrix} 3+2\sqrt{2} & 3 \\ 13 & 3+2\sqrt{2} \end{bmatrix} \underline{x}^* = \begin{bmatrix} 4 \\ 3+2\sqrt{2} \end{bmatrix}$$

Why use the QR Factorization?

Because if A is ill-conditioned, A^TA is even <u>more</u> ill-conditioned.

Suppose $A = \begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix}$.

$\|A\|_2 = 1$, $\|A^{-1}\|_2 = \frac{1}{\varepsilon}$,

so $k_2(A) = \frac{1}{\varepsilon}$.

If we look at $A^TA = \begin{pmatrix} \varepsilon^2 & 0 \\ 0 & 1 \end{pmatrix}$,

$\|A^TA\|_2 = 1$, $\|(A^TA)^{-1}\| = \frac{1}{\varepsilon^2}$,

so $k_2(A^TA) = \frac{1}{\varepsilon^2} >> \frac{1}{\varepsilon} = k_2(A)$.

Thm 2.12: S'pose $A \in \mathbb{R}^{m \times n}$, $m \geq n$.

Then A can be written in the form

$$A = \hat{Q}\,\hat{R},$$

where $\hat{R}$ is upper triangular, $n \times n$ matrix, and $\hat{Q}$ is an $m \times n$ which satisfies

$$\hat{Q}^T \hat{Q} = I_n = n \times n \text{ identity.}$$

If $\text{rank}(A) = n$, then $\hat{R}$ is nonsingular.

$\square$

We can use the QR Factorization to stably compute a solution to the least squares problem.

Thm 2.13 $A \in \mathbb{R}^{m \times n}$, $m \geq n$, $\text{rank}(A) = n$. $b \in \mathbb{R}^m$.

Then there exists a unique least squares solution $x^* \in \mathbb{R}^n$ which minimizes the function $y \longmapsto \|Ay - b\|_2$ over $y \in \mathbb{R}^n$.

x^* can be computed by solving

$$\hat{R} x^* = \hat{Q}^T b.$$

Proof: Note that $\hat{Q} \in \mathbb{R}^{m \times n}$, and

$$\hat{Q}^T \hat{Q} = I_n = n \times n \text{ identity.}$$

If $m = n$, then

$$\begin{aligned} x^* &= A^{-1} b \\ &= (\hat{Q}\hat{R})^{-1} b \\ &= \hat{R}^{-1} \hat{Q}^{-1} b \\ &= \hat{R}^{-1} \hat{Q}^T b \end{aligned}$$

implying that $\hat{R} x^* = \hat{Q}^T b$.

Take $m > n$: For $b \in \mathbb{R}^m$, we apply The Fundamental Thm of Linear Alg.

$$b = b_R + b_N, \quad b_R \in \text{Ran}(\hat{Q})$$

$$b_N \in \text{Null}(\hat{Q}^T).$$

Take x^* to be the solution of $\hat{R}x^* = \hat{Q}^T b$.

$$Ay - b = \hat{Q}\hat{R}y - b.$$

$$= \hat{Q}\hat{R}(y - x^*) + \hat{Q}\hat{R}x^* - b$$
$$= \hat{Q}\hat{R}(y - x^*) + \hat{Q}\hat{Q}^T b - b$$
$$= \hat{Q}\hat{R}(y - x^*) + \hat{Q}\hat{Q}^T(b_R + b_N) - b_R - b_N$$
$$= \hat{Q}\hat{R}(y - x^*) + \hat{Q}\hat{Q}^T b_R - b_R - b_N$$
$$= \hat{Q}\hat{R}(y - x^*) + \hat{Q}\hat{Q}^T\hat{Q}c - b_R - b_N$$

(For some $c \in \mathbb{R}^n$)

$$= \hat{Q}\hat{R}(y - x^*) + \hat{Q}c - b_R - b_N$$
$$= \hat{Q}\hat{R}(y - x^*) - b_N.$$

Now, you can show:

$$\|Ay - b\|_2^2 = \|\hat{R}(y - x^*)\|_2^2 + \|b_N\|_2^2$$

$$\geq \|b_N\|_2^2.$$

$\|Ay - b\|_2^2$ is minimized when $\hat{R}(y - x^*) = 0 \iff y = x^*$, since $\hat{R}$ is nonsingular. Thus, x^* taken as $\hat{R}x^* = \hat{Q}^T b$ solves the least squares problem.

