Math 232 Preamble for CA4 - Least squares fitting

We know that the system $A\mathbf{x} = \mathbf{b}$ is consistent if only if $\mathbf{b}$ is in the span of the columns of A ($\mathbf{b} \in \text{col}(A)$).

We can re-write the system $A\mathbf{x} = \mathbf{b}$ as $\mathbf{b} - A\mathbf{x} = \mathbf{0}$, so that if the system is consistent and $\mathbf{x}$ is a solution, then $\|\mathbf{b} - A\mathbf{x}\| = \|\mathbf{0}\| = 0$.

If $\mathbf{b} \notin \text{col}(A)$ then there is no solution; there is no vector $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{b}$ (the system is inconsistent).

When the system is inconsistent $A\mathbf{x}$ is *never* equal to $\mathbf{b}$ so

$$\|\mathbf{b} - A\mathbf{x}\| > 0 \quad \text{for all } \mathbf{x}$$

In this case we look for a vector $\hat{\mathbf{x}}$ that makes $\|\mathbf{b} - A\hat{\mathbf{x}}\|$ as small as possible (so $A\hat{\mathbf{x}}$ is 'close' to $\mathbf{b}$). In this sense we have an 'approximate' solution to $A\mathbf{x} = \mathbf{b}$;

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| \ \leq \ \|\mathbf{b} - A\mathbf{x}\| \quad \text{for all } \mathbf{x}$$

This $\hat{\mathbf{x}}$ will be our *best* approximate solution to the inconsistent system $A\mathbf{x} = \mathbf{b}$.

To summarize, we determined a way to 'solve' the system $A\mathbf{x} = \mathbf{b}$ when it is inconsistent ($\mathbf{b} \notin \text{col}(A)$) by solving instead the (consistent) system $A\mathbf{x} = \hat{\mathbf{b}}$ where $\hat{\mathbf{b}}$ is the closest vector to $\mathbf{b}$ in $\text{col}(A)$. This gives us the least squares solution of $A\mathbf{x} = \mathbf{b}$ (this is the definition of least squares solution to a linear system of equations.) Let's observe also that if the system $A\mathbf{x} = \mathbf{b}$ is consistent, this method finds the exact solution (right? because then $\hat{\mathbf{b}} = \mathbf{b}$).

It remains to solve $A\mathbf{x} = \hat{\mathbf{b}}$. Recall that

$$\mathbf{x} \ = \ \text{proj}_W\mathbf{x} \ + \ \text{proj}_{W^\perp}\mathbf{x} \ = \ \mathbf{x}_1 \ + \ \mathbf{x}_2$$

for any vector $\mathbf{x} \in \mathbf{R}^n$ and any subspace $W \subset \mathbf{R}^n$. So we write

$$\mathbf{b} \ = \ \hat{\mathbf{b}} \ + \ \mathbf{b}_2 \ = \ \text{proj}_W\mathbf{b} \ + \ \text{proj}_{W^\perp}\mathbf{b}$$

where $W = \text{col}(A)$. And then we observe that $W^\perp = \text{col}(A)^\perp = \text{null}(A^T)$. Since $\mathbf{b}_2 \in W^\perp$ (right?), we have that $A^T\mathbf{b}_2 = \mathbf{0}$. Writing $\mathbf{b}_2 = \mathbf{b} - \hat{\mathbf{b}} = \mathbf{b} - \text{proj}_{\text{col}(A)}\mathbf{b}$, we obtain

$$A^T(\mathbf{b} - \text{proj}_{\text{col}(A)}\mathbf{b}) \ = \ A^T(\mathbf{b} - A\hat{\mathbf{x}}) \ = \ \mathbf{0}$$

and re-arranging this we obtain

$$A^TA\hat{\mathbf{x}} \ = \ A^T\mathbf{b} \tag{1}$$

Equation (1) is called the **normal equation for** $\hat{\mathbf{x}}$. Note that A^TA and $A^T\mathbf{b}$ are 'easy' to compute, so solving this system (via row reduction) would be easier than solving $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$.

