247 DAY 19

6.5: Least-Squares Problems

Often in the real world, a system of equations that has no solution arises out of a model for a large dataset. A typical example is that of linear regression, in which one variable is assumed to depend linearly on another – but the data don't all lie perfectly along one line. In such a case, it is desirable to find the "best approximation" to a solution in some sense. The generally accepted solution to this problem is a so-called "least-squares solution."

Definition. Let A be an $m \times n$ matrix, $\mathbf{b}$ in $\mathbb{R}^m$. A *least-squares solution* to the (possibly inconsistent) equation $A\mathbf{x} = \mathbf{b}$ is a vector $\hat{\mathbf{x}}$ in R^n such that $\|\mathbf{b} - A\hat{\mathbf{x}}\|$ is as small as possible; that is, such that $\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$.

Since the set of all possible values of $A\mathbf{x}$ is the column space of A, we see that to minimize the desired distance, $A\hat{\mathbf{x}}$ should be the orthogonal projection of $\mathbf{b}$ onto $\operatorname{Col} A$. If we label this projection as $\hat{\mathbf{b}}$, we see that a least-squares solution is a solution of the equation $A\mathbf{x} = \hat{\mathbf{b}}$, which is consistent by design, since $\hat{\mathbf{b}}$ lives in $\operatorname{Col} A$.

Question: When is the least-squares solution unique?

How do we find the least-squares solution(s)? Well, notice that the vector $\mathbf{b} - \hat{\mathbf{b}}$ is orthogonal to $\operatorname{Col} A$, which is equivalent, as usual, to $A^T(\mathbf{b} - \hat{\mathbf{b}}) = \mathbf{0}$, or $A^T\hat{\mathbf{b}} = A^T\mathbf{b}$. If we write $\hat{\mathbf{b}}$ as $A\hat{\mathbf{x}}$, this gives the equation $A^TA\mathbf{x} = A^T\mathbf{b}$. The equations arising out of this matrix equation are called the *normal equations* for the equation $A\mathbf{x} = \mathbf{b}$, and any solution of the normal equations gives a least-squares solution.

Example. *Find the least-squares solution of* $A\mathbf{x} = \mathbf{b}$ *where* $A = \begin{bmatrix} 1 & -2 \\ -1 & 2 \\ 0 & 3 \\ 2 & 5 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ -4 \\ 2 \end{bmatrix}$.

Before we start, let's get our dimensions straight. What size will $\hat{\mathbf{x}}$ be? How about $\hat{\mathbf{b}}$? A^TA? The calculations come out $A^TA = \begin{bmatrix} 6 & 6 \\ 6 & 42 \end{bmatrix}$, $A^T\mathbf{b} = \begin{bmatrix} 6 \\ -6 \end{bmatrix}$, and $\hat{\mathbf{x}} = \begin{bmatrix} \frac{4}{3} \\ -\frac{1}{3} \end{bmatrix}$.

As we observed above, the least-squares solution $\hat{\mathbf{x}}$ is unique precisely when the columns of A are linearly independent. Since that implies that the equation $A^TA\mathbf{x} = A^T\mathbf{b}$ has a unique solution, and A^TA is a square matrix, it seems to me that this immediately implies that A^TA is invertible, by the Invertible Matrix Theorem, thus proving Theorem 14. The text refers to the exercises for a proof, but

I like mine better! In practice, as the author points out, formula (4) in Theorem 14 is not generally used, because calculating matrix inverses is computationally expensive.

In any event, the projection $\hat{\mathbf{b}} = A\hat{\mathbf{x}}$ is always unique, even when $\hat{\mathbf{x}}$ itself is not, so the *least-squares error* of the approximate solution, given by $\|\mathbf{b} - A\hat{\mathbf{x}}\|$, is always uniquely defined.

Example. In the above example, what was the least-squares error? Well, $A\hat{\mathbf{x}} = \begin{bmatrix} 2 \\ -2 \\ -1 \\ 1 \end{bmatrix}$, so the least-squares error is

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| = \sqrt{(3-2)^2 + (1+2)^2 + (-4+1)^2 + (2-1)^2} = 2\sqrt{5}.$$

For the purposes of dealing with numerical issues such as round-off error and ill-conditionedness of the matrix A, using QR-factorization is far superior to the above method – the idea being, as far as I can tell, that orthogonal vectors don't interfere with one another very much. How does this work? Well, remember that Q is a matrix with orthonormal columns whose column space is equal to the column space of A. Therefore, $QQ^T\mathbf{b}$ gives us the desired projection $\hat{\mathbf{b}}$ onto $\operatorname{Col} A$. This gives us the equation $A\hat{\mathbf{x}} = QQ^T\mathbf{b}$, or $QR\hat{\mathbf{x}} = QQ^T\mathbf{b}$, or $Q^TQR\hat{\mathbf{x}} = Q^TQQ^T\mathbf{b}$. Since $Q^TQ = I_n$, we see that $R\hat{\mathbf{x}} = Q^T\mathbf{b}$. Now, recall that R is upper-triangular with positive entries on the diagonal and is thus invertible. Multiplying both sides on the right by R^{-1} then gives Theorem 15. On the other hand, see the Numerical Note on p. 415: Typically, we just backsolve to find $\hat{\mathbf{x}}$, since R is in a good form for doing that.

Example. Return again to our example. Let's do this problem with QR-factorization. I used Maple to produce the factorization

$$A = QR = \begin{bmatrix} \frac{\sqrt{6}}{6} & -\frac{1}{2} \\ -\frac{\sqrt{6}}{6} & \frac{1}{2} \\ 0 & \frac{1}{2} \\ \frac{\sqrt{6}}{3} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \sqrt{6} & \sqrt{6} \\ 0 & 6 \end{bmatrix}.$$

Now, $Q^T\mathbf{b} = \begin{matrix} \frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{6} & 0 & \frac{\sqrt{6}}{3} \\ \frac{-1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{matrix} \begin{bmatrix} 3 \\ 1 \\ -4 \\ 2 \end{bmatrix} = \begin{bmatrix} \sqrt{6} \\ -2 \end{bmatrix}$. Let $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and back solve $R\mathbf{x} = \mathbb{Q}^t\mathbf{b}$. First we get $6x_2 = -2$, so $x_2 = -\frac{1}{3}$, and then dividing through by $\sqrt{6}$ gives $x_1 - \frac{1}{3} = 1$, so $x_1 = \frac{4}{3}$, recovering the result.