



Linear Algebra unit 4

Date: @today

Lane: *Linear Algebra - statistics*

Topic: Inner Product, Orthogonality, Projection Operators, Linear Regression as a projection operator.

What it covers:

- So far we've looked at *underdetermined* systems of equations. The same concepts from linear algebra can be applied to *overdetermined* systems.
- Inner product test of orthogonality. Cosine law, vector length, orthogonal complement, (direct sum), projections. linear regression operator.

Requires:

Definition of inner product and Vector spaces unit 1, linear independence, matrix operations.

Required by, used by:

Introduction to the linear regression as a projection operation for an overdetermined system.

Independence and Orthogonality

Both linear algebra and probability use the term "independence" and related concepts. How are these related? In linear algebra independence refers to perpendicular vectors, indicated by a zero inner product: $\langle u, v \rangle = 0$. Independent events in probability do not need to be conditioned when forming joint probabilities. It is always true that $P(A, B) = P(A | B)P(B)$. In the special case of independence, **B** does not condition **A** or vice versa: $P(A, B) = P(A)P(B)$. This is not to be confused with *mutual exclusivity* between events that entails probabilistic *dependence*.

There isn't a direct relationship between orthogonal vectors and probabilistically independent events. They are different things. However both concepts will be useful in different ways to explain the relationships among variables in regression.

The same holds for *linear independence* and probabilistic independence. First we need to clarify how orthogonality and linear independence are related.

Inner products

As mentioned two vectors whose inner product equals 0 are equivalently *orthogonal*, *perpendicular*, or *at right angles to each other*. These terms all mean the same. Orthogonality of a set of vectors implies linear independence. This is easy to see by taking the inner product (assuming we are in an *inner product space*) with one vector in the set over the linear combination:

$$\begin{aligned} &v_1(a_1v_1) + v_1(a_2v_2) + \dots v_1(a_nv_n) = 0 \\ &\langle v_1, v_1 \rangle a_1 + \langle v_1, v_2 \rangle a_2 + \dots \langle v_1, v_n \rangle a_n = 0 \end{aligned}$$

where by orthogonality, only the first term is non-zero. So working with orthogonal vectors simplifies linear combinations by making it possible to work with them one term at a time. Conversely linear independence does not imply orthogonality—most often any set of vectors “overlap” with each other. What about other cases? In general the inner product gives us the angle between vectors. When the vectors are aligned, lengths multiply. Thus $\langle v, v \rangle = \|v\|^2$. Otherwise its the cosine of the normed vectors:

$$\cos \theta = \langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \rangle = \frac{\langle u, v \rangle}{\|u\| \|v\|}$$

This can be derived from conventional trigonometric identities.

