Proposition

Let A be a matrix. Then $\mathbf{N}(A^T A) = \mathbf{N}(A)$.

Proof.

If $A\mathbf{x} = \mathbf{0}$, then of course $A^T A\mathbf{x} = \mathbf{0}$. Conversely, if $A^T A\mathbf{x} = \mathbf{0}$, then

$$0 = \mathbf{x} \cdot (A^T A\mathbf{x}) = \mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T A\mathbf{x} = (A\mathbf{x}) \cdot (A\mathbf{x}) \ ,$$

so $A\mathbf{x} = \mathbf{0}$ also. □

Corollary

If the columns of the matrix A are independent, then $A^T A$ is invertible.

Projection onto a subspace

V subspace of $\mathbb{R}^m$: For each $\mathbf{b}$ in $\mathbb{R}^m$, want to find the *orthogonal projection* $\mathbf{p}$ of $\mathbf{b}$ onto V, i.e., closest point of V to $\mathbf{b}$.

Function $\mathbf{b} \mapsto \mathbf{p}$ is linear transformation, so there is an $m \times m$ matrix P for which $P\mathbf{b} = \mathbf{p}$.

We know P takes points in V to themselves and points in $V^\perp$ to $\mathbf{0}$. Now we can build a basis for $\mathbb{R}^m$ by combining a basis for V and one for $V^\perp$; if we can find a matrix that takes the ones in V to themselves and the ones in $V^\perp$ to $\mathbf{0}$, then it must be P.

So make a matrix A ($m \times n$) out of the basis for V and another B ($m \times (m-n)$) out of the basis for $V^\perp$. Then

$$\begin{bmatrix} A & B \end{bmatrix} \overset{A^T}{\rightarrow} \begin{bmatrix} A^TA & O_{n\times(m-n)} \end{bmatrix} \overset{(A^TA)^{-1}}{\rightarrow} \begin{bmatrix} I_n & O_{n\times(m-n)} \end{bmatrix} \overset{A}{\rightarrow} \begin{bmatrix} A & O_{m\times(m-n)} \end{bmatrix}.$$

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

Notation:

- $\mathbf{e} = \mathbf{b} - \mathbf{p}$ — orthogonal to V
- $\widehat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$ — coefficients for writing $\mathbf{p}$ as a combination of the columns of A.

Example

Find the orthogonal projection $\mathbf{p}$ of $\mathbf{b} = (2, 0, 3)$ onto the V with basis $(1, 1, 1)$ and $(1, 0, -1)$, and find the projection matrix P, and check $\mathbf{e} = \mathbf{b} - \mathbf{p} \perp V$:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{bmatrix}, \qquad (A^T A)^{-1} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix}$$

$$\begin{aligned} P &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 5/6 & 1/3 & -1/6 \\ 1/3 & 1/3 & 1/3 \\ -1/6 & 1/3 & 5/6 \end{bmatrix}, \end{aligned}$$

Example (ctnd)

$$\begin{aligned}
\mathbf{p} &= P\mathbf{b} \\
&= \begin{bmatrix} 5/6 & 1/3 & -1/6 \\ 1/3 & 1/3 & 1/3 \\ -1/6 & 1/3 & 5/6 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 7/6 \\ 5/3 \\ 13/6 \end{bmatrix} \\
A^T(\mathbf{b} - \mathbf{p}) &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 5/6 \\ -5/3 \\ 5/6 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{aligned}$$

Special Case (silly)

A is invertible, i.e., $V = \mathbb{R}^m$: The closest point in V to any $\mathbf{b}$ in $\mathbb{R}^m$ is $\mathbf{b}$ itself, so $\mathbf{p} = \mathbf{b}$, $P = I$ and $\widehat{\mathbf{x}} = A^{-1}\mathbf{b}$.

