OLS example 1: fitting a line

Suppose we want to fit the best line through the three points:

- ($x_1 = 1.1$, $y_1 = -0.2$)
- ($x_2 = 0.5$, $y_2 = -0.3$)
- ($x_3 = 0.1$, $y_3 = 1.2$)

Will define the line as

$$y = \beta_1 + \beta_2 x$$

and collect the coefficients into the vector $\vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

Substituting the three data points gives the system:

- $y_1 = \beta_1 + \beta_2 x_1$
- $y_2 = \beta_1 + \beta_2 x_2$
- $y_3 = \beta_1 + \beta_2 x_3$

or

- $-0.2 = \beta_1 + 1.1\beta_2$
- $-0.3 = \beta_1 + 0.5\beta_2$
- $1.2 = \beta_1 + 0.1\beta_2$

The same system $A\vec{\beta} = [\, \vec{a}_1 \mid \vec{a}_2 \,]\, \vec{\beta} = \vec{b}$ in matrix form:

$$\begin{bmatrix} 1 & 1.1 \\ 1 & 0.5 \\ 1 & 0.1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -0.2 \\ -0.3 \\ 1.2 \end{bmatrix}$$

This system has

- 3 equations

⇔

data space = $\mathbb{R}^3$

• two unknowns β_1 and β_2

⇔

dimension of parameter space or column space(A) = 2

↓

$\vec{b}$ likely lies outside parameter space

exact solution likely does not exist

The image below shows

• $\vec{a}_1$ as the green arrow corresponding to the constant term in $y = \beta_1 + \beta_2 x$

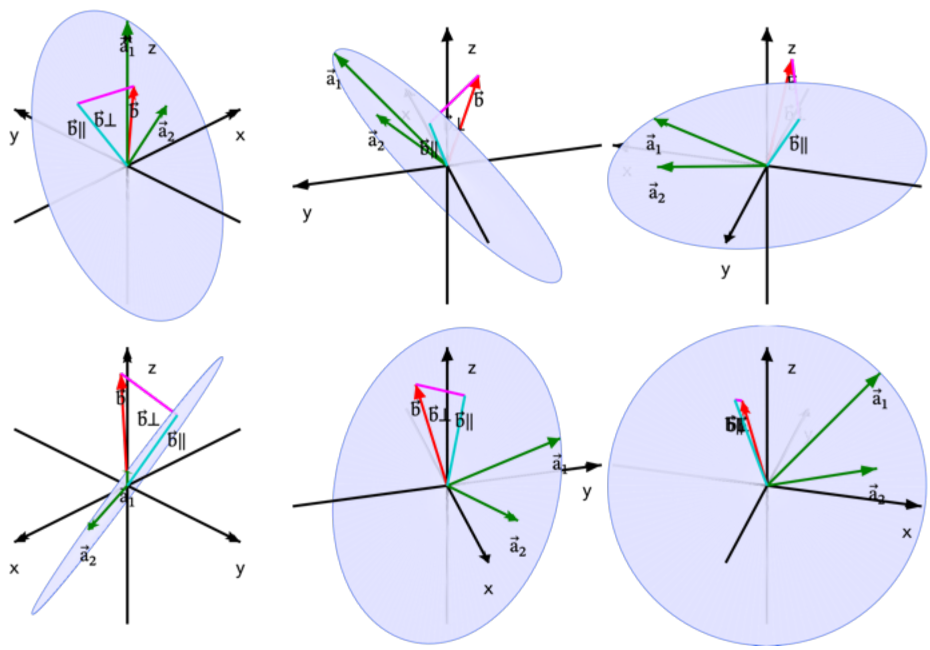
• $\vec{a}_2$ as the green arrow corresponding to the linear term in $y = \beta_1 + \beta_2 x$

• the span of $\vec{a}_1$ & $\vec{a}_2$ as a plane (allowed by linear independence of $\vec{a}_1$ & $\vec{a}_2$)

• $\vec{b}$ as the red arrow outside the span of $\vec{a}_1$ & $\vec{a}_2$

↓

an exact solution indeed does not exist



One possible way to compute the least-squares solution is to solve the normal equations derived in the 'Four subspaces' chapter

