The Method of Least Squares

This is an important application of the orthogonality material to data analysis / modeling.

Setup: Suppose you have a matrix equation $Ax=b$ that is (probably) inconsistent. What is the

best approximate solution?

What does that even mean?

distance from $A\hat{x}$ to b

Def: $\hat{x}$ is a least squares solution of $Ax=b$ if $\|b-A\hat{x}\|$ is minimized over all vectors $\hat{x}$. This minimal value of $\|b-A\hat{x}\|$ is called the error.

In other words, $\hat{x}$ is a least squares solution $\iff$ $A\hat{x}$ is as close as possible to b.

Let's set $\hat{b}=A\hat{x}$ for the moment. We know

$$A\hat{x}=\hat{b} \text{ has a solution} \iff \hat{b}\in \text{Col}(A).$$

Since $\hat{b}=A\hat{x}$ is as close as possible to b, this means that $\hat{b}$ is the closest vector to b in $\text{Col}(A)$. We have a name for that!

$\hat{b}$ = the orthogonal projection of b onto $\text{Col}(A)$.

So a least-squares solution of $Ax=b$ is a solution of the consistent matrix equation $A\hat{x}=b_V$, $V=\text{Col}(A)$.

How do we compute $\hat{x}$?

→ Maybe we should compute b_V first? We could do this by solving the normal equation

$$A^TA\hat{x}=A^Tb$$

Then $b_V=A\hat{x}$ for any solution $\hat{x}$.

But that just means $\hat{x}$ is a solution of $A\hat{x}=b_V$, which is a least-squares solution of $Ax=b$!

Summary: These three things are the same:

Least-squares solutions of $Ax=b$	=	Solutions of $A\hat{x}=b_V$ ($V=\text{Col}(A)$)	=	Solutions of $A^TA\hat{x}=A^Tb$

How to Find the Least-Squares Solutions of $Ax=b$:

(1) Solve the normal equation $A^TA\hat{x}=A^Tb$.

Any solution $\hat{x}$ is a least-squares solution, and $b_V=A\hat{x}$ ($V=\text{Col}(A)$).

The error is the length of $b - A\hat{x} = b - b_V = b_{V^\perp}$:

$$\text{error} = \|b_{V^\perp}\|$$

(b – the b-vector we wanted; b_V – the b-vector we got)

NB: Minimizing $\|b_{V^\perp}\| \rightleftharpoons$ minimizing $\|b_{V^\perp}\|^2$

So saying that $A\hat{x}$ is as close as possible to b means:

The least-squares solution(s) minimize $\|b_{V^\perp}\|^2$

If $b_{V^\perp} = (a,b,c)$ then $\|b_{V^\perp}\|^2 = a^2+b^2+c^2$: these are the "squares" in "least squares".

Eg: Find the least-squares solution & error of $Ax=b$ for

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \quad b = \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}$$

We have to solve $A^TA\hat{x} = A^Tb$.

$$A^TA = \begin{pmatrix} \text{column} \\ \text{dot} \\ \text{products} \end{pmatrix} = \begin{pmatrix} 5 & 3 \\ 3 & 3 \end{pmatrix} \qquad A^Tb = \begin{pmatrix} 0 \\ 6 \end{pmatrix}$$

$$\left(\begin{array}{cc|c} 5 & 3 & 0 \\ 3 & 3 & 6 \end{array}\right) \xrightarrow{\text{ref}} \left(\begin{array}{cc|c} 1 & 0 & -3 \\ 0 & 1 & 5 \end{array}\right) \rightsquigarrow \hat{x} = \begin{pmatrix} -3 \\ 5 \end{pmatrix}$$

$$b_{V^\perp} = b - A\hat{x} = \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 5 \\ 3 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ 1 \end{pmatrix}$$

$$\text{error}^2 = \|b_{V^\perp}\|^2 = 1^2 + (-3)^2 + 1^2 = 6$$

[DEMO]

Eg: Find the least-squares solutions & error of $Ax=b$ for

$$A=\begin{pmatrix}1&-1&-1\\2&1&4\\1&-1&-1\end{pmatrix} \qquad b=\begin{pmatrix}1\\1\\1\end{pmatrix}$$

We have to solve $A^TA\hat{x}=A^Tb$.