Special Case (not silly)

A is a single nonzero vector $\mathbf{a}$, i.e., V is a line: Then $A^T A = \mathbf{a} \cdot \mathbf{a}$ is a nonzero number, so its inverse is just its reciprocal, and

$$P = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a} \cdot \mathbf{a}} .$$

Moreover,

$$\mathbf{p} = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a} \cdot \mathbf{a}}\mathbf{b} = \frac{\mathbf{a}(\mathbf{a}^T\mathbf{b})}{\mathbf{a} \cdot \mathbf{a}} = \frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{a} \cdot \mathbf{a}}\mathbf{a} , \quad \text{so} \quad \widehat{\mathbf{x}} = \frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{a} \cdot \mathbf{a}} ,$$

because in this case $\widehat{\mathbf{x}}$ is a single number, the coefficient of $\mathbf{a}$ when writing $\mathbf{p}$.

Example

Find the projection $\mathbf{p}$ of $\mathbf{b} = (1, 2, 3)$ onto (the span of) $\mathbf{a} = (1, 0, -1)$ and the corresponding $\widehat{\mathbf{x}}$, and check that $\mathbf{e} = \mathbf{b} - \mathbf{p}$ is orthogonal to $\mathbf{a}$:

$$P = \frac{1}{1+0+1} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix}$$

$$\mathbf{p} = P \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} ,$$

$$\widehat{\mathbf{x}} = \frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{a} \cdot \mathbf{a}} = \frac{1+0+(-3)}{1+0+1} = -1 , \qquad \mathbf{a} \cdot \mathbf{e} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = 0$$

Idempotent matrix

Any projection matrix P, onto the subspace V, say, multiplies all the vectors in $\mathbb{R}^m$ into V, and if we multiply again by P, they don't change. Thus,

$$P^2\mathbf{b} = P\mathbf{b} \text{ for every } \mathbf{b} \text{ in } \mathbb{R}^m ,$$

Therefore, $P^2 = P$. The term for a matrix that equals its own square is *idempotent.*

Same problem, two forms: Form 1

The projection problem: Given vectors $(q_1, q_2, q_3), (r_1, r_2, r_3), (s_1, s_2, s_3)$ in $\mathbb{R}^3$, find the point $(\hat{s}_1, \hat{s}_2, \hat{s}_3) = g(q_1, q_2, q_3) + h(r_1, r_2, r_3)$ closest to (s_1, s_2, s_3), i.e., minimizing

$$\begin{aligned} E &= (s_1 - \hat{s}_1)^2 + (s_2 - \hat{s}_2)^2 + (s_3 - \hat{s}_3)^2 \\ &= (s_1 - (gq_1 + hr_1))^2 + (s_2 - (gq_2 + hr_2))^2 + (s_3 - (gq_3 + hr_3))^2 \ . \end{aligned}$$

In our earlier notation $\mathbf{s} = \mathbf{b}$, $\mathbf{q}$ and $\mathbf{r}$ are the columns of A, and we seek $\widehat{\mathbf{s}} = \mathbf{p}$ and $\widehat{\mathbf{x}} = (g, h)$.

Same problem, two forms: Form 2

The least squares problem: Given points $(q_1, r_1, s_1), (q_2, r_2, s_2), (q_3, r_3, s_3)$ in $\mathbb{R}^3$, with s on the vertical axis, find the subspace, given by the equation $\hat{s} = gq + hr$, that minimizes the sum of the squares of the vertical distances from the given points to the corresponding points on the subspace, i.e., minimizing

$$\begin{aligned} E &= (s_1 - \hat{s}_1)^2 + (s_2 - \hat{s}_2)^2 + (s_3 - \hat{s}_3)^2 \\ &= (s_1 - (gq_1 + hr_1))^2 + (s_2 - (gq_2 + hr_2))^2 + (s_3 - (gq_3 + hr_3))^2 \ . \end{aligned}$$

They are both the same problem because they both come from the same general setup: A linear system $A\mathbf{x} = \mathbf{b}$ with more equations than variables (i.e., $m > n$), so there is no exact solution.

So we want the "best" solution, as measured by least squares.