Now, if A is $m \times n$, then A^TA is $n \times n$ so it may be invertible. If it is then we can solve (1) for $\hat{\mathbf{x}}$;

$$\hat{\mathbf{x}} \;=\; (A^TA)^{-1}A^T\mathbf{b} \qquad (2)$$

and we have a unique least squares solution $\hat{\mathbf{x}}$. If A^TA is singular (non-invertible), then there are many least squares solutions.

There are many, many applications of least squares solutions (statistics, engineering, finance, ...).

Example 1: $A = \begin{bmatrix} 1 & -1 \\ 3 & 2 \\ -2 & 4 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 4 \\ 1 \\ 3 \end{bmatrix}$. Check that $\mathbf{b} \notin \text{col}(A)$, so the system $A\mathbf{x} = \mathbf{b}$ is inconsistent. Let's find the least squares solution(s).

Step 1: Compute A^TA;

$$A^TA = \begin{bmatrix} 1 & 3 & -2 \\ -1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 3 & 2 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} 14 & -3 \\ -3 & 21 \end{bmatrix}$$

.

Step 2: $\det(A^TA) \neq 0$ so (A^TA) invertible and therefore there is only *one* least squares solution. Compute $(A^TA)^{-1}$;

$$(A^TA)^{-1} \;=\; \frac{1}{285}\begin{bmatrix} 21 & 3 \\ 2 & 14 \end{bmatrix}$$

Step 3: Compute $\hat{\mathbf{x}}$;

$$\hat{\mathbf{x}} \;=\; (A^TA)^{-1}A^T\mathbf{b} \;=\; \frac{1}{285}\begin{bmatrix} 21 & 3 \\ 2 & 14 \end{bmatrix}\begin{bmatrix} 1 & 3 & -2 \\ -1 & 2 & 4 \end{bmatrix}\begin{bmatrix} 4 \\ 1 \\ 3 \end{bmatrix} \;=\; \begin{bmatrix} 17/95 \\ 143/285 \end{bmatrix}$$

This is the least squares solution.
$\hat{\mathbf{b}} = A\hat{\mathbf{x}} = (-0.323,\ 1.540,\ 1.65)$. It may seem like $\hat{\mathbf{b}}$ is far from $\mathbf{b}$, $\|\hat{\mathbf{b}} - \mathbf{b}\| = 4.56$, but $\hat{\mathbf{b}}$ is still the closest vector to $\mathbf{b}$ in $\text{col}(A)$, and we used it, $\hat{\mathbf{b}}$, instead of $\mathbf{b}$ to solve the *consistent* system $A\mathbf{x} = \hat{\mathbf{b}}$ and found the solution $\hat{\mathbf{x}}$.

Example 2: $A = \begin{bmatrix} 3 & 2 & -1 \\ 1 & -4 & 3 \\ 1 & 10 & -7 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$. Again, $\mathbf{b} \notin \text{col}(A)$.

$$A^TA = \begin{bmatrix} 11 & 12 & -7 \\ 12 & 120 & -84 \\ -7 & -84 & 59 \end{bmatrix}, \quad A^T\mathbf{b} = \begin{bmatrix} 5 \\ 22 \\ -15 \end{bmatrix}$$

Here, $\det(A^TA) = 0$ so we have to solve the normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$ by row reduction to find $\hat{\mathbf{x}}$;

$$\begin{bmatrix} 11 & 12 & -7 & 5 \\ 12 & 120 & -84 & 22 \\ -7 & -84 & 59 & -15 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 0 & \frac{1}{7} & \frac{2}{7} \\ 0 & 1 & -\frac{5}{7} & \frac{13}{84} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The solution set is

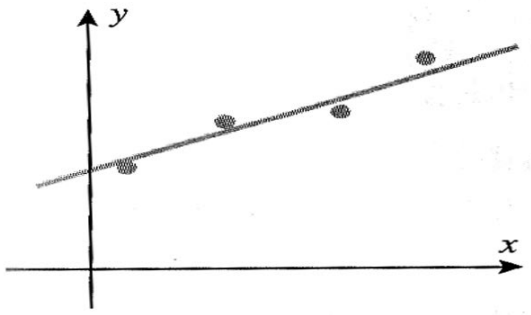
$$\hat{\mathbf{x}} \;=\; \mathbf{v}_0 + t\mathbf{v}_1, \quad \mathbf{v}_0 = (\frac{2}{7}, \frac{13}{84}, 0), \; \mathbf{v}_1 = (-\frac{1}{7}, \frac{5}{7}, 1)$$

and so there are many least squares solutions. For each of them we have that $A\hat{\mathbf{x}} = \hat{\mathbf{b}} = (5, 22, -15)$.

