LINEAR ALGEBRA. LECTURE 7

Projection matrices and least squares

Projections

Last lecture, we learned that $P = A(A^TA)^{-1}A^T$ is the matrix that projects a vector $\mathbf{b}$ onto the space spanned by the columns of A. If $\mathbf{b}$ is perpendicular to the column space, then it's in the left nullspace $N(A^T)$ of A and $P\mathbf{b} = \mathbf{0}$. If $\mathbf{b}$ is in the column space then $\mathbf{b} = A\mathbf{x}$ for some $\mathbf{x}$, and $P\mathbf{b} = \mathbf{b}$.

A typical vector will have a component $\mathbf{p}$ in the column space and a component $\mathbf{e}$ perpendicular to the column space (in the left nullspace); its projection is just the component in the column space.

The matrix projecting $\mathbf{b}$ onto $N(A^T)$ is $I - P$:

$$\begin{aligned} \mathbf{e} &= \mathbf{b} - \mathbf{p} \\ \mathbf{e} &= (I - P)\mathbf{b}. \end{aligned}$$

Naturally, $I - P$ has all the properties of a projection matrix.

Least squares

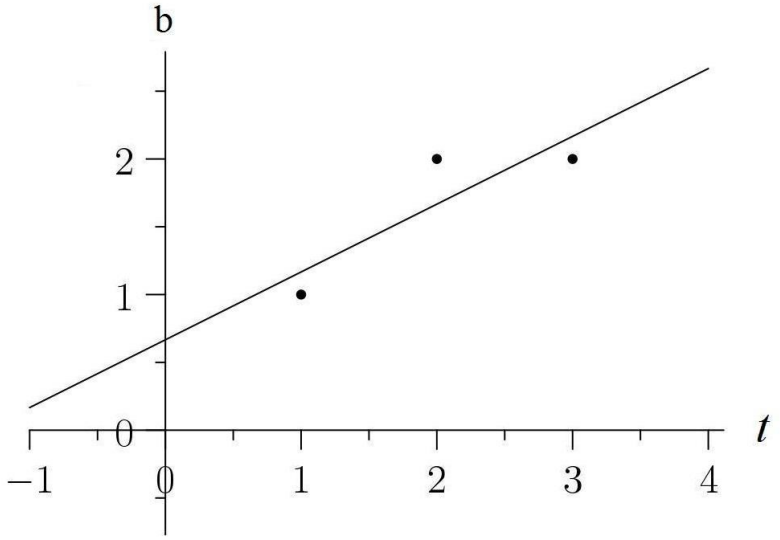

Figure 1: Three points and a line close to them.

We want to find the closest line $b = C + Dt$ to the points $(1,1)$, $(2,2)$, and $(3,2)$. The process we're going to use is called *linear regression*; this technique is most useful if none of the data points are *outliers*.

By "closest" line we mean one that minimizes the error represented by the distance from the points to the line. We measure that error by adding up the squares of these distances. In other words, we want to minimize $||A\mathbf{x} - \mathbf{b}||^2 = ||\mathbf{e}||^2$.

If the line went through all three points, we'd have:

$$\begin{aligned} C + D &= 1 \\ C + 2D &= 2 \\ C + 3D &= 2, \end{aligned}$$

but this system is unsolvable. It's equivalent to $A\mathbf{x} = \mathbf{b}$, where:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} C \\ D \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}.$$

There are two ways of viewing this. In the space of the line we're trying to find, e_1, e_2 and e_3 are the vertical distances from the data points to the line. The components p_1, p_2 and p_3 are the values of $C + Dt$ near each data point; $\mathbf{p} \approx \mathbf{b}$.

In the other view we have a vector $\mathbf{b}$ in $\mathbb{R}^3$, its projection $\mathbf{p}$ onto the column space of A, and its projection $\mathbf{e}$ onto $N(A^T)$.

We will now find $\hat{\mathbf{x}} = \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix}$ and $\mathbf{p}$. We know:

$$\begin{aligned} A^T A\hat{\mathbf{x}} &= A^T\mathbf{b} \\ \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix} &= \begin{bmatrix} 5 \\ 11 \end{bmatrix}. \end{aligned}$$

From this we get the *normal equations*:

$$\begin{aligned} 3\hat{C} + 6\hat{D} &= 5 \\ 6\hat{C} + 14\hat{D} &= 11. \end{aligned}$$

We solve these to find $\hat{D} = 1/2$ and $\hat{C} = 2/3$.