And that comes from solving the square system $A^T A\mathbf{x} = A^T\mathbf{b}$ (multiply both sides by A^T, to "kill" all the vectors perpendicular to the columns of A, including the difference between $\mathbf{b}$ and its projection $\mathbf{p}$ on the column space).

So the best approximate solution is

$$\widehat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b} \ ,$$

and $\mathbf{p} = A\widehat{\mathbf{x}}$ is as close as you can get to $\mathbf{b}$ in the column space of A.

Same question, answered by (multivariate) calculus: Why is the solution to $A^T A\mathbf{x} = A^T\mathbf{b}$ the best approximation to a solution to $A\mathbf{x} = \mathbf{b}$?

To find $\mathbf{x}$-values $\widehat{\mathbf{x}}$ that minimize

$$E = ||\mathbf{b} - A\mathbf{x}||^2 = \sum_i (b_i - \sum_j a_{ij} x_j)^2 \ ,$$

take partial derivatives with respect to x_k's:

$$\frac{\partial E}{\partial x_k} = \sum_i 2(b_i - \sum_j a_{ij} x_j)(-a_{ik}) \ ;$$

set all equal to 0 and simplify. Result is a system:

$$\sum_i \sum_j a_{ik} a_{ij} x_j = \sum_i a_{ik} b_i$$

which is $A^T A\mathbf{x} = A^T\mathbf{b}$.

The least squares problem comes from: Given data values with n "input" (or "explanatory", or "predictor") variables $x_1, x_2, \ldots, x_n$ and one "output" (or "response") variable y, find the coefficients c_i of the linear equation

$$\hat{y} = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n$$

that best approximates the data. So:

- the x-values of the data points become the constants, the entries in the matrix X (which used to be A),
- the y-values of the data points become $\mathbf{b}$,
- the c's become the coefficients for writing the approximation $\widehat{\mathbf{y}}$ for $\mathbf{y}$ in terms of x-values of the data points — in other words, what $\widehat{\mathbf{x}}$ used to be.

Example 1

Find the plane $\hat{y} = cx_1 + dx_2$ that is the least squares approximation to the data points (0,1,4), (1,0,5), (1,1,8), (-1,2,7):

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ -1 & 2 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 4 \\ 5 \\ 8 \\ 7 \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} 3 & -1 \\ -1 & 6 \end{bmatrix}^{-1} = (1/17)\begin{bmatrix} 6 & 1 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{c} = (X^TX)^{-1}X^T\mathbf{y} = (1/17)\begin{bmatrix} 6 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 6 \\ 26 \end{bmatrix} = \begin{bmatrix} 62/17 \\ 84/17 \end{bmatrix};$$

so the best approximation to these points is
$\hat{y} = (62/17)x_1 + (84/17)x_2$.

Example 2

Find the line $\hat{y} = mx + b$ in $\mathbb{R}^2$ that is the least squares approximation to (i.e., is the *regression line* for) the data points (0,1), (1,2), (2,5), (3,6):

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 5 \\ 6 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 14 & 6 \\ 6 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} 0.2 & -0.3 \\ -0.3 & 0.7 \end{bmatrix}$$

$$\mathbf{c} = (X^T X)^{-1} X^T \mathbf{y} = \begin{bmatrix} 0.2 & -0.3 \\ -0.3 & 0.7 \end{bmatrix} \begin{bmatrix} 42 \\ 24 \end{bmatrix} = \begin{bmatrix} 1.8 \\ 0.8 \end{bmatrix};$$

so the best approximation to these points is $\hat{y} = 1.8x + 0.8$.

