Patch Kessler

March 16, 2006

Notes on Least Squares

Goal: Let $\mathbf{a}_1, \ldots, \mathbf{a}_n$ be a set of vectors in some vector space V. Given $\mathbf{b} \in V$, we would like to find n scalars x_i so that $\mathbf{b}$ equals the linear combination $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n =: \Sigma x_i\mathbf{a}_i$.

Snag: Unfortunately, when $\mathbf{b} \notin \text{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$, it is impossible to achieve the **Goal**. For any collection of scalars x_i, the error vector $\mathbf{e} := \Sigma x_i\mathbf{a}_i - \mathbf{b}$ will be nonzero.

Resolution: Instead of achieving $\mathbf{e} = \mathbf{0}$, we settle for the next best thing; we find scalars x_i for which the length of the error vector $\mathbf{e}$ is as small as possible.

The square of the *length* of $\mathbf{e}$ is defined to be $\langle \mathbf{e}, \mathbf{e} \rangle$ where $\langle \cdot, \cdot \rangle$ is an inner product on V. In addition to providing us with length, the inner product tells us when vectors are *orthogonal*, for instance $\mathbf{e}$ and $\mathbf{a}_1$ are said to be orthogonal if $\langle \mathbf{a}_1, \mathbf{e} \rangle = 0$. We might suspect from this suggestive geometric language that the length of the error vector $\mathbf{e} := \Sigma x_i\mathbf{a}_i - \mathbf{b}$ is as small as possible when $\mathbf{e}$ is orthogonal to the plane spanned by the $\mathbf{a}_i$'s, that is, when $\langle \mathbf{a}_i, \mathbf{e} \rangle = 0$ for each $\mathbf{a}_i$. In fact, it is straightforward to show that this is true[1]. When $n = 2$, these so called normal equations $\langle \mathbf{a}_i, \mathbf{e} \rangle = \langle \mathbf{a}_i, x_1\mathbf{a}_1 + x_2\mathbf{a}_2 - \mathbf{b} \rangle = 0$ are given by

$$\left.\begin{array}{l} \langle \mathbf{a}_1, \mathbf{e} \rangle = 0 \longrightarrow x_1\langle \mathbf{a}_1, \mathbf{a}_1 \rangle + x_2\langle \mathbf{a}_1, \mathbf{a}_2 \rangle = \langle \mathbf{a}_1, \mathbf{b} \rangle \\ \langle \mathbf{a}_2, \mathbf{e} \rangle = 0 \longrightarrow x_1\langle \mathbf{a}_2, \mathbf{a}_1 \rangle + x_2\langle \mathbf{a}_2, \mathbf{a}_2 \rangle = \langle \mathbf{a}_2, \mathbf{b} \rangle \end{array}\right\} \longrightarrow \begin{bmatrix} \langle \mathbf{a}_1, \mathbf{a}_1 \rangle & \langle \mathbf{a}_1, \mathbf{a}_2 \rangle \\ \langle \mathbf{a}_2, \mathbf{a}_1 \rangle & \langle \mathbf{a}_2, \mathbf{a}_2 \rangle \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \langle \mathbf{a}_1, \mathbf{b} \rangle \\ \langle \mathbf{a}_2, \mathbf{b} \rangle \end{bmatrix}$$

allowing us to solve for the x_i's. For general n, the proceedure is the same, with normal equations given by

$$\langle \mathbf{a}_i, \mathbf{e} \rangle = \langle \mathbf{a}_i, x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n - \mathbf{b} \rangle = 0 \tag{1}$$

for each $\mathbf{a}_i$, immediately giving a linear system for the n x_i's.

$$\begin{bmatrix} \langle \mathbf{a}_1, \mathbf{a}_1 \rangle & \langle \mathbf{a}_1, \mathbf{a}_2 \rangle & \cdots & \langle \mathbf{a}_1, \mathbf{a}_n \rangle \\ \langle \mathbf{a}_2, \mathbf{a}_1 \rangle & \langle \mathbf{a}_2, \mathbf{a}_2 \rangle & \cdots & \langle \mathbf{a}_2, \mathbf{a}_n \rangle \\ \vdots & \vdots & & \vdots \\ \langle \mathbf{a}_n, \mathbf{a}_1 \rangle & \langle \mathbf{a}_n, \mathbf{a}_2 \rangle & \cdots & \langle \mathbf{a}_n, \mathbf{a}_n \rangle \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \langle \mathbf{a}_1, \mathbf{b} \rangle \\ \langle \mathbf{a}_2, \mathbf{b} \rangle \\ \vdots \\ \langle \mathbf{a}_n, \mathbf{b} \rangle \end{bmatrix} \tag{2}$$

Examples:

1. Let $V = \mathbb{R}^2$, $\mathbf{a}_1 = [1 \;\; 0.5]^T$, and $\mathbf{b} = [1 \;\; 1]^T$. With $n = 1$, (2) reduces to $\langle \mathbf{a}_1, \mathbf{a}_1 \rangle x_1 = \langle \mathbf{a}_1, \mathbf{b} \rangle$ which gives the x_1 for which $x_1\mathbf{a}_1$ is as close as possible to $\mathbf{b}$. Before turning this page, decide for yourself what the value of x_1 should be.

