Inner Products and Least Squares

Preliminaries

The goal of these notes is to present the *method of least squares*. The simplest way to view this conceptually is using the inner product. We will need a few observations using inner products. In what we do, we will be careful to use only the general properties of an inner product, not those that are only special to R^n. This way our results will hold even in situations where the inner product is defined by an integral or some other rule.

OBSERVATION 1. If X_0 is a vector with the property that $\langle X_0, Y\rangle = 0$ for all vectors Y, then $X_0 = 0$. In other words, the only vector X_0 that is perpendicular to all vectors is the zero vector. The proof is simple. Since Y can be any vector, we make the special choice $Y = X_0$. Then $0 = \langle X_0, Y\rangle = \langle X_0, X_0\rangle = \|X_0\|^2$ so $X_0 = 0$.

OBSERVATION 2 *adjoint.* Let A be a matrix, not necessarily square. We define the **adjoint** of A, written A^*, to be the matrix that satisfies the following identity for any choice of vectors X and Y

$$\langle X, AY\rangle = \langle A^*X, Y\rangle. \tag{1}$$

Although this may appear strange, the matrix A^* arises frequently in applications. By writing X and Y in coordinates, one finds

$$\langle X, AY\rangle = \sum_i x_i \left(\sum_j a_{ij}y_j\right) = \sum_j \left(\sum_i a_{ij}x_i\right) y_j = \langle A^T X, Y\rangle.$$

Thus A^* is just A^T, the *transpose* of A. The *only* reason the transpose is important is because one so frequently needs the identity (1).

A matrix is called *symmetric* (or *self-adjoint*) if it equals its adjoint: $A = A^*$. It is called *anti-symmetric (or skew-adjoint)* if $A^* = -A$.

OBSERVATION 3 *orthogonal projection onto a line.* Let X and Y be given vectors. We would like to write Y in the form $Y = cX + V$, where V is perpendicular to X. Then the vector cX is the **orthogonal projection** of Y in the line determined by the vector X.

How can we find the constant c and the vector V? We use the only fact we know: that V is supposed to be perpendicular to X. Thus we take the inner product of $Y = cX + V$ with X and conclude that $\langle X, Y\rangle = c\langle X, X\rangle$, that is

$$c = \frac{\langle X, Y\rangle}{\|X\|^2}.$$

Now that we know c, we can simply define V by the obvious formula $V = Y - cX$.

At first this may seem circular. To convince your self that this works, let $X = (1,1)$, and $Y = (2,3)$. Then compute c and V and draw a sketch showing X, Y, cX, and V.

Since $cX \perp V$, we can use the Pythagorean Theorem to conclude that

$$\|Y\|^2 = c^2\|X\|^2 + \|V\|^2 \geq c^2\|X\|^2.$$

From this, using the explicit value of c found above we obtain the *Schwarz inequality*

$$|\langle X, Y\rangle| \leq \|X\|\|Y\|.$$

Notice that this was done without trigonometry. It used only the properties of the inner product.

OBSERVATION 4 *orthogonal projection into a subspace.* If a linear space has an inner product and S is a subspace of it, we can discuss the orthogonal projection of a vector into that subspace. Given a vector Y, if we can write

$$Y = U + V,$$

where U is in S and V is perpendicular to S, then we call U the projection of Y into S and V the projection of Y perpendicular to S. The notation $U = P_S Y. V = P_S^{\perp} Y$ is frequently used for this projection U. By the Pythagorean theorem

$$\|Y\|^2 = \|U\|^2 + \|V\|^2, \qquad (U = P_S Y,\ V = P_S^{\perp} Y).$$

It is easy to show that *the projection $P_S Y$ is closer to Y than any other point in S*. In other words,

$$\|Y - P_S Y\| \leq \|Y - X\| \qquad \text{for all } X \text{ in } S.$$

To see this, given any $X \in S$ write $Y - X = (Y - P_S Y) + (P_S Y - X)$ and observe that $Y - P_S Y$ is perpendicular to S while $P_S Y$ and X, and hence $P_S Y - X$ are in S. Thus by the Pythagorean Theorem

$$\|Y - X\|^2 = \|Y - P_S Y\|^2 + \|P_S Y - X\|^2 \geq \|Y - P_S Y\|^2.$$

This is what we asserted.

Least Squares

Say you measure the same quantity c four times and get the numbers c_1, c_2, c_3 and c_4. What should you use as your best estimate of the number c? One approach is to pick the number c to minimize the square of the error

$$\text{Error}(c) = (c_1 - c)^2 + (c_2 - c)^2 + (c_3 - c)^2 + (c_4 - c)^2.$$

By a calculation, perhaps using calculus, this gives the *mean* or "average"

$$c = \frac{c_1 + c_2 + c_3 + c_4}{4}.$$

The essential reason the mean c is a "good" measure is that it minimizes this Error(c).

Now we move to a more complicated problem. Say we are given n experimental data points (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ and want to find the straight line $y = a + bx$ that fits this data best. How should be proceed? Ideally we want to pick the coefficients a and b so that

$$\begin{aligned} a + bx_1 &= y_1 \\ a + bx_2 &= y_3 \\ &\cdots \\ a + bx_n &= y_n. \end{aligned}$$

However, these are n equations for the two unknowns a, b, and it is unlikely that we can solve them exactly. Following the suggestion of the simpler situation we just considered, we can pick a, b to minimize the error

$$\text{Error}(a, b) = (a + bx_1 - y_1)^2 + (a + bx_2 - y_2)^2 + \cdots + (a + bx_n - y_n)^2.$$

One can find a and b using calculus. But one gets more insight by using the inner product. We write the above equations in matrix notation as

$$AV = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \cdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ y_n \end{pmatrix} = Y,$$

that is, $AV = Y$. Then

$$\text{Error}(V) = \|AV - Y\|^2.$$

Thus, we want to pick V so that $W = AV$ is as close as possible to Y. Notice that W must be in the image of A. From OBSERVATION 4 above, we want to let W be the orthogonal projection of Y into the image of A.

How can we compute this? Notice that $AV - Y$ will then be perpendicular to the image of A. In other words, $AV - Y$ will be perpendicular to all vectors of the form AU for any vector U. Thus by OBSERVATION 2

$$0 = \langle AU, AV - Y\rangle = \langle U, A^*(AV - Y)\rangle.$$

But now since the right side holds for *all* vectors U we can apply OBSERVATION 1 to conclude that

$$A^*AV = A^*Y. \tag{2}$$

These are the **normal equations** for V and are what we are seeking.

Although this may seem abstract, it is easy to compute this explicitly.

$$A^*A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \cdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{pmatrix}.$$

The computation of A^*Y is equally straightforward so the normal equations are two equations in two unknowns:

$$\begin{pmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum y_j \\ \sum x_j y_j \end{pmatrix}.$$

These can be solved using high school algebra.

Identical methods can be used to find, for instance, the quadratic polynomial $y = a + bx + cx^2$ that best fits some data, or the plane $z = a + bx + cy$ that best fits given data. The technique of least squares is widely used in all area where one has experimental data.