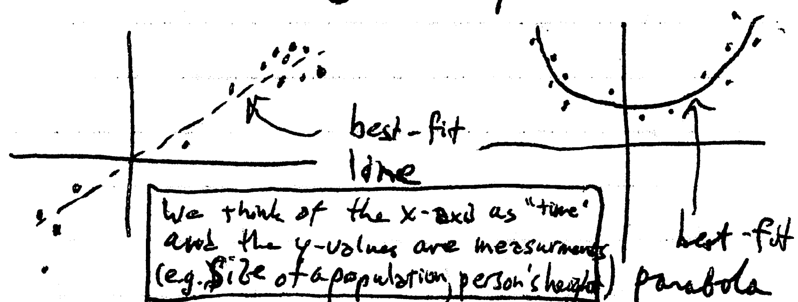
An Application of Orthogonality: Least Squares Approximation

Given a set of data points in $\mathbb{R}^2$, one often wants to find a line, or curve, which approximates them as closely as possible.



To find such a best-fit curve, we need to make the notion precise: how do we decide which curve is best?

and has good statistical properties...

Def'n: Given a set of points $(x_1,y_1),\dots,(x_k,y_k)\in\mathbb{R}^2$, the best-fit line $y=mx+b$ is the line minimizing the sum of the $\boxed{\text{squares}}$ of the distances from y_i to mx_i+b: $\sum_{i=1}^{k}\left(y_i-(mx_i+b)\right)^2$.

[using squares here makes the problem tractable]

This distance measures how far off our actual measurements are from the approximating line $y=mx+b$.

Observation: If we write $u=\begin{pmatrix} b \\ m \end{pmatrix}$ and $A=\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix}$, then the vector

Au gives the y-coords mx_i+b on the line $y=mx+b$:

$$\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix}\begin{pmatrix} b \\ m \end{pmatrix} = \begin{pmatrix} b+x_1m \\ \vdots \\ b+x_km \end{pmatrix}.$$

So minimizing $\sum\left(y_i-(mx_i+b)\right)^2$ is the same as minimizing $|\vec{y}-Au|=\sqrt{\sum_i\left(y_i-(b+x_im)\right)^2}$.

So we need to find $\vec{u}\in\mathbb{R}^2$ s.t. $|y-Au|$ is as small as possible.

Here Au is a vector in the range of A, so we are just looking for a vector in range(A) as close to y as possible:

Theorem 8 (Section 4.9): If $H\subseteq\mathbb{R}^n$ is any subspace, and $\vec{y}\in\mathbb{R}^n$ any vector, then $\text{proj}_H(\vec{y})$ is the closest point in H to $\vec{y}$:

$$|\vec{y}-\text{proj}_H(\vec{y})|<|\vec{y}-\vec{h}|$$

for any other $\vec{h}\in H$.

This means that we need to find $u\in\mathbb{R}^2$ such that

$$Au=\text{proj}_{\text{range}(A)}(\vec{y})$$

How can we find $\text{proj}_{\text{range}(A)}(\vec{y})$?

<u>Lemma</u>: The vector $\text{proj}_H(\vec{y})$ is the unique vector in H s.t.

$$\vec{h}\cdot\left(\vec{y} - \text{proj}_H(\vec{y})\right) = 0$$

for all $\vec{h} \in H$.

<u>Pf</u>: Choose an o.n. basis $\{h_1, \dots, h_m\}$ for H. Then by Thm 4 (Section 4.9)

$$h\cdot\left(\vec{y} - \text{proj}_H(\vec{y})\right) = \left(\sum (h_i\cdot h) h_i\right)\cdot\left(y - \text{proj}_H(\vec{y})\right)$$

$$= \left(\sum (h_i\cdot h) h_i\right)\left(y - \sum (y_i\cdot h_i) h_i\right) \quad \underbrace{h_i\cdot h_j}_{\substack{0 \text{ if } i\neq j,\\ 1 \text{ else}}}$$

$$= \sum (h_i\cdot h)(h_i\cdot y) - \sum_i\sum_j (h_i\cdot h)(h_j\cdot y)\, h_i\cdot h_j$$

$$= \sum (h_i\cdot h)(h_i\cdot y) - \sum_i (h_i\cdot h)(h_i\cdot y) = 0.$$

On the other hand, choose an o.n. basis $\{h_1, \dots, h_m, v_{m+1}, \dots, v_n\}$ for $\mathbb{R}^n$ extending our basis for H. Then if $\bar{y} \in H$ also satisfies $(y - \bar{y})\cdot h = 0 \ \forall\, h \in H$, we have

$$y - \bar{y} = \sum ((y-\bar{y})\cdot h_i) h_i + \sum ((y-\bar{y})\cdot\vec{v}_i)\,\vec{v}_i = \sum ((y-\bar{y})\cdot\vec{v}_i)\,\vec{v}_i$$