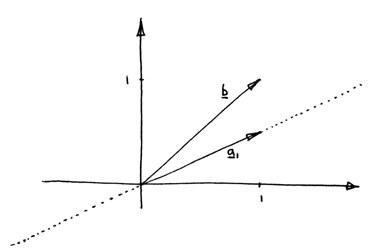

Figure 1: What value of x_1 causes $x_1\mathbf{a}_1$ to be as close as possible to $\mathbf{b}$?

[1] If our x_i's are chosen so that the error $\mathbf{e} = \Sigma x_i\mathbf{a}_i - \mathbf{b}$ is orthogonal to each $\mathbf{a}_i$, then adding any combination $\mathbf{a}$ of the $\mathbf{a}_i$'s to $\Sigma x_i\mathbf{a}_i$ gives a new error $\mathbf{e} + \mathbf{a}$, which has a squared length $\langle \mathbf{e} + \mathbf{a}, \mathbf{e} + \mathbf{a} \rangle = \langle \mathbf{e}, \mathbf{e} \rangle + 2\langle \mathbf{e}, \mathbf{a} \rangle + \langle \mathbf{a}, \mathbf{a} \rangle = \langle \mathbf{e}, \mathbf{e} \rangle + \langle \mathbf{a}, \mathbf{a} \rangle \geq \langle \mathbf{e}, \mathbf{e} \rangle$ with equality iff $\mathbf{a} = \mathbf{0}$.

Trick Question!! :-) The value of x_1 depends on the inner product $\langle\cdot,\cdot\rangle$, which we haven't yet defined! Different inner products give different concepts of length, and thus different values of x_1 for which $x_1\mathbf{a}_1 - \mathbf{b}$ has as small a "length" as possible.

– if the inner product is given by $\langle \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \rangle = u_1 v_1 + u_2 v_2$, then $x_1 = \frac{6}{5}$.

– if the inner product is given by $\langle \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \rangle = u_1 v_1 + 2u_2 v_2$, then $x_1 = \frac{4}{3}$.

2. Let $V = \mathbb{R}^3$, $\mathbf{a}_1 = [1 \ \ 0 \ \ 0]^T$, $\mathbf{a}_1 = [0 \ \ 1 \ \ 0]^T$, and $\mathbf{b} = [1 \ \ 1 \ \ 1]^T$, and let $\langle\cdot,\cdot\rangle$ be the standard dot product. Our job as always is to find the linear combination of $\mathbf{a}_i$'s that is closest to $\mathbf{b}$. The concepts of *length* and *orthogonality* associated with this inner product coencide with the everyday meanings of these terms, and so we can solve this problem almost at once. First, observe that the projection of $\mathbf{b}$

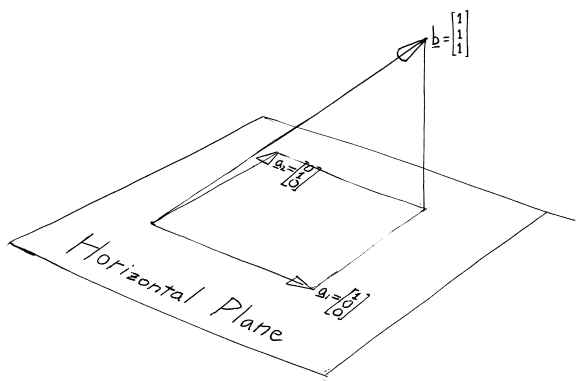

Figure 2: A picture of $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{b}$, and the horizontal plane.

into the horizontal plane (spanned by $\mathbf{a}_1$ and $\mathbf{a}_2$) is given by $[1 \ \ 1 \ \ 0]^T = 1 \cdot \mathbf{a}_1 + 1 \cdot \mathbf{a}_2$. It is immediately obvious that this projection is the closest vector in the horizontal plane to $\mathbf{b}$, and so we are done. Because the solution is trivial, the problem can be used as a vehicle for developing an intuition for the formal mathematics. Solve for the x_i's rigorously using the steps from the previous page, as if you had no intuition for the vector space and inner product. For instance, imagine that $V = \mathbb{R}^{1000}$, or better yet, that V is a vector space of functions, and that $\langle\cdot,\cdot\rangle$ is some truly bizzare inner product, like those that Professor Neu likes to put on the homework.

Comments:

- Fitting a straight line $\eta(\xi) = x_1 + x_2\xi$ through 1000 data points (ξ_i, η_i) is a least squares problem with the same structure as example 2, but that takes place in $\mathbb{R}^{1000}$ instead of $\mathbb{R}^2$.

- The exciting thing for me about linear algebra is that all sorts of objects qualify as vectors besides lists of scalars. For instance, the continuous functions on $[a, b]$ comprise a vector space. We can define an inner product on this space, and use it to get a linear combination of a set of functions (vectors) that is as close as possible to some other function (vector). Of course "close" depends on how we've defined the inner product, and the possibilities are endless. For instance we can define the product so that agreement of function derivatives counts for more than agreement of function values or vice versa. We can even count agreement at the endpoints a and b, as in Problem Set 6.

- We can find a linear combination of functions (vectors) that is as close as possible to the solution function (vector) of a BVP **even if we don't know what that solution is**! This is because when we set up (2) to find the coefficients x_i, the various matrix entries become integrals from the BVP problem statement. This is exactly what happened in Problem Set 6.