Lecture 10: Least Squares Problems

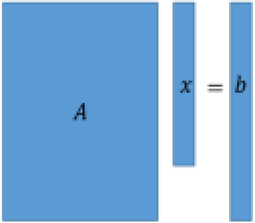
June 18, 2025

Outline

1. Least Squares
2. Method 1: Normal Equations
3. Method 2: QR Factorization
    1. QR for Least Squares

Least Squares

- This lecture discusses solving problems with more equations (rows) than variables.
- The problem is solved “as well as possible” since the system is over-determined (more equations than necessary).
- That is, least squares problems solve the equation $Ax = b$, where A is taller that it is wide.
- We end up with a solution that is over-determined.



Least Squares

- Least squares problems were first posed and formulated by Gauss around 1795 (though published first by Legendre 1805).
- The method of least squares is often found in applications, e.g., finding a line or polynomial to fit a large set of data/observations.
- Mathematically, we want to minimize the magnitude of the **residual vector** $r = b - Ax$.

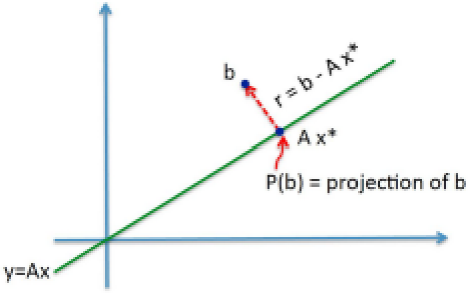
$$\min_{x \in \mathbb{R}^n} ||b - Ax||_2^2, \text{ for } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, m \geq n.$$

- In general, we can not achieve $r = 0$.
- Notice this differs from minimizing (square, non-singular) linear systems, for which we use:

$$F(x) = \frac{1}{2}x^T Ax - x^T b.$$

Least Squares

- A geometric interpretation of least squares problems is as follows.
- We find the closest point to b on the $y = Ax$ hyperplane.
- In other words, find the "projection" of b onto the range of A.
- Notice that residual vector r is orthogonal to y (see figure below).



Least Squares

Theorem 1

Let $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, m \geq n$*, and A full rank. A vector* $x \in \mathbb{R}^n$ *minimizes*

$$||r||_2^2 = ||b - Ax||_2^2$$

if and only if $r \perp range(A)$.

Task: Collin to add the proof of Theorem 1 into the Lecture Notes. Theorem 1 implies

$$\begin{aligned} r^T A &= \vec{0} \\ \Leftrightarrow A^T r &= \vec{0} \\ \Leftrightarrow A^T(b - Ax) &= \vec{0} \\ \Leftrightarrow A^T b &= A^T Ax. \end{aligned}$$

Least Squares

High Level View of Where We Are Going

- By end of Lecture 10: existence of Q, computed from A.
- By end of Lecture 11: existence of R, computed from A, Q.
- Not proved in Slides / Notes yet: uniqueness of Q, R, given A.
  - A Brief Note From Wikipedia: If A has full rank, then the QR factorization is unique, provided we require the diagonal elements of R to be positive.

Least Squares

- The equations $A^T Ax = A^T b$ are known as the **normal equations**.
- Solving the normal equations (which is a square system) gives the **least squares solution**.
- This motivates the definition of the pseudo-inverse.

Definition 1.1

$A^+ = (A^T A)^{-1} A^T$ *is called the (Moore-Penrose)* **pseudo-inverse** *of A.*

- The least squares solution satisfies

$$x = A^+ b = (A^T A)^{-1} A^T b.$$

Least Squares

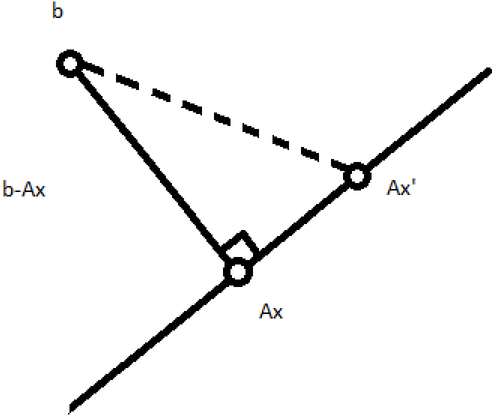
Fact: Any perturbation of x from this solution yields a higher residual norm. To see this let $x' = x + e$, where x is the least squares solution and e is some perturbation. Then,

$$
\begin{aligned}
& \|b - Ax'\|_2^2 \\
= \;& (b - Ax')^T (b - Ax'), \\
= \;& (b - Ax - Ae)^T (b - Ax - Ae), \\
= \;& (b - Ax)^T (b - Ax) - 2(Ae)^T (b - Ax) + (Ae)^T (Ae), \\
= \;& (b - Ax)^T (b - Ax) - 2e^T A^T (b - Ax) + \|Ae\|_2^2, \\
= \;& \|b - Ax\|_2^2 + \|Ae\|_2^2 - 2e^T \cancelto{0}{(A^T b - A^T A x)}, \quad (x \text{ is the LS soln})
\end{aligned}
$$

$\Rightarrow \|b - Ax'\|_2^2 > \|b - Ax\|_2^2$ for any $e \neq 0$.