Example 3

The heights h in feet of a ball t seconds after it is thrown upward are (0,0.2), (0.2,3.0), (0.4,5.3), (0.6,7.0), (0.8,7.5), (1.0,5.5). The height should be related to the time by $h = a + bt + ct^2$ where c is negative. What are best approximations for the values of a, b, c, the initial height, initial velocity, and half the gravitational constant respectively. Using R:

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0.2 & 0.04 \\ 1 & 0.4 & 0.16 \\ 1 & 0.6 & 0.36 \\ 1 & 0.8 & 0.64 \\ 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 0.2 \\ 3.0 \\ 5.3 \\ 7.0 \\ 7.5 \\ 5.5 \end{bmatrix}$$

Example 3 (ctnd)

$$(T^T T)^{-1} = \begin{bmatrix} 0.8214286 & -2.946429 & 2.232143 \\ -2.9464286 & 18.169643 & -16.741071 \\ 2.2321429 & -16.741071 & 16.741071 \end{bmatrix}$$

$$\begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{bmatrix} = (T^T T)^{-1} T^T \mathbf{h} = \begin{bmatrix} -0.08571429 \\ 19.88571429 \\ -13.92857143 \end{bmatrix}$$

Apparently the ball was not thrown on Earth; it was thrown out of a shallow hole on a smaller planet.

Orthonormal Bases

Vectors are *orthonormal* iff they are pairwise orthogonal and they all have length 1.

If the columns of the matrix Q (traditionally that letter) and Q is $m \times n$ where $m \geq n$, then $Q^T Q = I_n$.

In particular, if Q is $n \times n$ — and still has orthonormal columns — then Q is invertible with inverse its transpose; it is then called *orthogonal.* (Why not "orthonormal", I don't know, and apparently neither does Strang.)

What good is an orthonormal basis?

Suppose $\mathbf{q}_1, \ldots, \mathbf{q}_n$ is an orthonormal basis for $\mathbb{R}^n$. Then for any vector $\mathbf{v}$ in $\mathbb{R}^n$, suppose we have written $\mathbf{v}$ in terms of the $\mathbf{q}$'s:

$$\mathbf{v} = c_1\mathbf{q}_1 + \cdots + c_n\mathbf{q}_n \ .$$

Then for each k from 1 to n:

$$\begin{aligned} \mathbf{q}_k^T\mathbf{v} &= c_1\mathbf{q}_k^T\mathbf{q}_1 + \cdots + c_n\mathbf{q}_k^T\mathbf{q}_n \\ &= c_1(0) + \cdots + c_k(1) + \cdots + c_n(0) = c_k \ . \end{aligned}$$

So we can find the coefficients to write $\mathbf{v}$ in terms of $\mathbf{q}_1, \ldots, \mathbf{q}_n$ just by taking dot products with the $\mathbf{q}$'s.

We want to start with a basis $\mathbf{a}_1, \ldots, \mathbf{a}_n$ for $\mathbb{R}^n$ and build a new basis $\mathbf{q}_1, \ldots, \mathbf{q}_n$ that consists of orthonormal vectors.

$$\mathbf{q}_1 = \mathbf{a}_1 / ||\mathbf{a}_1|| \ .$$

$$\mathbf{e}_2 = \mathbf{a}_2 - \frac{\mathbf{q}_1^T \mathbf{a}_2}{\mathbf{q}_1^T \mathbf{q}_1} \mathbf{q}_1 = \mathbf{a}_2 - (\mathbf{q}_1^T \mathbf{a}_2) \mathbf{q}_1 \ ,$$

$$\text{so } \mathbf{q}_2 = \mathbf{e}_2 / ||\mathbf{e}_2|| \ ;$$

$$\mathbf{e}_3 = \mathbf{a}_3 - (\mathbf{q}_1^T \mathbf{a}_3) \mathbf{q}_1 - (\mathbf{q}_2^T \mathbf{a}_3) \mathbf{q}_2 \ ,$$

$$\text{so } \mathbf{q}_3 = \mathbf{e}_3 / ||\mathbf{e}_3|| \ .$$

And so on.

