Recitation 7

October 1, 2024

1 Recap

1.1 Orthogonal projection

Given a subspace S of some vector space V, the **orthogonal projection** $p = Pb$ of some vector $b \in V$ is the unique vector $p \in S$ such that $b - p \in S^\perp$. We call the linear operator P that gives $p = Pb$ for any b the "projection matrix" for S.

- $P^2 = P$ since P doesn't change vectors already in S. We can also immediately see that $C(P) = S$ and $N(P) = S^\perp$. $I - P$ is the projection matrix onto $S^\perp$, since $(I - P)b = b - p = e$.
- If $S = C(A)$, then $p = A\hat{x}$ where $\hat{x}$ solves the "normal equations" $A^TA\hat{x} = A^Tb$. If A has full column rank, then A^TA turns out to be invertible (see below), and hence $P = A(A^TA)^{-1}A^T$.
- If $S = C(Q)$ where the columns of Q are an *orthonormal* basis ($Q^TQ = I$), then the above simplifies to $Pb = Q\hat{x}$ where $\hat{x} = Q^Tb$, or equivalently $P = QQ^T$.
- If $S = C(Q)$ where $Q = (q_1)$ is a matrix with a *single* unit-vector column q_1 (i.e. S is a 1d subspace), then P further simplifies to the rank-1 matrix $q_1q_1^T$ (projection onto a line!).

We showed that the orthogonal projection $p = Pb$ is the *closest* vector to b in S. That is, for $S = C(A)$, $p = A\hat{x}$ *minimizes* the "error norm" $\|b - Ax\|$ over all possible x: the **least-square solution** $\hat{x}$ is an *approximate* solution to $Ax = b$ in cases where $b \notin C(A)$. Furthermore, we showed how this can be used for least-square fitting of models to data: Ax represents a "model" (e.g. a polynomial) with unknown coefficients x (e.g. the polynomial coefficients), and b represents the dependent variables of the data we are trying to fit. The least-square solution $\hat{x}$ are the best-fit parameters of the model for the data.

1.2 Gram–Schmidt and QR

Given a set of linearly independent vectors $a_1, a_2, a_3, \ldots$ (the columns of a matrix A with full column rank), we can *construct* an orthonormal basis $q_1, q_2, q_3, \ldots$ for $C(A)$ with the following "Gram–Schmidt" algorithm:

1. $q_1 = a_1 / \underbrace{\|a_1\|}_{r_{11}}$ (just normalize the first vector). Now, $\text{span}\{q_1\} = \text{span}\{a_1\}$.
2. $v_2 = (I - q_1 q_1^T) a_2 = a_2 - q_1 \underbrace{(q_1^T a_2)}_{r_{12}}$ (project a_2 to $v_2 \perp q_1$) and $q_2 = v_2 / \underbrace{\|v_2\|}_{r_{22}}$ (normalize). Now, $\text{span}\{q_1, q_2\} = \text{span}\{a_1, a_2\}$.
3. $v_3 = (I - q_1 q_1^T - q_2 q_2^T) a_3 = a_3 - q_1 \underbrace{(q_1^T a_3)}_{r_{13}} - q_2 \underbrace{(q_2^T a_3)}_{r_{23}}$ (project a_3 to $v_3 \perp q_1, q_2$) and $q_3 = v_3 / \underbrace{\|v_3\|}_{r_{33}}$ (normalize). Now, $\text{span}\{q_1, q_2, q_3\} = \text{span}\{a_1, a_2, a_3\}$.

Here, we have labelled the coefficients in the algorithm because it turns out that you can interpret this as a *factorization* of A (much like we interpreted Gaussian elimination as an LU factorization), the ("thin") **QR factorization**:

$$\underbrace{\begin{pmatrix} a_1 & a_2 & a_3 & \cdots \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} q_1 & q_2 & q_3 & \cdots \end{pmatrix}}_{Q} \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots \\ & r_{22} & r_{23} & \cdots \\ & & r_{33} & \cdots \\ & & & \ddots \end{pmatrix}}_{R}$$

in which A is $m \times n$ (assumed full column rank, "tall" $m \geq n$), Q is $m \times n$ with orthonormal columns ($Q^T Q = I$), and R is $n \times n$ upper-triangular and invertible. The *reason* for the upper-triangular structure is that, by construction, $\text{span}\{q_1, \ldots, q_k\} = \text{span}\{a_1, \ldots, a_k\}$ for any k: each column of A is made up of that column of Q and preceding columns only, and vice versa.

1.3 New material

In class, Prof. Johnson claimed that if A has full column rank, then $A^T A$ is invertible. The key fact to show this is that $\boxed{N(A) = N(A^T A)}$ for *any* matrix A.

Proof: Easy: if $x \in N(A)$, then $Ax = 0$ and hence $A^T Ax = 0$ and $x \in N(A^T A)$. Tricky: if $x \in N(A^T A)$, then $A^T Ax = 0$, hence $x^T A^T Ax = 0 = (Ax)^T (Ax) = \|Ax\|^2$. But the only way we can have $\|Ax\| = 0$ is if $Ax = 0$, hence $x \in N(A)$. Q.E.D.