Examples for Computing Assignment 4; Start here!

Example 3: Finding the least squares curve to a set of data points in $\mathbf{R}^2$.

We are given a set of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where we are considering the x as inputs and the y as outputs; so, one input and one output. We ask: What is the relation between the inputs and outputs? That is, can we find a function $f(x)$ such that the data set is represented by $y = f(x)$? To begin addressing that question, we first look for a linear function $y = f(x) = a + bx$, whose graph (a striaght line) passes close to all the data points.



That is, we want to determine a and b so that the total (vertical) 'error' between the line $y = a + bx$ and the data points is smallest;

$$\text{residual error} \;=\; [y_1 - (a + bx_1)]^2 + [y_2 - (a + bx_2)]^2 + \;\cdots\; + \;[y_n - (a + bx_n)]^2$$

Notice that *if* all the data points did lie on the same line, $y = a + bx$, then

$$\begin{aligned} y_1 &= a + bx_1 \\ y_2 &= a + bx_2 \\ \vdots &\quad \vdots \\ y_n &= a + x_n \end{aligned}$$

which in matrix form would be a system of equations;

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad M\mathbf{v} = \mathbf{y}$$

In other words, this system of equations, $M\mathbf{v} = \mathbf{y}$, would be <u>consistent</u> since $\mathbf{y} \in \text{col}(M)$, and we could solve it to determine $\mathbf{v} = (a, b)$ and get the equation for the straight line.

However, if the data points do *not* lie along a straight line, then the system $M\mathbf{v} = \mathbf{y}$ is inconsistent, and so we look for a least squares solution. The line $y = a + bx$ we find this way is called the **least squares line of best fit** (linear regression) and if you plot this line with the date, it appears to be the 'best' fitting straight line to the data (passes 'closest' to all the data points, in the sense of minimizing the residual error above).

That is, we solve the normal equations $M^T M\mathbf{v} = M^T\mathbf{y}$ (here, M is playing the role of A above, $\mathbf{y}$ is playing the role of $\mathbf{b}$, and $\mathbf{v}$ the role of $\mathbf{x}$). If the x-coordinates of the data points are not all the same (which is usually the case), $M^T M$ will be invertible, and thus there will be a unique least squares solution;

$$\mathbf{v} = (M^T M)^{-1} M^T \mathbf{y}$$

Let's try it; four data points $(0, 1), (1, 3), (2, 4), (3, 4)$. Then

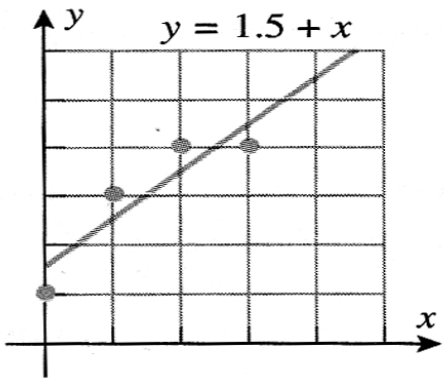
$$M = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}$$

$$M^T M = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}, \quad (M^T M)^{-1} = \frac{1}{10} \begin{bmatrix} 7 & -3 \\ -3 & 2 \end{bmatrix}$$

and so,

$$\mathbf{v} = (M^T M)^{-1} M^T \mathbf{y} = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$$

The least squares ('best fit') line to the data is $y = 1.5 + x$;



A least squares fit of a line to the data points.

The same procedure can be used to find least squares *curves* other than straight lines. For example, if you wanted to fit a quadratic equation (parabola) to your data points $(x_1, y_1), \ldots, (x_n, y_n)$, you would look for an equation $y = a_0 + a_1 x + a_2 x^2$ that 'best' fits the data. Let's try it; five data points $(0.10, -0.18), (0.20, 0.31), (0.30, 1.03), (0.40, 2.48), (0.50, 3.73)$.