And what is the process of going back from the **q**'s to the **a**'s? Using the dot product fact that we noted earlier, we get

$$\mathbf{a}_k = \sum_{i=1}^{n} (\mathbf{q}_i^T \mathbf{a}_k)\mathbf{q}_i \ .$$

If we write this in terms of matrices, we get

$$\begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \dots & \mathbf{q}_n \end{bmatrix} \begin{bmatrix} \mathbf{q}_1^T\mathbf{a}_1 & \mathbf{q}_1^T\mathbf{a}_1 & \dots & \mathbf{q}_1^T\mathbf{a}_1 \\ \mathbf{q}_2^T\mathbf{a}_1 & \mathbf{q}_2^T\mathbf{a}_2 & \dots & \mathbf{q}_2^T\mathbf{a}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_n^T\mathbf{a}_1 & \mathbf{q}_n^T\mathbf{a}_2 & \dots & \mathbf{q}_n^T\mathbf{a}_n \end{bmatrix} .$$

If we call A the matrix with columns the $\mathbf{a}$'s, Q the matrix with columns the $\mathbf{q}$'s, and R the matrix of dot products, then $A = QR$. Moreover, look at the entries $\mathbf{q}_i^T \mathbf{a}_j$ in the lower triangle of R, i.e., $i > j$. Then $\mathbf{a}_j$ can be written as a combination of the $\mathbf{q}$'s up to j; and $\mathbf{q}_i$ is orthogonal to the earlier $\mathbf{q}$'s, so it is orthogonal to $\mathbf{a}_j$. So the lower triangle is all 0's, i.e., R is upper triangular.

Example

The vector $(1, -2, 2)/3$ has length 1. We want to find an basis for $\mathbb{R}^3$ that includes it, and then write the corresponding A as QR where Q is orthogonal and R is upper triangular:

The given vector together with (0,1,0) and (0,0,1) is a basis for $\mathbb{R}^3$, so we can use this basis

$$\mathbf{a}_1 = \mathbf{q}_1 = \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix}, \; \mathbf{a}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \; \mathbf{a}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 1/3 & 0 & 0 \\ -2/3 & 1 & 0 \\ 2/3 & 0 & 1 \end{bmatrix}.$$

Example (ctnd)

$$\mathbf{e}_2 = \mathbf{a}_2 - (\mathbf{q}_1^T \mathbf{a}_2)\mathbf{q}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - (-2/3) \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix} = \begin{bmatrix} 2/9 \\ 5/9 \\ 4/9 \end{bmatrix},$$

$$\mathbf{q}_2 = \begin{bmatrix} 2/\sqrt{45} \\ 5/\sqrt{45} \\ 4/\sqrt{45} \end{bmatrix}$$

and ...

Example (ctnd again)

$$
\begin{aligned}
\mathbf{e}_3 &= \mathbf{a}_3 - (\mathbf{q}_1^T \mathbf{a}_3)\mathbf{q}_1 - (\mathbf{q}_2^T \mathbf{a}_3)\mathbf{q}_2 \\
&= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - (2/3) \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix} - (4/\sqrt{45}) \begin{bmatrix} 2/\sqrt{45} \\ 5/\sqrt{45} \\ 4/\sqrt{45} \end{bmatrix} \\
&= \begin{bmatrix} -2/5 \\ 0 \\ 1/5 \end{bmatrix}, \\
\mathbf{q}_3 &= \begin{bmatrix} -2/\sqrt{5} \\ 0 \\ 1/\sqrt{5} \end{bmatrix}
\end{aligned}
$$

Example (ctnd yet again)

We could check that the **q**'s are pairwise orthogonal and each have length 1. We get $A = QR$ where

$$A = \begin{bmatrix} 1/3 & 0 & 0 \\ -2/3 & 1 & 0 \\ 2/3 & 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 1/3 & 2/\sqrt{45} & -2/\sqrt{5} \\ -2/3 & 5/\sqrt{45} & 0 \\ 2/3 & 4/\sqrt{45} & 1/\sqrt{5} \end{bmatrix},$$

$$R = \begin{bmatrix} 1 & -2/3 & 2/3 \\ 0 & 5/\sqrt{45} & 4/\sqrt{45} \\ 0 & 0 & 1/\sqrt{5} \end{bmatrix}.$$