We could also have used calculus to find the minimum of the following function of two variables:

$$e_1^2 + e_2^2 + e_3^2 = (C + D - 1)^2 + (C + 2D - 2)^2 + (C + 3D - 2)^2.$$

Either way, we end up solving a system of linear equations to find that the closest line to our points is $b = \frac{2}{3} + \frac{1}{2}t$.

This gives us:

i	p_i	e_i
1	7/6	−1/6
2	5/3	1/3
3	13/6	−1/6

or $\mathbf{p} = \begin{bmatrix} 7/6 \\ 5/3 \\ 13/6 \end{bmatrix}$ and $\mathbf{e} = \begin{bmatrix} -1/6 \\ 2/6 \\ -1/6 \end{bmatrix}$. Note that $\mathbf{p}$ and $\mathbf{e}$ are orthogonal, and also that $\mathbf{e}$ is perpendicular to the columns of A.

The matrix $A^T A$

We've been assuming that the matrix $A^T A$ is invertible. Is this justified?

If A has independent columns, then $A^T A$ is invertible.

To prove this we assume that $A^T A\mathbf{x} = \mathbf{0}$, then show that it must be true that $\mathbf{x} = \mathbf{0}$:

$$\begin{aligned} A^T A\mathbf{x} &= \mathbf{0} \\ \mathbf{x}^T A^T A\mathbf{x} &= \mathbf{x}^T\mathbf{0} \\ (A\mathbf{x})^T(A\mathbf{x}) &= \mathbf{0} \\ A\mathbf{x} &= \mathbf{0}. \end{aligned}$$

Since A has independent columns, $A\mathbf{x} = \mathbf{0}$ only when $\mathbf{x} = \mathbf{0}$.

As long as the columns of A are independent, we can use linear regression to find approximate solutions to unsolvable systems of linear equations. The columns of A are guaranteed to be independent if they are *orthonormal*, i.e. if they are perpendicular unit vectors like $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, or like $\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$ and $\begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$.

Orthogonal matrices and Gram-Schmidt

In this lecture we finish introducing orthogonality. Using an orthonormal basis or a matrix with orthonormal columns makes calculations much easier. The Gram-Schmidt process starts with any basis and produces an orthonormal basis that spans the same space as the original basis.

Orthonormal vectors

The vectors $\mathbf{q}_1, \mathbf{q}_2, ...\mathbf{q}_n$ are *orthonormal* if:

$$\mathbf{q}_i^T \mathbf{q}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$$

In other words, they all have (normal) length 1 and are perpendicular (ortho) to each other. Orthonormal vectors are always independent.

Orthonormal matrix

If the columns of $Q = \begin{bmatrix} \mathbf{q}_1 & ... & \mathbf{q}_n \end{bmatrix}$ are orthonormal, then $Q^T Q = I$ is the identity.

Matrices with orthonormal columns are a new class of important matrices to add to those on our list: triangular, diagonal, permutation, symmetric, reduced row echelon, and projection matrices. We'll call them "orthonormal matrices".

A square orthonormal matrix Q is called an *orthogonal matrix*. If Q is square, then $Q^T Q = I$ tells us that $Q^T = Q^{-1}$.

For example, if $Q = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ then $Q^T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$. Both Q and Q^T are orthogonal matrices, and their product is the identity.

The matrix $Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ is orthogonal. The matrix $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ is not, but we can adjust that matrix to get the orthogonal matrix $Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

We can use the same tactic to find some larger orthogonal matrices called *Hadamard matrices*:

$$Q = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

An example of a rectangular matrix with orthonormal columns is:

$$Q = \frac{1}{3} \begin{bmatrix} 1 & -2 \\ 2 & -1 \\ 2 & 2 \end{bmatrix}.$$

We can extend this to a (square) orthogonal matrix:

$$\frac{1}{3}\begin{bmatrix} 1 & -2 & 2 \\ 2 & -1 & -2 \\ 2 & 2 & 1 \end{bmatrix}.$$

These examples are particularly nice because they don't include complicated square roots.

Orthonormal columns are good

Suppose Q has orthonormal columns. The matrix that projects onto the column space of Q is:

$$P = Q(Q^TQ)^{-1}Q^T.$$

If the columns of Q are orthonormal, then $Q^TQ = I$ and $P = QQ^T$. If Q is square, then $P = I$ because the columns of Q span the entire space.