Thus

$$\bar{y} = y - \sum ((y-\bar{y})\cdot\vec{v}_i)\,\vec{v}_i$$

$$= \sum (y_i\cdot \vec{h}_i)\vec{h}_i + \cancel{\sum (y_i\cdot\vec{v}_i)\vec{v}_i}$$
$$\quad - \cancel{\sum (y\cdot\vec{v}_i)\cdot\vec{v}_i} + \sum (\bar{y}\cdot\vec{v}_i)\cdot\vec{v}_i$$

$$= \text{proj}_H(y) + \sum (\bar{y}\cdot v_i)\cdot v_i$$

But $\bar{y}\cdot v_i = 0$ b/c $\bar{y} \in H$:
$\bar{y} = \sum c_i h_i$ for some c_i, so
$\bar{y}\cdot v_i = \sum c_i (h_i\cdot v_i) = 0$. □

Note that this Lemma implies that $\text{proj}_H(y)$ doesn't depend on the chosen o.n. basis $\{h_1, \dots, h_m\}$ for Y.

Now recall that we want to find

$$\text{proj}_{\text{range}(A)}(y),$$

where $A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$ (the pts (x_i, y_i) are our data points).

The lemma says we just need a vector $\bar{y} \in$ range (A) s.t.

$$(Au)\cdot(y-\bar{y}) = 0 \text{ for all } u \in \mathbb{R}^2.$$

(i.e. $y-\bar{y}$ is orthogonal to the range of A).

Letting $\bar{y} = A\bar{u}$, we can write the above scalar product in matrix form:

$$Au\cdot(y-A\bar{u}) = (Au)^t(y-A\bar{u})$$

$$\overset{\text{old HW exercise}}{=} (u^tA^t)(y-A\bar{u})$$

$$= u^t(A^ty - A^tA\bar{u})$$

$$= u\cdot(A^ty - A^tA\bar{u}).$$

So if this is zero for every $u \in \mathbb{R}^2$, then the vector

$$A^ty - A^tA\bar{u}$$

must be zero, i.e. $A^tA\bar{u} = A^ty$.

Claim: A^tA is invertible so long as not all the x_i are identical.

Pf: Since not all the x_i are the same, the cols of $A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix}$ are lin. indep, so $N(A) = \{0\}$.

Since A^tA is square we just need to show that $N(A^tA) = 0$.

Say $A^tAz = 0$. Then Az is orthogonal to the rows of A^t, i.e. the columns of A:

$$\begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_k \end{pmatrix}(Az) = \begin{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\cdot Az \\ \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}\cdot Az \end{pmatrix}.$$

Let $\{a_1, a_2\}$ be an o.n. basis for the range of A. Then $Az = \sum c_i a_i$ for some c_i. Since Az is orthog. to the columns of A, $Az\cdot a_j = 0$ as well. So

$$\left(\sum c_i a_i\right)\cdot a_j = c_j, \text{ meaning } c_1 = c_2 = 0.$$

So $Az = 0$, and now $N(A) = \{0\} \Rightarrow z = 0$. □

$\boxed{\text{We now assume at least 2 } x_i \text{ are distinct.}}$

Since A^tA is inv'ble, we now have

$$A^tA\bar{u} = A^ty$$

$$\Rightarrow \quad \bar{u} = (A^tA)^{-1}A^ty$$

[Note: $(A^tA)^{-1} \neq A^{-1}(A^t)^{-1}$ b/c A^{-1} and $(A^t)^{-1}$ do not exist.]

So we can easily find the vector $\bar{u} = \begin{pmatrix} \bar{b} \\ \bar{m} \end{pmatrix}$ for which

$$y = \bar{m}x + \bar{b}$$

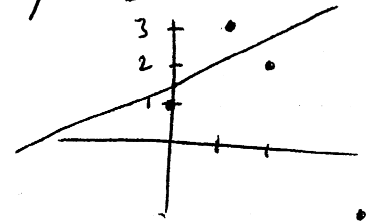
is the best-fit line for the data pts (x_i, y_i).

