Orthogonal projections and Least Squares

A. Eremenko

September 7, 2024

In this section, all vector spaces are real and of finite dimension. Let U be a subspace of a vector space V.

For every $\mathbf{x} \in V$ we define the *projection* of $\mathbf{x}$ onto U as the vector $\mathbf{y} \in U$ which minimizes the distance $\|\mathbf{x} - \mathbf{y}\|$.

Theorem 1. *In a space of finite dimension, a projection of a vector* $\mathbf{x}$ *onto a subspace* U *exists and is unique.*

This theorem is proved using Calculus.

Proof. Let $\delta = \inf_{\mathbf{u} \in U} \|\mathbf{x} - \mathbf{u}\|$. This means that there is a sequence $\mathbf{u}_n \in U$ such that

$$\|\mathbf{x} - \mathbf{u}_n\| \to \delta. \tag{1}$$

Then the distance from $\mathbf{x}$ to the middle of the segment $[\mathbf{u}_m, \mathbf{u}_n]$ satisfies

$$\|\mathbf{x} - (\mathbf{u}_m + \mathbf{u}_n)/2\| \leq \frac{1}{2}\|\mathbf{x} - \mathbf{u}_m| + \frac{1}{2}\|\mathbf{x} - \mathbf{u}_n\|,$$

by the triangle inequality. On the other hand,

$$\|\mathbf{x} - (\mathbf{u}_m + \mathbf{u}_n)/2\| \geq \delta$$

by definition of δ. Therefore

$$\limsup_{m,n,\to\infty} \|\mathbf{x} - (\mathbf{u}_m + \mathbf{u}_n)/2\| = \delta.$$

Now we use the polarization identity from the previous lecture:

$$\|\mathbf{u}_n - \mathbf{u}_m\|^2 = 2\|\mathbf{x} - \mathbf{u}_m\|^2 + 2\|\mathbf{x} - \mathbf{u}_n\|^2 - 4\|\mathbf{x} - (\mathbf{u}_m + \mathbf{u}_n)/2\|^2,$$

and conclude that the LHS tends to 0. Therefore $(\mathbf{u}_n)$ is a Cauchy sequence and must have a limit u. This limit belongs to U since U is a closed set, and by passing to the limit in (1) we obtain

$$\|\mathbf{x} - \mathbf{u}\| = \delta.$$

Now we prove the uniqueness. Suppose that some $\mathbf{x}$ has two projections $\mathbf{w}_1, \mathbf{w}_2$ in U. Then the triangle $(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2)$ is a plane isosceles triangle with equal sides $[\mathbf{x}, \mathbf{w}_1]$ and $[\mathbf{x}, \mathbf{w}_2]$, and the height of such triangle (which is $\|\mathbf{x} - (\mathbf{w}_1 + \mathbf{w}_2)/2\|$ is evidently shorter then the sides. This shows that we must have $\mathbf{w}_1 = \mathbf{w}_2$.

Our goal is to find an explicit formula for projections. From this formula the theorem can be also derived.

Proposition 1. *If* $\mathbf{y}$ *is the projection of* $\mathbf{x}$ *on* U *then* $\mathbf{y} \in U$ *and* $(\mathbf{x}-\mathbf{y}, \mathbf{u}) = 0$ *for all* $\mathbf{u} \in U$*.*

Proof. Let $\mathbf{y}$ be the projection of $\mathbf{x}$ on U, and consider the distance from $\mathbf{x}$ to $\mathbf{y} - t\mathbf{u}$, where $\mathbf{u}$ is an arbitrary vector in U and t is a real number. Since $\mathbf{y}$, by definition, is the closest to $\mathbf{x}$ vector in U, and $\mathbf{y} - t\mathbf{u}$ is in U, the function

$$\|\mathbf{x} - (\mathbf{y} - t\mathbf{u})\|^2 = \|(\mathbf{x} - \mathbf{y}) + t\mathbf{u}\|^2 = (\mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y}) + 2t(\mathbf{x} - \mathbf{y}, \mathbf{u}) + t^2(\mathbf{u}, \mathbf{u}),$$

considered as a function of t must attain the minimum for $t = 0$. This function is a quadratic polynomial, and the minimum is attained at $t = 0$ if and only if $(\mathbf{x} - \mathbf{y}, \mathbf{u}) = 0$, as claimed.

This proposition allows us to derive the formula for the projection. We choose to describe U as the column space of some $n \times m$ matrix A. (The dimension of our space is n). Our condition that $(\mathbf{x} - \mathbf{y}, \mathbf{u}) = 0$ for all $\mathbf{u} \in U$ translates to

$$\mathbf{x} - \mathbf{y} \in U^{\perp}, \tag{2}$$

by definition of $U^{\perp}$. We have $U = C(A)$, and in the previous lecture we described the orthogonal complement of the column space: it is the null space of A^T. So (2) is equivalent to

$$A^T(\mathbf{x} - \mathbf{y}) = 0.$$

now we have $\mathbf{y} \in U = C(A)$ which means that $\mathbf{y}$ is a linear combination of columns of A, in other words, $\mathbf{y} = A\mathbf{w}$ for some $w \in \mathbf{R}^m$. Thus

$$A^T\mathbf{x} = A^T A\mathbf{w}, \quad \text{for some} \quad \mathbf{w} \in \mathbf{R}^m. \tag{3}$$

Since the projection exists, this equation must have a solution $\mathbf{w}$. Using this solution, we obtain the formula for the projection

$$\mathbf{y} = A\mathbf{w}. \tag{4}$$

So the recipe to find the projection $\mathbf{y}$ of $\mathbf{x}$ onto the column space of A is to solve (3) and plug it to (4).