$$A^TA=\begin{pmatrix}\text{column}\\\text{dot}\\\text{products}\end{pmatrix}=\begin{pmatrix}6&0&6\\0&3&6\\6&6&18\end{pmatrix} \qquad A^Tb=\begin{pmatrix}4\\-1\\2\end{pmatrix}$$

$$\left(\begin{array}{ccc|c}6&0&6&4\\0&3&6&-1\\6&6&18&2\end{array}\right) \xrightarrow{\text{ref}} \left(\begin{array}{ccc|c}1&0&1&2/3\\0&1&2&-1/3\\0&0&0&0\end{array}\right) \xrightarrow{\text{PVF}} \hat{x}=\begin{pmatrix}2/3\\-1/3\\0\end{pmatrix}+x_3\begin{pmatrix}-1\\-2\\1\end{pmatrix}$$

In this case there are infinitely many least-squares solutions.

Error: $b_{\perp}=b-A\hat{x}$ for any solution $\hat{x}$.

Let's take $\hat{x}=\begin{pmatrix}2/3\\-1/3\\0\end{pmatrix} \rightsquigarrow A\hat{x}=\begin{pmatrix}1\\1\\1\end{pmatrix} \rightsquigarrow b_{\perp}=0$

So the error is zero — that means $A\hat{x}=b$, ie, $Ax=b$ was consistent after all!

Observation 1:

> $Ax=b$ has a unique least-squares solution $\iff$ A has FCR

That's because least-squares solutions are just normal solutions of $Ax=b_{\perp}$.

Otherwise $Ax=b$ has infinitely many least-squares solutions $\hat{x}$. This means $\|b-A\hat{x}\|$ is minimized for any such $\hat{x}$, since

$$b_V = A\hat{x} \text{ for any least-squares solution } \hat{x}.$$

Of course, there can't be zero least-squares solutions! $A\hat{x}=b_V$ is always consistent.

Observation 2:

> If $Ax=b$ is consistent then
> $$\begin{pmatrix}\text{least-squares}\\ \text{solutions of } Ax=b\end{pmatrix} = \begin{pmatrix}\text{ordinary solutions}\\ \text{of } Ax=b\end{pmatrix}$$

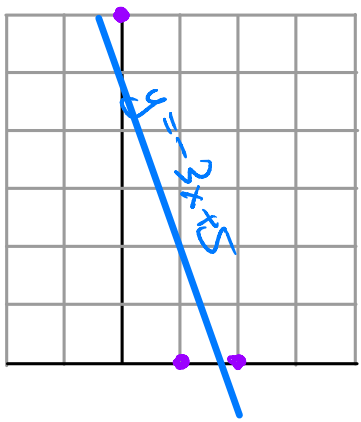
If $Ax=b$ is consistent then $b \in V = \text{Col}(A)$, so $b=b_V$ (it's the closest vector to itself in V), so

$$\begin{pmatrix}\text{solutions of}\\ A\hat{x}=b_V\end{pmatrix} = \begin{pmatrix}\text{solutions of}\\ Ax=b\end{pmatrix}$$

Least Squares and Data Modeling

Eg (linear regression):

Find the best-fit line $y = Cx + D$ through the data points $\begin{pmatrix}0\\6\end{pmatrix}$ $\begin{pmatrix}1\\0\end{pmatrix}$ $\begin{pmatrix}2\\0\end{pmatrix}$



> If $\begin{pmatrix}0\\6\end{pmatrix}$ lies on $y = Cx + D$ then substituting $x=0, y=6$ would give $6 = 0 \cdot C + D$. Likewise for the other data points.

So we want to solve:

$\begin{pmatrix}0\\6\end{pmatrix}$: $6 = C \cdot 0 + D$
$\begin{pmatrix}1\\0\end{pmatrix}$: $0 = C \cdot 1 + D$
$\begin{pmatrix}2\\0\end{pmatrix}$: $0 = C \cdot 2 + D$

in the unknowns C & D.

Matrix equation: $Ax = b$ for

$$A = \begin{pmatrix}0 & 1\\1 & 1\\2 & 1\end{pmatrix} \quad b = \begin{pmatrix}6\\0\\0\end{pmatrix} \quad x = \begin{pmatrix}C\\D\end{pmatrix}$$

NB: the data points are not collinear, so there is no exact solution! (perhaps measurement error...)