Least Squares

Thus, any other point x' yields a larger residual, as seen geometrically below.



Least Squares

We will consider the following two solution strategies (for now):

1. Normal equations: Find and solve normal equations $A^T Ax = A^T b$ to find x (e.g., via Cholesky).
    1. It is proved in the Lecture Notes that, if A has full rank, then $A^T A$ is SPD, so that it has a Cholesky factor.
2. QR Factorization: Construct a factorization $A = QR$ (with certain properties) and instead solve $Rx = Q^T b$ for x.

Method 1: Normal Equations

- In this subsection we will look at the normal equations solution.
- In the next subsection will discuss QR factorizations.
- We solve $A^T Ax = A^T b$ directly by computing the Cholesky factorization $A^T A = GG^T$, with G lower triangular.
- Then, we compute x by forward/backward solves.
- The complexity of this approach has flops to form $A^T A \approx mn^2$ and $GG^T \approx \frac{1}{3}n^3$.
- Therefore, the total flops $\approx mn^2 + \frac{1}{3}n^3$.

Method 1: Normal Equations

Consider the application of polynomial fitting with least squares. We want to find a polynomial of the form:

$$p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_{n-1} t^{n-1},$$

that best fits the set of 2D points given by (t_i, y_i) for $i = 1, \ldots, m$, with $m > n$. Each data point yields one equation. The coefficients $a_0, a_1, \ldots, a_{n-1}$ are the unknowns. The matrix problem is

$$\begin{bmatrix} 1 & t_{\color{red}1} & t_1^2 & \cdots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^{n-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Method 1: Normal Equations

As a concrete example we are given the set of (t_i, y_i) points $\{(0,0),(0,-1),(2,1),(2,0),(4,2),(4,1)\}$. We want to find the best fit line $y = a_0 + a_1 t$ using normal equations. We have,

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 1 & 2 \\ 1 & 4 \\ 1 & 4 \end{bmatrix}, \qquad x = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \\ 2 \\ 1 \end{bmatrix}.$$

Method 1: Normal Equations

To obtain the solution we construct

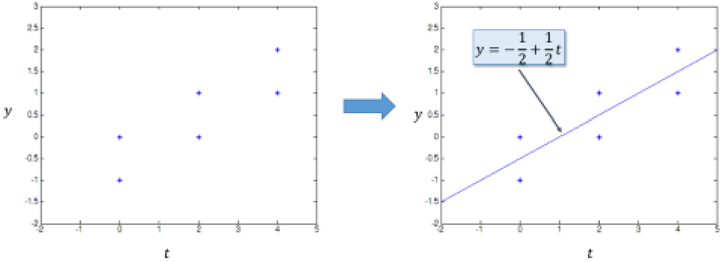
$$A^T A = \begin{bmatrix} 6 & 12 \\ 12 & 40 \end{bmatrix}, \quad \text{and } A^T b = \begin{bmatrix} 3 \\ 14 \end{bmatrix}.$$

Solving $A^T A x = A^T b$ gives

$$\begin{aligned} a_0 &= -\frac{1}{2}, \\ a_1 &= \frac{1}{2}, \text{ so} \\ p &= -\frac{1}{2} + \frac{1}{2} t. \end{aligned}$$

Method 1: Normal Equations

The figure below shows the points and the line of best fit given by p.



Method 2: QR Factorization

- The previous subsection discussed the first method for solving least squares problems, i.e., via the normal equations.
- This lecture discusses a second approach using QR factorization.
- The QR factorization decomposes a matrix A into an orthogonal matrix Q and a triangular matrix R.

Method 2: QR Factorization

- Some properties of **orthogonal matrices** are discussed next.
- They are needed to give a theorem for the existence of a QR factorization at the end of this section.

Definition 3.1

A square matrix Q is **orthogonal** *if $Q^{-1} = Q^T$ (i.e., $Q^T Q = QQ^T = I$).*

Theorem 2

If Q is orthogonal, then $\|Qx\|_2 = \|x\|_2$.

Proof.

$\|Qx\|^2 = (Qx)^T(Qx) = x^T Q^T Qx = x^T x = \|x\|^2$. □

Remark: Permutation matrices, first seen during matrix re-ordering, are examples of orthogonal matrices.