Remark: Why is it better to solve

$\hat{R}x = \hat{Q}^T b$, instead of the original normal equations $A^T A x = A^T b$. ?

Typically, $\kappa_2(\hat{R}) << \kappa_2(A^T A)$.

In particular, with $A = \hat{Q}\hat{R}$,

$$\begin{aligned} A^T A &= (\hat{Q}\hat{R})^T \hat{Q}\hat{R} \\ &= \hat{R}^T \hat{Q}^T \hat{Q} \hat{R} \\ &= \hat{R}^T \hat{R}, \end{aligned}$$

Showing that $\kappa_2(A) = \kappa_2(\hat{R})$.
In particular, the conditioning of the system $\hat{R}x = \hat{Q}^T b$ is the same as the matrix A.

Eigenvalues and Eigenvectors

Definition 5.1: Suppose $A \in \mathbb{R}^{n \times n}$.

If $\lambda \in \mathbb{C}$ satisfies

$$A x = \lambda x$$

and $\underline{x \neq 0}$, then λ is called an eigenvalue of A. x is the associated eigenvector.

Remark: Why do we care about eigenvalues and eigenvectors?

Consider a general dynamical system

$$\frac{dx}{dt} = Ax, \quad A \in \mathbb{R}^{n \times n}$$

$$x(0) = x_0.$$

A solution can be rewritten in the form:

$$x(t) = x(0) \underbrace{\exp(At)}_{\text{matrix exponential}}.$$

If we can "diagonalize" A, $A = \mathcal{Q} D \mathcal{Q}^{-1}$,

here $\mathcal{Q}$ is a matrix with eigenvectors in the columns, D is a diagonal matrix with the eigenvalues on the diagonal,

then the system of ODEs can be rewritten

$$\frac{dx}{dt} = \mathcal{Q} D \mathcal{Q}^{-1} x$$

$$\Leftrightarrow \quad \frac{d\,\mathcal{Q}^{-1}x}{dt} = D\,\mathcal{Q}^{-1}x.$$

Defining $y = \mathcal{Q}^{-1}x$, we obtain a new, decoupled set of ODEs:

$$\frac{dy}{dt} = Dy, \quad \underline{\text{or}}:$$

$$\frac{dy_i}{dt} = \lambda_i y_i,$$

and the eigenvalues tell us which components of y grow or decay.

Gerschgorin Thms

Definition 5.3 $A \in \mathbb{C}^{n\times n}$, $n \geq 2$., $A = (a_{ij})$

D_i = Gerschgorin discs

$$= \{ z \in \mathbb{C} : |z - a_{ii}| \leq R_i \}.$$

with $$R_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|.$$

Thm 5.4 $A \in \mathbb{C}^{n\times n}$, $n \geq 2$. All of the eigenvalues of A are in $D = \bigcup_{i=1} D_i$,

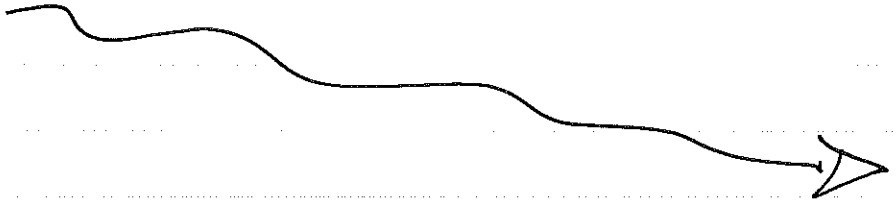
D_i = Gerschgorin disc.

Pf: let (x, λ) be an eigenpair. by definition

$$\sum_{j=1}^{n} a_{ij} x_j = \lambda x_i, \quad i = 1, \ldots, n.$$

Take x_k to be the component with largest absolute value: $|x_j| \leq |x_k|, j = 1, \ldots, n.$

Then



$$|\lambda - a_{kk}|\,|x_k| = |\lambda x_k - a_{kk} x_k|$$

$$= \left| \sum_{j=1}^{n} a_{kj} x_j - a_{kk} x_k \right|$$

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

$$\leq |x_k| \, R_k .$$

This shows that $|\lambda - a_{kk}| \leq R_k$, i.e. λ is in D_k.