We found the least-squares solution a few pages ago:

$\hat{x} = \begin{pmatrix}-3\\5\end{pmatrix} = \begin{pmatrix}C\\D\end{pmatrix} \rightsquigarrow$ best-fit line $y = -3x + 5$

[DEMO]

Important question: What exactly did we minimize?

The answer is always $\|b - A\hat{x}\|^2 = \|b_{v^\perp}\|^2$, but where can I see that in the context of the original problem?

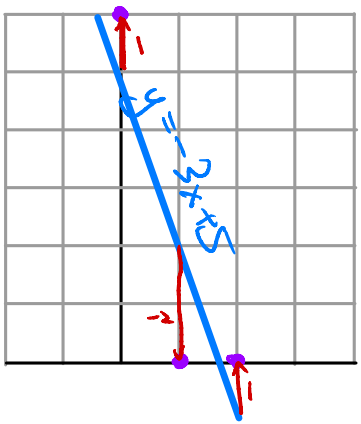
$$b = \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} y\text{-values of the data points} \\ \text{(the } y\text{-values we wanted)} \end{pmatrix}$$

$$A\hat{x} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} -3 \\ 5 \end{pmatrix} = \begin{pmatrix} -3\cdot 0 + 5 \\ -3\cdot 1 + 5 \\ -3\cdot 2 + 5 \end{pmatrix} = \begin{pmatrix} y\text{-values of } y = -3x+5 \\ \text{evaluated at the } x\text{-values} \\ \text{of the data points} \\ \text{(the } y\text{-values we got)} \end{pmatrix}$$

$$b_{v^\perp} = b - A\hat{x} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} \text{vertical distances} \\ \text{from } y = -3x+5 \text{ to} \\ \text{the data points} \end{pmatrix}$$

$$= \begin{pmatrix} \text{differences between} \\ \text{expected and} \\ \text{observed values} \end{pmatrix}$$



We minimized the sum of the squares of the vertical distances (= the error^2).

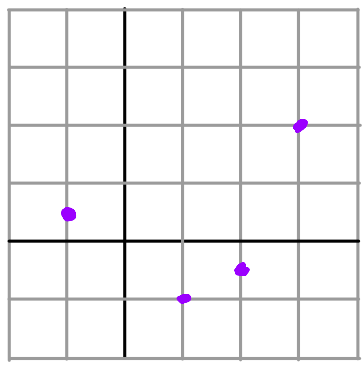
What if our data aren't supposed to lie on a line?

Eg (best-fit parabola):

Find the best-fit parabola $y = Bx^2 + Cx + D$ through the data points

$$\begin{pmatrix} -1 \\ 1/2 \end{pmatrix} \quad \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 2 \\ -1/2 \end{pmatrix} \quad \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

As before, we substitute the x- & y-values of our data points into our equation:



$$\begin{pmatrix} -1 \\ 1/2 \end{pmatrix}: \quad \frac{1}{2} = B(-1)^2 + C(-1) + D$$

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}: \quad -1 = B(1)^2 + C(1) + D$$

$$\begin{pmatrix} 2 \\ -1/2 \end{pmatrix}: \quad -\frac{1}{2} = B(2)^2 + C(2) + D$$

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix}: \quad 2 = B(3)^2 + C(3) + D$$

Matrix equation: $Ax = b$ for

$$A = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} 1/2 \\ -1 \\ -1/2 \\ 2 \end{pmatrix} \qquad x = \begin{pmatrix} B \\ C \\ D \end{pmatrix}$$

Let's find the least-squares solution.

$$A^TA = \begin{pmatrix} 99 & 35 & 15 \\ 35 & 15 & 5 \\ 15 & 5 & 4 \end{pmatrix} \qquad A^Tb = \begin{pmatrix} 31/2 \\ 7/2 \\ 1 \end{pmatrix}$$

$$\left(\begin{array}{ccc|c} 99 & 35 & 15 & 31/2 \\ 35 & 15 & 5 & 7/2 \\ 15 & 5 & 4 & 1 \end{array}\right) \xrightarrow{\text{rref}} \left(\begin{array}{ccc|c} 1 & 0 & 0 & 53/88 \\ 0 & 1 & 0 & -379/440 \\ 0 & 0 & 1 & -41/44 \end{array}\right)$$

$$\hat{x} = \begin{pmatrix} 53/88 \\ -379/440 \\ -41/44 \end{pmatrix} = \begin{pmatrix} B \\ C \\ D \end{pmatrix} \rightsquigarrow y = \frac{53}{88}x^2 - \frac{379}{440}x - \frac{41}{44}$$