Orthogonal complements



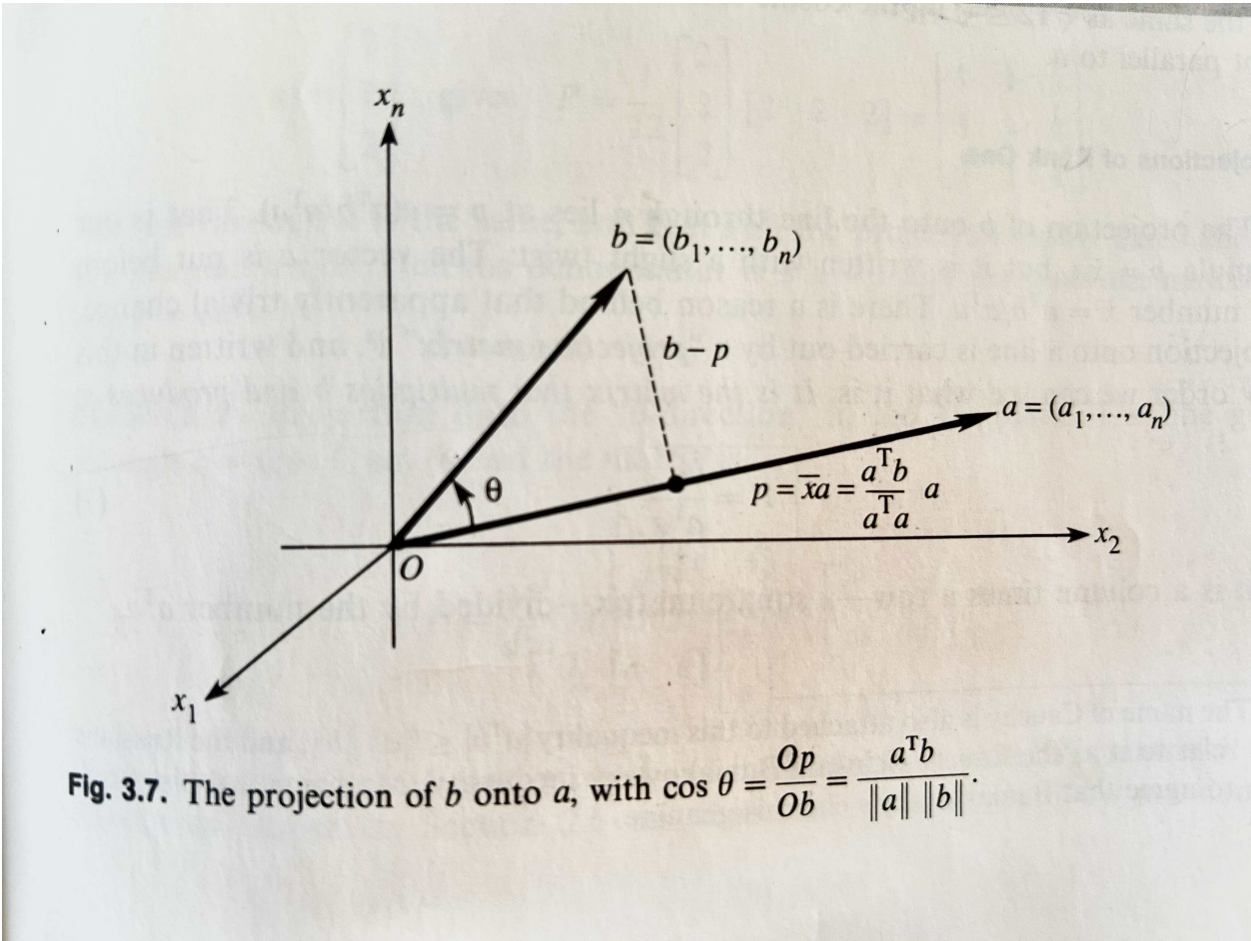
The *orthogonal complement* of vectors v in a subspace \mathbf{S} is a subspace of vectors orthogonal to \mathbf{S} .

$$S^\perp = \{s : \langle s, v \rangle = 0\}$$

We’ve already seen examples of these, such as the row space and null space. S^\perp partitions the vector space into two mutually exclusive subspaces. The vector space V is the *direct sum* of S and its complement. This is written: $V = S \oplus S^\perp$.

Projections

To project the vector b on the vector a (e.g. the line formed by a): $p = \frac{\langle a, b \rangle}{\langle a, a \rangle} a$. If a is unit length, the the projection is just a rescaling of the length of a by the cosine of the included angle. Hence the projection cannot be longer than the original vector: $\|p\| \leq \|b\|$.



Strang, “Linear Algebra and its Applications 3rd Ed.” p. 147

Projection on a set of orthogonal vectors

The general problem is to project a line, b , onto a subspace, where it is not contained. This decomposes nicely if the subspace is defined by a set of orthonormal vectors.

Each component is computed separately, which add to form the complete projection.