Fact: The eigenvalues of a matrix A are roots of the characteristic polynomial $p(\lambda) = \det(A - \lambda I)$.

Fact: There is a one-to-one correspondence between eigenvalue problems and root-finding problems.

i.e. If $p(z) = z^m + a_{m-1} z^{m-1} + \cdots + a_1 z + a_0$,

then the roots of p are actually the eigenvalues of the matrix

$$\begin{bmatrix} 0 & & & -a_0 \\ 1 & \ddots & & \vdots \\ & \ddots & 0 & -a_{m-2} \\ & & 1 & -a_{m-1} \end{bmatrix}$$

The above matrix is called the companion matrix for p.

Thm (Abel 1824).
For any $m \geq 5$, there is a polynomial $p(z)$ of degree m with rational coeff's that has a real root with the property that it cannot be written in any expression involving rational numbers, $+$, $-$, $*$, $\div$, and k^{th} roots.

Further issue: determining roots of a characteristic polynomial, or more generally any polynomial, is ill-conditioned.

Example: Consider the 2×2 identity matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

writing down the characteristic polynomial, we get $p(\lambda) = (\lambda-1)^2 = \lambda^2 - 2\lambda + 1$.

If we perturb the constant coeff in this polynomial by 10^{-4}:

$$\tilde{p}(\lambda) = \lambda^2 - 2\lambda + 0.9999,$$

the roots of this perturbed poly are:

$$\tilde{p}(\lambda) = (\lambda - 0.99)(\lambda - 1.01),$$

i.e. they are perturbed by 10^{-2} from the original roots of $p(\lambda)!!!$

Power method for computing eigenvalues and eigenvectors.

Let's assume $A \in \mathbb{R}^{m \times m}$ is a real, symmetric matrix.

Definition: (Rayleigh Quotient).

$$r(x) = \frac{x^T A x}{x^T x}.$$

Note that if x is an eigenvector of A, then $r(x) = \lambda$ is its corresponding eigenvalue.

Via taking derivatives, we can see that $\nabla r(x) = \frac{2}{x^T x}(Ax - r(x)x)$.

So, if x is an eigenvector of A, then $\nabla r(x) = 0$. And, if $\nabla r(x) = 0$ with $x \neq 0$, then x is an eigenvector of A, with eigenvalue $r(x)$.

Remark: let q_J be an eigenvector of A. we know that $\nabla r(q_J) = 0$.

By Taylor expansion, we have:

$$r(x) - r(q_J) = \nabla r(q_J)(x - q_J) + (x - q_J)^T \nabla^2 r(\xi)(x - q_J)$$

for some ξ on the line between x and q_J.

This shows that

$$|r(x) - r(q_J)| \leq C \|x - q_J\|^2,$$

since $\nabla r(q_J) = 0$.

The idea is if we have a good estimate for an ~~eigenvalue~~ eigenvector, we can get a quadratically accurate estimate for the corresponding eigenvalue. (ii).

Let $v^{(0)}$ with $\|v^{(0)}\| = 1$ be some initial guess for an eigenvector of A. The power method, or power iteration, is defined as:

---

for $k = 1, 2, \ldots$

$$w = A v^{(k-1)} \qquad \text{apply } A$$
$$v^{(k)} = w / \|w\| \qquad \text{normalize}$$
$$\lambda^{(k)} = (v^{(k)})^T A v^{(k)} \qquad \text{Rayleigh quotient.}$$

The claim is that $\lambda^{(k)}$ converges to an eigenvalue of A and $v^{(k)}$ converges to the corresponding eigenvector, as $k \to \infty$. This iteration should converge to the largest eigenvalue in absolute value.

Thm: Suppose $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_m| \geq 0$

and $q_1^T v^{(0)} \neq 0$. Then the power iteration above creates iterates satisfying

$$\|v^{(k)} - (\pm q_1)\| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right), \quad |\lambda^{(k)} - \lambda_1| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right).$$