It follows that if A has full column rank, i.e. if $N(A) = \{\vec{0}\}$, then $A^T A$ also has full column rank. But since $A^T A$ is square, this means it is invertible.

2 Exercises

1. Since $N(A) = N(A^TA)$, explain why:
   (a) (Using orthogonal complements) why $C(A^T) = C(A^TA)$.
   (b) Why, therefore, $A^TA\hat{x} = A^Tb$ must *always* have a solution $\hat{x}$, even if A^TA is *not* invertible.
   (c) If A^TA is *not* invertible, $\hat{x}$ is not unique. But why is the projection $p = A\hat{x}$ still unique?

2. Find the projection matrix P onto the column space $C(A)$ for $A = \begin{pmatrix} 3 & 6 & 6 \\ 4 & 8 & 8 \end{pmatrix}$. Look closely at the matrix before you plunge into calculations!

3. (From Strang section 4.3.) Consider the two lines in 3d defined by the points $\mathcal{P}(x) = \begin{pmatrix} x \\ x \\ x \end{pmatrix}$ and $\mathcal{Q}(y) = \begin{pmatrix} y \\ 3y \\ -1 \end{pmatrix}$. We want to choose x and y to minimize $\|\mathcal{P}(x) - \mathcal{Q}(y)\|^2$.
   (a) Express this problem in matrix form as minimizing $\|A\vec{x} - \vec{b}\|^2$ for $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}$, some matrix A, and some vector $\vec{b}$.
   (b) Find the values $\hat{x}, \hat{y}$ that minimizes the distance.
   (c) The line connecting the closest points, i.e. connecting $\mathcal{P}(\hat{x})$ and $\mathcal{Q}(\hat{y})$ is perpendicular to _____ ?

3 Solutions

1. (a) $C(A^T) = N(A)^\perp = N(A^TA)^\perp = C((A^TA)^T) = C(A^TA)$.

   (b) $A^TA\hat{x} = A^Tb$ must *always* have a solution $\hat{x}$ because the right-hand-side $A^Tb \in C(A^T) = C(A^TA)$ is in the column space of the matrix A^TA on the left-hand side.

   (c) The solution is not unique because $\hat{x}+v$ is also a solution for any $v \in N(A^TA) = N(A)$. But this gives the *same* projection $p = A(\hat{x}+v) = A\hat{x} + \cancelto{0}{Av} = A\hat{x}$.

2. By inspection, all of the columns are parallel, so the column space is spanned by $a = \begin{pmatrix}3\\4\end{pmatrix}$. In this case, our projection formula simplifies to

$$P = a(a^Ta)^{-1}a^T = \frac{aa^T}{a^Ta} = \frac{\begin{pmatrix}3 & 4\end{pmatrix}\begin{pmatrix}3\\4\end{pmatrix}}{3^2+4^2} = \frac{1}{25}\begin{pmatrix}9 & 12\\12 & 16\end{pmatrix}$$

   since a^Ta is a *scalar* that we can easily invert and pull out of the expression.

3. (From Strang section 4.3.) Consider the two lines in 3d defined by the points $\mathcal{P}(x) = \begin{pmatrix}x\\x\\x\end{pmatrix}$ and $\mathcal{Q}(y) = \begin{pmatrix}y\\3y\\-1\end{pmatrix}$. We want to choose x and y to minimize $\|\mathcal{P}(x) - \mathcal{Q}(y)\|^2$.

   (a) We have $\mathcal{P}(x) = x\begin{pmatrix}1\\1\\1\end{pmatrix}$ and $\mathcal{Q}(y) = y\begin{pmatrix}1\\3\\0\end{pmatrix} + \begin{pmatrix}0\\0\\-1\end{pmatrix}$, so the difference is:

$$\mathcal{P}(x) - \mathcal{Q}(y) = x\begin{pmatrix}1\\1\\1\end{pmatrix} - y\begin{pmatrix}1\\3\\0\end{pmatrix} - \begin{pmatrix}0\\0\\-1\end{pmatrix} = \underbrace{\begin{pmatrix}1 & -1\\1 & -3\\1 & 0\end{pmatrix}}_{A}\underbrace{\begin{pmatrix}x\\y\end{pmatrix}}_{\vec{x}} - \underbrace{\begin{pmatrix}0\\0\\-1\end{pmatrix}}_{b},$$

   which defines our A and $\vec{b}$.

   (b) The minimum of $\|A\vec{x} - \vec{b}\|$ is found by solving the normal equations $A^TA\vec{x} = A^Y\vec{b}$. Plugging our A and $\vec{b}$ in gives:

$$\underbrace{\begin{pmatrix}3 & -4\\-4 & 10\end{pmatrix}}_{A^TA}\begin{pmatrix}\hat{x}\\\hat{y}\end{pmatrix} = \begin{pmatrix}-1\\0\end{pmatrix},$$

   which has solution $\hat{x} = -5/7$ and $\hat{y} = -2/7$.

   (c) The line connecting the closest points, i.e. connecting $\mathcal{P}(\hat{x})$ and $\mathcal{Q}(\hat{y})$ is perpendicular to the **column space** $C(A)$ (because this is orthogonal projections), which means that it is perpendicular to **both lines**.