Method 2: QR Factorization

Note that left multiplication by an orthogonal Q corresponds to

$$\begin{cases} \text{rotation if } \det(Q) = 1, \\ \text{reflection if } \det(Q) = -1. \end{cases}$$

Definition 3.2

A set of vectors are **orthonormal** *if they are mutually orthogonal and each vector has norm* $= 1$.

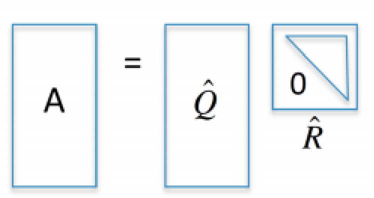
- For example, the columns of an orthogonal matrix are orthonormal.
- The columns of an $n \times n$ orthogonal matrix, Q, form an orthonormal basis of $\mathbb{R}^n$.
- Be careful not to confuse the following:
    1. orthogonal **vectors** need not be unit length,
    2. an orthogonal **matrix** has columns that are **orthonormal**.

Method 2: QR Factorization

The following theorem gives the existence of a QR factorization.

Theorem 3

Suppose $A \in \mathbb{R}^{m\times n}$ has full rank. Then there exists a unique matrix $\hat{Q} \in \mathbb{R}^{m\times n}$ satisfying $\hat{Q}^T\hat{Q} = I$ (i.e., with orthonormal columns) and a unique upper triangular matrix $\hat{R} \in \mathbb{R}^{n\times n}$ with positive diagonals $(r_{i,i} > 0)$ such that $A = \hat{Q}\hat{R}$.



Note that because $\hat{Q}$ is non-square in general here, it is **not** necessarily an orthogonal matrix.

Method 2: QR Factorization

Example: Let

$$\hat{Q} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{bmatrix},$$

so that

$$\hat{Q}^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

Then

$$\begin{aligned} \hat{Q}\hat{Q}^T &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ &\neq I_3. \end{aligned}$$

Method 2: QR Factorization - QR for Least Squares

Consider the least squares problem:

$$\min_{x} \|Ax - b\|^2.$$

We will try to make $\|Ax - b\|^2$ as small as possible by re-expressing $Ax - b$ in terms of the QR factorization, and adjusting x.
Only x can be adjusted because A and b are defined (given) by the problem.

Method 2: QR Factorization - QR for Least Squares

Consider separating $Ax - b$ into orthogonal components using $\hat{Q}\hat{R}$. That is, using $\hat{Q}$ and $\hat{R}$, split $Ax - b$ into two orthogonal components:

$$\begin{aligned} Ax - b &= \hat{Q}\hat{R}x - b, \\ &= \hat{Q}\hat{R}x - (\underbrace{\hat{Q}\hat{Q}^T - \hat{Q}\hat{Q}^T}_{=0} + I)b, \\ &= \underbrace{\hat{Q}(\hat{R}x - \hat{Q}^T b)} - \underbrace{(I - \hat{Q}\hat{Q}^T)b}. \end{aligned}$$

We claim that these two vectors are orthogonal.

Method 2: QR Factorization - QR for Least Squares

These vectors are orthogonal if and only if their inner product is zero. We can verify that the inner product is zero as follows:

$$\begin{aligned}
& \left[\hat{Q}(\hat{R}x - \hat{Q}^T b)\right]^T \left[(I - \hat{Q}\hat{Q}^T)b\right] \\
= \;& (\hat{R}x - \hat{Q}^T b)^{T} \hat{Q}^T (I - \hat{Q}\hat{Q}^T) b, \\
= \;& (\hat{R}x - \hat{Q}^T b)^{T} (\hat{Q}^T - \underbrace{\hat{Q}^T \hat{Q}}_{=I} \hat{Q}^T) b, \quad (\hat{Q}\text{'s columns orthonormal}) \\
= \;& (\hat{R}x - \hat{Q}^T b)^{T} (\underbrace{\hat{Q}^T - \hat{Q}^T}_{=0}) b = 0.
\end{aligned}$$

Note that since $\hat{Q}$ is not square, $\hat{Q}\hat{Q}^T \neq I$, but $\hat{Q}^T\hat{Q} = I$.