Many equations become trivial when using a matrix with orthonormal columns. If our basis is orthonormal, the projection component $\hat{x}_i$ is just $\mathbf{q}_i^T\mathbf{b}$ because $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$ becomes $\hat{\mathbf{x}} = Q^T\mathbf{b}$.

Gram-Schmidt

With elimination, our goal was "make the matrix triangular". Now our goal is "make the matrix orthonormal".

We start with two independent vectors $\mathbf{a}$ and $\mathbf{b}$ and want to find orthonormal vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ that span the same plane. We start by finding orthogonal vectors $\mathbf{A}$ and $\mathbf{B}$ that span the same space as $\mathbf{a}$ and $\mathbf{b}$. Then the unit vectors $\mathbf{q}_1 = \frac{\mathbf{A}}{||\mathbf{A}||}$ and $\mathbf{q}_2 = \frac{\mathbf{B}}{||\mathbf{B}||}$ form the desired orthonormal basis.

Let $\mathbf{A} = \mathbf{a}$. We get a vector orthogonal to $\mathbf{A}$ in the space spanned by $\mathbf{a}$ and $\mathbf{b}$ by projecting $\mathbf{b}$ onto $\mathbf{a}$ and letting $\mathbf{B} = \mathbf{b} - \mathbf{p}$. ($\mathbf{B}$ is what we previously called $\mathbf{e}$.)

$$\mathbf{B} = \mathbf{b} - \frac{\mathbf{A}^T\mathbf{b}}{\mathbf{A}^T\mathbf{A}}\mathbf{A}.$$

If we multiply both sides of this equation by $\mathbf{A}^T$, we see that $\mathbf{A}^T\mathbf{B} = 0$.

What if we had started with three independent vectors, $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$? Then we'd find a vector $\mathbf{C}$ orthogonal to both $\mathbf{A}$ and $\mathbf{B}$ by subtracting from $\mathbf{c}$ its components in the $\mathbf{A}$ and $\mathbf{B}$ directions:

$$\mathbf{C} = \mathbf{c} - \frac{\mathbf{A}^T\mathbf{c}}{\mathbf{A}^T\mathbf{A}}\mathbf{A} - \frac{\mathbf{B}^T\mathbf{c}}{\mathbf{B}^T\mathbf{B}}\mathbf{B}.$$

For example, suppose $\mathbf{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$. Then $\mathbf{A} = \mathbf{a}$ and:

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} - \frac{\mathbf{A}^T\mathbf{b}}{\mathbf{A}^T\mathbf{A}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} - \frac{3}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}. \end{aligned}$$

Normalizing, we get:

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{3} & 0 \\ 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{2} \end{bmatrix}.$$

The column space of Q is the plane spanned by $\mathbf{a}$ and $\mathbf{b}$.

When we studied elimination, we wrote the process in terms of matrices and found $A = LU$. A similar equation $A = QR$ relates our starting matrix A to the result Q of the Gram-Schmidt process. Where L was lower triangular, R is upper triangular.

We started with a matrix A, whose columns were a, b, c. We ended with a matrix Q, whose columns are q_1, q_2, q_3. What is the relation between those matrices? The matrices A and Q are m by n when the n vectors are in m-dimensional space, and there has to be a third matrix that connects them.

The idea is to write the a's as combinations of the q's. The vector b is a combination of the orthonormal q_1 and q_2, and we know what combination it is:

$$b = (q_1^{\mathrm{T}}b)q_1 + (q_2^{\mathrm{T}}b)q_2.$$

Every vector in the plane is the sum of its q_1 and q_2 components. Similarly c is the sum of its q_1, q_2, q_3 components: $c = (q_1^{\mathrm{T}}c)q_1 + (q_2^{\mathrm{T}}c)q_2 + (q_3^{\mathrm{T}}c)q_3$. If we express that in matrix form we have ***the new factorization*** $A = QR$:

$$\textbf{QR factors} \qquad A = \begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix} \begin{bmatrix} q_1^{\mathrm{T}}a & q_1^{\mathrm{T}}b & q_1^{\mathrm{T}}c \\ & q_2^{\mathrm{T}}b & q_2^{\mathrm{T}}c \\ & & q_3^{\mathrm{T}}c \end{bmatrix} = QR$$

Notice the zeros in the last matrix! R is *upper triangular* because of the way Gram-Schmidt was done. The first vectors a and q_1 fell on the same line. Then q_1, q_2 were in the same plane as a, b. The third vectors c and q_3 were not involved until step 3.