The formula simplifies we choose A smartly, namely *with linearly independent columns*, that is $r(A) = m$.

Proposition 2. *If the columns of A are linearly independent then $A^T A$ is invertible.*

So for this case, (3) and (4) give this projection formula

$$\mathbf{y} = A(A^T A)^{-1} A^T \mathbf{x}. \tag{5}$$

One immediate conclusion is that *projection is a linear operator*, and the matrix of this operator is

$$A(A^T A)^{-1} A^T,$$

where A is a matrix whose columns make a basis of the subspace onto which we project.

Proof of Proposition 2. If A is $n \times m$ then $A^T A$ is $m \times m$, One necessary and sufficient condition for invertibility of a square matrix its nullspace is trivial. So we want to prove that the equation $A^T A\mathbf{x} = \mathbf{0}$ has only trivial solution. Let $\mathbf{y} = A\mathbf{x}$. Then $\mathbf{y} \in C(A)$ and also $\mathbf{y} \in N(A^T)$. But these two spaces are orthogonal complements of each other (see the previous lecture) therefore $\mathbf{y} = 0$. Finally $\mathbf{x} = 0$ since columns of A are linearly independent.

Example. Projection on a line. Let this line be spanned by some vector $\mathbf{a}$. In our terminology, this line is the column space of the $n \times 1$ matrix $\mathbf{a}$. Then by (5) the projection $\mathbf{y}$ of $\mathbf{x}$ is

$$\mathbf{y} = \mathbf{a}(\mathbf{a}^T\mathbf{a})^{-1}\mathbf{a}^T\mathbf{x}.$$

Now $\mathbf{a}^T\mathbf{a} = \|\mathbf{a}\|^2$ and $\mathbf{a}^T\mathbf{x}$ is $(\mathbf{a}, \mathbf{x})$, so

$$\mathbf{y} = \mathbf{a}\frac{(\mathbf{a}, \mathbf{x})}{\|\mathbf{a}\|^2}$$

the formula which is probably familiar to you; it simplifies when $\mathbf{a}$ is a unit vector, $\|\mathbf{a}\| = 1$. Using $(\mathbf{a}, \mathbf{x}) = \|\mathbf{a}\|\|\mathbf{x}\| \cos\alpha$, where α is the angle between $\mathbf{x}$ and $\mathbf{a}$ we obtain

$$\mathbf{y} = \mathbf{a}(\mathbf{a}, x) = \mathbf{a}\|\mathbf{x}\| \cos\alpha, \quad \text{when} \quad \|\mathbf{a}\| = 1.$$

Exercise. Compare our general formula for the projection matrix with the special case that we earlier derived for projection on a line in $\mathbf{R}^2$ and show that they give the same result.

Exercise. Show that every projection matrix $P = A(AA^T)^{-1}A^T$ is a) orthogonal, ($P^T = P$) and b) satisfies $P^2 = I$, as we checked earlier for projections on lines in the plane.

Actually these two properties characterize projections.

In the previous lecture, we discussed orthogonal complement, and proved that for every subspace U of a vector space V equipped with an inner product, and every vector $\mathbf{x} \in V$, we have the orthogonal decomposition

$$\mathbf{x} = \mathbf{u} + \mathbf{w}, \quad \text{where} \quad \mathbf{u} \in U \quad \text{and} \quad \mathbf{w} \in U^\perp.$$

It is easy to see that this $\mathbf{u}$ is nothing but projection of $\mathbf{x}$ onto U. Indeed, $\mathbf{u} \in U$ and $\mathbf{w} = \mathbf{x} - \mathbf{u}$ is orthogonal to every vector in U by definition of $U^\perp$. So the criterion in Theorem 1 is satisfied. Similarly, $\mathbf{w}$ is the projection of $\mathbf{x}$ onto $U^\perp$.

Method of Least Squares. Suppose we have a system

$$A\mathbf{x} = \mathbf{y}$$

with $m \times n$ matrix A which has no solution, the most interesting case is that $m > n$ and $r(A) = n$, so A has linearly independent columns. In practical applications the absence of solutions may be due to the errors in the experimental data. Then we may want to find "the best guess" for $\mathbf{x}$, based on these data. A typical example is a set of points (t_k, y_k), $1 \leq k \leq m$, in the plane, which are supposed to lie on a line. But they do not lie on a line (because y_k are the results of some measurements which are prone to various errors). We want to find a line $y = at + b$ which fits the experimental data best. So we want to find a and b. For them we have a system of m equations:

$$at_k + b = y_k, \quad 1 \le k \le m$$

with matrix

$$A = \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \dots & \dots \\ t_m & 1 \end{pmatrix}.$$

If all t_k are distinct, which is natural to assume, then the columns of A are linearly independent. *The Least Square Solution* is, by definition such $\hat{\mathbf{x}}$ that $A\hat{\mathbf{x}}$ is the projection of $\mathbf{y}$ onto the columns space of A. In other words, it minimizes the distance

$$\|A\mathbf{x} - \mathbf{y}\|$$

instead of solving the system $A\mathbf{x} = \mathbf{y}$ exactly.

According to the previous theory, it is obtained in the following way:

Compute $A^T A$ and $A^T\mathbf{y}$, and solve $A^T A\hat{\mathbf{x}} = A^T\mathbf{y}$. This always has a solution $\hat{\mathbf{x}}$. If columns of A are linearly independent this solution is

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{y}.$$

In the case of fitting a line $at + b$ this system $A^T A\mathbf{x} = \mathbf{y}$ is 2×2.

Remark. In practice, solving $A^T A\hat{x} = A^T\mathbf{b}$ by row operations is easier than inverting $A^T A$.