Method 2: QR Factorization - QR for Least Squares

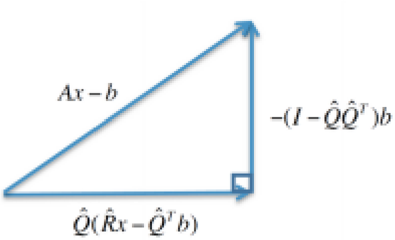
The goal of the least squares problem is to minimize the size of $r = b - Ax$. We can only modify x to make the vector $r = b - Ax$ as short as possible. By Pythagoras we have

$$\begin{aligned} \|Ax - b\|^2 &= \|\hat{Q}(\hat{R}x - \hat{Q}^T b)\|^2 + \|(I - \hat{Q}\hat{Q}^T)b\|^2, \\ &= \underbrace{\|(\hat{R}x - \hat{Q}^T b)\|^2}_{\text{select } x \text{ to minimize}} + \underbrace{\|(I - \hat{Q}\hat{Q}^T)b\|^2}_{\text{can't adjust}}. \end{aligned}$$

Method 2: QR Factorization - QR for Least Squares

The orthogonal components can be visualized as shown in the figure below. The norm is minimized when the first term is 0. So the least squares solution is

$$\hat{R}x = \hat{Q}^T b \qquad \Rightarrow \qquad x = \hat{R}^{-1}\hat{Q}^T b.$$



Method 2: QR Factorization - QR for Least Squares

We can also relate this solution to pseudoinverse and normal equations as follows.

Pseudoinverse

The pseudoinverse written in terms of QR factors is

$$\begin{aligned}
A^{+} &= (A^T A)^{-1} A^T \\
&= ((\hat{Q}\hat{R})^T(\hat{Q}\hat{R}))^{-1}(\hat{Q}\hat{R})^T \\
&= (\hat{R}^T \underbrace{\hat{Q}^T \hat{Q}}_{=I} \hat{R})^{-1}(\hat{R}^T \hat{Q}^T) \\
&= (\hat{R}^T \hat{R})^{-1} \hat{R}^T \hat{Q}^T \\
&= \hat{R}^{-1} \underbrace{(\hat{R}^T)^{-1} \hat{R}^T}_{=I} \hat{Q}^T \\
&= \hat{R}^{-1} \hat{Q}^T.
\end{aligned}$$

Method 2: QR Factorization - QR for Least Squares

Normal Equations

Then consider the normal equations

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

Method 2: QR Factorization - QR for Least Squares

Two Sizes of QR Factorization

- So far we have only seen the **reduced ("economy size")** version of QR factorization.
- Specifically, $A = \hat{Q}\hat{R}$ where $\hat{Q} \in \mathbb{R}^{m \times n}$ and $\hat{R} \in \mathbb{R}^{n \times n}$.
- The "full" version of QR adds extra orthonormal columns to make Q square (and thus makes Q a true orthogonal matrix).
- Extra zero rows in R are also added to match the dimensions (See below).

Method 2: QR Factorization - QR for Least Squares

A full QR factorization is achieved by appending $m - n$ additional orthonormal columns to Q. First define

$$\hat{Q}_{m-n} \equiv \begin{bmatrix} q_{n+1} & q_{n+2} & \cdots & q_m \end{bmatrix}.$$

Then we have

$$\begin{bmatrix} A \end{bmatrix}_{m\times n} = \begin{bmatrix} \hat{Q} | \hat{Q}_{m-n} \end{bmatrix}_{m\times m} \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}_{m\times n},$$

which is often useful for theoretical purposes.

Method 2: QR Factorization - QR for Least Squares

Computing the (reduced) QR Factorization

- To compute the reduced QR factorization we let

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix},$$

  where a_i are the columns of A.
- The columns span the column space of the matrix.
- So we want to find a set of orthonormal column vectors, $\{q_i\}$ spanning the same space.
- That is, $\text{span}\{q_1, q_2, \dots, q_j\} = \text{span}\{a_1, a_2, \dots, a_j\}$, for $j = 1, \dots, n$.

Method 2: QR Factorization - QR for Least Squares

- For this we can use Gram-Schmidt orthogonalization.
- We already saw a variant of Gram-Schmidt earlier for constructing A-orthogonal search directions in conjugate gradient.
- The same general idea is used here:
    - use columns of A as proposed vectors to be orthogonalized into Q,
    - build each new vector q_j by orthogonalizing a_j with respect to all previous q vectors, $\{q_1, q_2, \ldots, q_{j-1}\}$, and then normalize q_j.

Method 2: QR Factorization - QR for Least Squares

The entries for R can be calculated once we know Q by considering the general form

$$\begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ q_1 & q_2 & \dots & q_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{bmatrix}.$$

Written out fully, this is

$$\begin{aligned} a_1 &= r_{11}q_1, \\ a_2 &= r_{12}q_1 + r_{22}q_2, \\ a_3 &= r_{13}q_1 + r_{23}q_2 + r_{33}q_3, \\ &\vdots \\ a_n &= r_{1n}q_1 + r_{2n}q_2 + \cdots + r_{nn}q_n. \end{aligned}$$

The next lecture will discuss this approach (using Gram-Schmidt) of computing the QR factorization in more detail.