Linear regression

When n is greater than p the system of linear equations has more equations than unknowns and there will not be a solution vector that solves the system. Linear regression takes the approach to minimize how far off the regression value is from a solution. We can picture that as a projection of the desired value - the b vector into the column space of the A matrix. We define the error term (or the "residual") as the difference that cannot be placed in A 's column space: $b - Ax$. So this "projects away" the error term. Geometrically the error is minimized when the error vector is orthogonal to the columns of A . Pre-multiplying by the transpose must therefore map into zero (e.g. into the left-nullspace of A): $A^T(b - Ax) = 0$.

This is trivial when A is square and full rank. Then the error term will be zero and b has a solution. In the overdetermined case the column space is insufficient to map into b .

The general projection operator

From the orthogonality of the error and A 's column space we get $A^Tb = A^T Ax$. So $(A^T A)^{-1} A^T b = \bar{x}$. And since $p = A\bar{x}$:

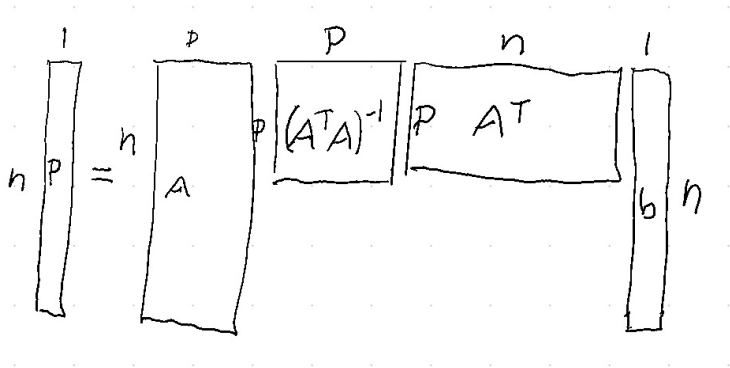
$$P = A(A^T A)^{-1} A^T$$

So that the projection p of b on the subspace is $p = Pb$, where the column space of P is the subspace we are projecting onto, P puts b into A 's column space. Note the projection "solution" p is a linear function of b . The term $(A^T A)^{-1}$ is called the Gram Matrix.

💡 If A 's columns are linearly independent (even if its a tall matrix), then (unlike A) the "Gram matrix" $A^T A$ is square, symmetric and full rank.

In general, any matrix that is a projection operator satisfies $P = P^2, P = P^T$. This is the definition of a projection operator. It is only interesting if it's row space is not full rank — e.g. it is it's null space that it "projects away." Projection operators divide the vector space into orthogonal complements. It satisfies the condition that $V = P \oplus (I - P)$.

To see how this works, consider each step in the P operator. Starting from right to left: A^T "squeezes" b into A 's column space, then the Gram matrix turns this into a set of coefficients \bar{x} that combine the data in the columns of A to form "points" in A 's column spac



See the notebook "lin_reg1.ipynb" for an example of linear regression in one dimension.

Other derivations of Linear Regression

1. **By optimization** Conventionally Linear Regression is defined by minimizing the quadratic error term.
2. **As probabilistic inference.** As a the probability of a predicted value conditional on the data.

We will cover these in upcoming lectures.

References

Read Strang, ch 3

Read Boyd Applied Lin Alg. Section 3.4 (We will do Chapter 12 later.)


Here are some videos:


Khan Academy - Linear Algebra projection operators.

Introduction to projections | Matrix transformations | Linear Algebra | Khan Academy

Determining the projection of a vector on a line

Watch the next lesson: <https://www.khanacademy.org/math/linear->

 <https://www.youtube.com/watch?v=27vT-NWuw0M>



Projections Introduction


Here's G. Stang's lecture on projections:

15. Projections onto Subspaces

MIT 18.06 Linear Algebra, Spring 2005

Instructor: Gilbert Strang

View the complete course: <http://ocw.mit.edu/18-06S05>

 https://www.youtube.com/watch?v=Y_Ac6KiQ1t0