Example: If our data pts are $(0,1), (1,3), (2,2)$ then $A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$ and $A^tA = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}$.

So $\bar{u} = (A^tA)^{-1}A^t\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 5 & -3 \\ -3 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$
$= \frac{1}{6}\begin{pmatrix} 5 & -3 \\ -3 & 3 \end{pmatrix}\begin{pmatrix} 6 \\ 7 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 9 \\ 3 \end{pmatrix}$
$= \begin{pmatrix} 3/2 \\ 1/2 \end{pmatrix}$

Hence the best-fit line is

$$y = \tfrac{1}{2}x + \tfrac{3}{2};$$



Some data sets may be better approximated by a polynomial of higher degree, say

$$p(x) = a_0 + a_1x + \cdots + a_nx^n.$$

Say our data pts are $(x_1, y_1), \ldots, (x_k, y_k)$. Letting $A = \begin{pmatrix} 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_k & & x_k^n \end{pmatrix}$, we have

$$A\begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_k & & x_k^n \end{pmatrix}\begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \Sigma a_i x_1^i \\ \vdots \\ \Sigma a_i x_k^i \end{pmatrix}$$

So minimizing $\sum_{i=1}^{k}(p(x_i) - y_i)^2$ is the same as minimizing

$$\|Au - \vec{y}\|$$

over all $\vec{u} = \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^n$.

The minimum is again given by

$$\bar{u} = (A^t A)^{-1} A^t y.$$

All that we used in this argument was that $N(A) = 0$, or that $A^t A$ is invertible.

<u>Lemma</u>: $N\begin{pmatrix} 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_k & & x_k^n \end{pmatrix} = 0$ as long as at least $n+1$ of the x_i are distinct.

<u>Pf</u>: Say $\vec{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$. Then $A\vec{c} = \begin{pmatrix} c_1 + c_2 x_1 + \cdots + c_n x_1^n \\ \vdots \\ c_1 + c_2 x_k + \cdots + c_n x_k^n \end{pmatrix}$

So $\vec{c} \in N(A) \iff x_1, \ldots, x_k$ are roots of the poly. $p(x) = \sum_{i=1}^{n} c_i x^i$.

But this is a degree n polynomial and at least $n+1$ of the x_i are distinct. So if all the x_i are roots, then $p(x) = 0$, i.e. $c_1 = \cdots = c_n = 0$. Hence $N(A) = \{0\}$. $\square$

With the same data pts as before, the best-fit quadratic is $a_0 + a_1 x + a_2 x^2$ where $\bar{u} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}$ is given by $\bar{u} = (A^t A)^{-1} A^t y$ where

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1^2 \\ 1 & 2 & 2^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

Since A, A^t are square, we can compute

$$(A^t A)^{-1} = A^{-1} (A^t)^{-1}$$

more easily.

Row reducing shows

$$(A \mid I) \sim \left(I \,\middle|\, \begin{matrix} 1 & 0 & 0 \\ -3/2 & 2 & -1/2 \\ 1/2 & -1 & 1/2 \end{matrix} \right)$$

$$(A^t \mid I) \sim \left(I \,\middle|\, \begin{matrix} 1 & -3/2 & 1/2 \\ 0 & 2 & -1 \\ 0 & -1/2 & 1/2 \end{matrix} \right)$$

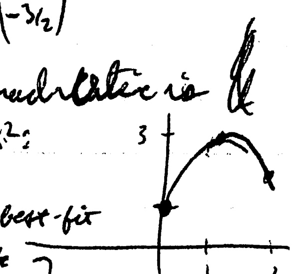
so $(A^t A)^{-1} = \begin{pmatrix} 1 & -3/2 & 1/2 \\ -3/2 & 13/2 & -3 \\ 1/2 & -3 & 3/2 \end{pmatrix}$,

and $\bar{u} = (A^t A)^{-1} A^t \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$

$$= \begin{pmatrix} 1 & 0 & 0 \\ -3/2 & 2 & -1/2 \\ 1/2 & -1 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3.5 \\ -3/2 \end{pmatrix}$$

and the best fit quadratic is

$$y = 1 + 3.5x + (-3/2)x^2;$$



[With just 3 data pts, the best-fit quadratic passes through all of them.]