[DEMO]

Question: What did we minimize? Where can we see the error^2? We always minimize $\|b - A\hat{x}\|^2$.

$$b = \begin{pmatrix} 1/2 \\ -1 \\ -1/2 \\ 2 \end{pmatrix} = \begin{pmatrix} y\text{-values of the data points} \\ \text{(the } y\text{-values we wanted)} \end{pmatrix}$$

$$A\hat{x} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} \begin{pmatrix} 53/88 \\ -379/440 \\ -41/44 \end{pmatrix} = \begin{pmatrix} \frac{53}{88}(-1)^2 - \frac{379}{440}(-1) - \frac{41}{44} \\ \frac{53}{88}(1)^2 - \frac{379}{440}(1) - \frac{41}{44} \\ \frac{53}{88}(2)^2 - \frac{379}{440}(2) - \frac{41}{44} \\ \frac{53}{88}(3)^2 - \frac{379}{440}(3) - \frac{41}{44} \end{pmatrix}$$

$$= \begin{pmatrix} y\text{-values of } y = \frac{53}{88}x^2 - \frac{379}{440}x - \frac{41}{44} \\ \text{evaluated at the } x\text{-values} \\ \text{of the data points} \\ \text{(the } y\text{-values we got)} \end{pmatrix}$$

$$b_{\perp} = b - A\hat{x} = \begin{pmatrix} \text{vertical distances} \\ \text{from the graph to} \\ \text{the data points} \end{pmatrix}$$

The error^2 has the same interpretation as in the previous example.

The previous two examples were very similar. In fact, if you want to find a best-fit function of the form

$$y = Af(x) + Bg(x) + Ch(x) + \cdots$$

where $A, B, C, \ldots$ are unknowns and $f, g, h, \ldots$ are any functions, then you do the same thing: plug in the x- and y-values of your data points $\rightsquigarrow$ linear equations in $A, B, C, \ldots$

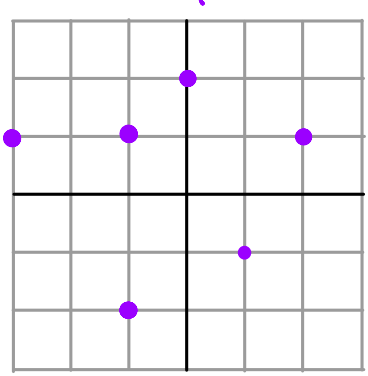
Eg (best-fit trigonometric function): [DEMO] [ILA]

This "based-on-a-true-story" example of Gauss was in L1. It has a somewhat different character.

Eg (best-fit ellipse):

An asteroid has been observed at these coordinates:

$\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ $\begin{pmatrix} -1 \\ -2 \end{pmatrix}$ $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$ $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$



Question: What is the most likely orbit? Will the asteroid crash into Earth?

Fact (Kepler): The orbit is an ellipse.

Equation for an Ellipse:

$$x^2 + By^2 + Cxy + Dx + Ey + F = 0$$

Substitute our x- & y-values of my data points:

$\begin{pmatrix}0\\2\end{pmatrix}$: $(0)^2 + (2)^2B + (0)(2)C + (0)D + (2)E + F = 0$

$\begin{pmatrix}2\\1\end{pmatrix}$: $(2)^2 + (1)^2B + (2)(1)C + (2)D + (1)E + F = 0$

$\begin{pmatrix}1\\-1\end{pmatrix}$: $(1)^2 + (-1)^2B + (1)(-1)C + (1)D + (-1)E + F = 0$

$\begin{pmatrix}-1\\-2\end{pmatrix}$: $(-1)^2 + (-2)^2B + (-1)(-2)C + (-1)D + (-2)E + F = 0$

$\begin{pmatrix}-3\\1\end{pmatrix}$: $(-3)^2 + (1)^2B + (-3)(1)C + (-3)D + (1)E + F = 0$

$\begin{pmatrix}-1\\1\end{pmatrix}$: $(-1)^2 + (1)^2B + (-1)(1)C + (-1)D + (1)E + F = 0$

the x^2 summands are constants!
Move them to the right of the $=$ sign.