In this case,

$$\begin{aligned} y_1 &= a_0 + a_1 x_1 + a_2 x_1^2 \\ y_2 &= a_0 + a_1 x_2 + a_2 x_2^2 \\ y_3 &= a_0 + a_1 x_3 + a_2 x_3^2 \\ \vdots &\quad \vdots \\ y_n &= a_0 + a_1 x_n + a_2 x_n^2 \end{aligned}$$

which in matrix form would be a system of equations;

$$M = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

and we would solve for $\mathbf{v}$ as above. For this set of data points we have,

$$M = \begin{bmatrix} 1 & 0.10 & 0.01 \\ 1 & 0.20 & 0.40 \\ 1 & 0.30 & 0.09 \\ 1 & 0.40 & 0.16 \\ 1 & 0.50 & 0.25 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} -0.18 \\ 0.31 \\ 1.03 \\ 2.48 \\ 3.73 \end{bmatrix}$$

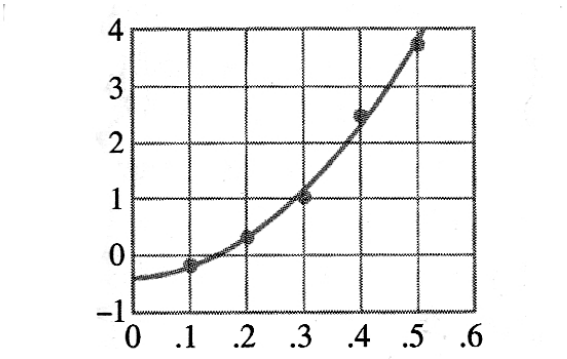
$$M^T M = \begin{bmatrix} 5 & 1.5 & 0.55 \\ 1.5 & 0.55 & 0.225 \\ 0.55 & 0.225 & 0.0979 \end{bmatrix}$$

and so,

$$\mathbf{v} = (M^T M)^{-1} M^T \mathbf{y} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -0.40 \\ 0.35 \\ 16.08 \end{bmatrix}$$

with our least squares quadratic

$$y = -0.40 + 0.35x + 16.08x^2$$



A least squares fit of a quadratic to the data points.

Example 4: Finding the least squares surface to a set of data points in $\mathbf{R}^3$

Now we have two inputs (x, y) and one output z; (x_i, y_i, z_i). So our data set is a collection of points in $\mathbf{R}^3$ and we wish to fit a *surface* through (or close to) this set of points. Again we ask: What is the relation between the inputs and outputs? That is, can we find a function $f(x, y)$ such that the data set is represented by $z = f(x, y)$? To begin addressing that question, we first look for a linear function $z = f(x, y) = a + bx + cy$, whose graph (a plane!) passes close to all the data points.

Notice that *if* all the data points did lie on the same plane, $z = a + bx + cy$, then

$$\begin{aligned} z_1 &= a + bx_1 + cy_1 \\ z_2 &= a + bx_2 + cy_2 \\ &\vdots \quad\quad \vdots \\ z_n &= a + bx_n + cy_n \end{aligned}$$

which in matrix form would be a system of equations;

$$M = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ & & \\ \vdots & \vdots & \vdots \\ & & \\ 1 & x_n & y_n \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \\ \vdots \\ \\ z_n \end{bmatrix}$$

This system is likely inconsistent (i.e., not all the points will lie on a plane), so we solve the normal equation. Suppose for the data set we find this solution,

$$\mathbf{v} = \left(M^T M\right)^{-1} M^T \mathbf{z} = \begin{bmatrix} -131.7 \\ 3.5 \\ 2.2 \end{bmatrix}$$

Then, our least squares plane to this data is (our 'model' of the data)

$$z = -131.7 + 3.5x + 2.2y$$

(Variations on a theme: Just in the case for Example 3, we could instead use *nonlinear* surfaces to fit the data. It would still turn out to be a least squares (linear!) problem solved by the normal equations. The modeler has to decide which type of surface to fit the data with!)