This is a matrix equation $Ax=b$ for

$$A=\begin{pmatrix}4&0&0&2&1\\1&2&2&1&1\\1&-1&1&-1&1\\4&2&-1&-2&1\\1&-3&-3&1&1\\1&-1&-1&1&1\end{pmatrix} \quad b=-\begin{pmatrix}0\\4\\1\\1\\9\\1\end{pmatrix} \quad x=\begin{pmatrix}B\\C\\D\\E\\F\end{pmatrix}$$

The least-squares solution is

$$\hat{x}=\left(\frac{405}{266}, -\frac{89}{133}, \frac{201}{133}, -\frac{123}{266}, -\frac{687}{133}\right)$$

$$\rightsquigarrow \quad x^2+\frac{405}{266}y^2-\frac{89}{133}xy+\frac{201}{133}x-\frac{123}{266}y-\frac{687}{133}=0$$

[DEMO]

Question: What did we minimize this time?

Always $\|b-A\hat{x}\|^2$, or $\|-b+A\hat{x}\|^2$.

$$-b+A\hat{x} = \begin{pmatrix} 0^2 \\ 2^2 \\ 1^2 \\ (-1)^2 \\ (-3)^2 \\ (-1)^2 \end{pmatrix} + \begin{pmatrix} 4 & 0 & 0 & 2 & 1 \\ 1 & 2 & 2 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 4 & 2 & -1 & -2 & 1 \\ 1 & -3 & -3 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 405/266 \\ -89/133 \\ 201/133 \\ -123/266 \\ -687/133 \end{pmatrix}$$

$$= \begin{pmatrix} (0)^2 + \frac{405}{266}(2)^2 - \frac{89}{133}(0)(2) + \frac{201}{133}(0) - \frac{123}{266}(2) - \frac{687}{133} \\ (2)^2 + \frac{405}{266}(1)^2 - \frac{89}{133}(2)(1) + \frac{201}{133}(2) - \frac{123}{266}(1) - \frac{687}{133} \\ (1)^2 + \frac{405}{266}(-1)^2 - \frac{89}{133}(1)(-1) + \frac{201}{133}(1) - \frac{123}{266}(-1) - \frac{687}{133} \\ (-1)^2 + \frac{405}{266}(-2)^2 - \frac{89}{133}(-1)(-2) + \frac{201}{133}(-1) - \frac{123}{266}(-2) - \frac{687}{133} \\ (-3)^2 + \frac{405}{266}(1)^2 - \frac{89}{133}(-3)(1) + \frac{201}{133}(-3) - \frac{123}{266}(1) - \frac{687}{133} \\ (-1)^2 + \frac{405}{266}(1)^2 - \frac{89}{133}(-1)(1) + \frac{201}{133}(-1) - \frac{123}{266}(1) - \frac{687}{133} \end{pmatrix}$$

$$= \begin{pmatrix} \text{what you get by substituting the } x\text{- and} \\ y\text{-values of the data points into} \\ x^2 + \frac{405}{266}y^2 - \frac{89}{133}xy + \frac{201}{133}x - \frac{123}{266}y - \frac{687}{133} \end{pmatrix}$$

We wanted our data points to satisfy

$$x^2 + \frac{405}{266}y^2 - \frac{89}{133}xy + \frac{201}{133}x - \frac{123}{266}y - \frac{687}{133} = 0$$

so in this case we're minimizing the distance from 0.

Alternatively, we're trying to fit the data points

$$\begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \begin{pmatrix} -1 \\ -2 \\ 0 \end{pmatrix} \begin{pmatrix} -3 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

to the equation

$$z = x^2 + By^2 + Cxy + Dx + Ey + F$$

so we're minimizing the sum of the squares of the vertical distances from the graph to the data points again.

[DEMO]

Upshot: Least squares always minimizes $\|b - A\hat{x}\|^2$; it's up to you to interpret what that means in the context of the problem you're trying to solve.