$A^T A \vec{\beta} = A^T \vec{b}$

↓

$$\vec{\beta} = (A^{T} A)^{-1} A^{T} \vec{b}$$

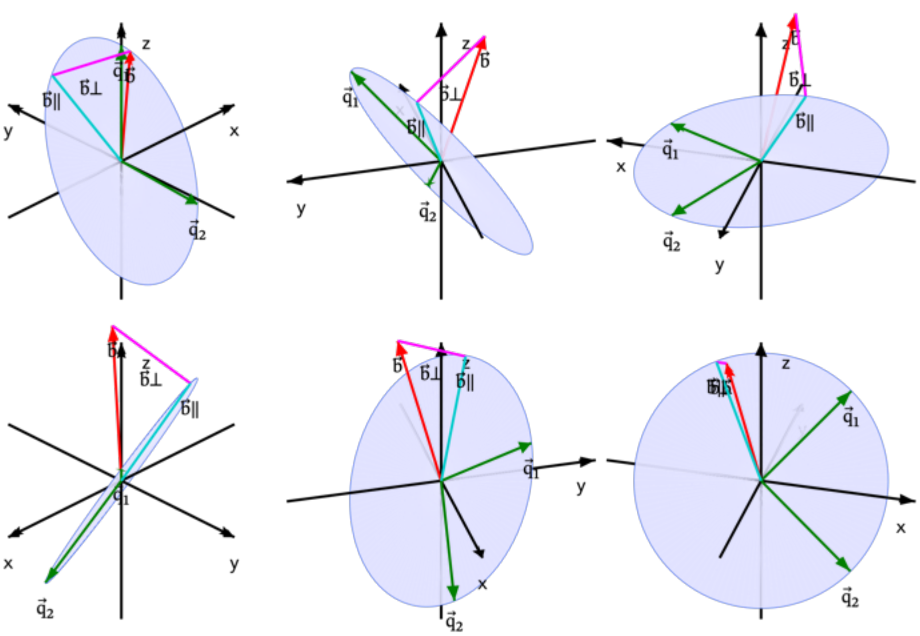
Since forming and inverting $(A^{T} A)$ may amplify numerical error,
we orthogonalize the column space of A and solve the triangular system $R \vec{\beta} = Q^{T} \vec{b}$

Converting $\vec{a}_1$ & $\vec{a}_2$ into an orthonormal basis yields

$$Q = [\, \vec{q}_1 \mid \vec{q}_2 \,] = \begin{bmatrix} \approx 0.577 & \approx 0.749 \\ \approx 0.577 & \approx -0.094 \\ \approx 0.577 & \approx -0.656 \end{bmatrix}$$

Same image with $\vec{q}_1$ & $\vec{q}_2$ in place of $\vec{a}_1$ & $\vec{a}_2$:

- $\vec{q}_1$ is the same as normalized $\vec{a}_1$
- $\vec{q}_2$ is orthogonal to $\vec{q}_1$
- the plane spanned by $\vec{q}_1$ & $\vec{q}_2$ is the same as the plane spanned by $\vec{a}_1$ & $\vec{a}_2$
- $\vec{b}$ is unchanged
- the cyan vector $\vec{b}_{||} = Q Q^{T} \vec{b}$ or the projection of $\vec{b}$ onto $\text{span}(\vec{q}_1, \vec{q}_2)$



As shown on the previous page, we compute

upper-triangular matrix $R = Q^T A = \begin{bmatrix} \approx 1.732 & \approx 0.981 \\ 0 & \approx 0.712 \end{bmatrix}$

and solve $R\vec{\beta} = Q^T\vec{b}$ without inverting Q:

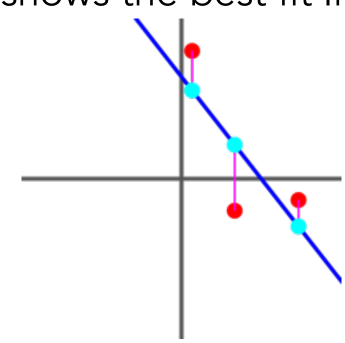
- $\beta_1 \approx 0.957$ (constant term in $y = \beta_1 + \beta_2 x$)
- $\beta_2 \approx -1.276$ (linear term in $y = \beta_1 + \beta_2 x$)

- $\vec{b}$ represents the three actual 'y' values of the dataset
- Each vector in the plane $\text{span}(\vec{a}_1, \vec{a}_2) = \text{span}(\vec{q}_1, \vec{q}_2)$ represents the three predicted values of some line $y = \beta_1 + \beta_2 x$ evaluated at the three sample points
- The cyan vector $\vec{b}\|$ represents the predicted values for the best fit line
- The 3D magenta segment represents the residual vector

$$\vec{b}\perp = \vec{b} - \vec{b}\|$$
$$= \vec{b} - Q\,Q^T\,\vec{b}$$
$$= \vec{b} - A\,\vec{\beta}$$

- The sum of squared residuals $|\,\vec{b} - A\,\vec{\beta}\,|^2 \approx 0.581$ is the quantity the algorithm minimizes

Image below shows the best fit line $y = \beta_1 + \beta_2 x$



- red dots are the 3 input values

- cyan dots are the 3 values predicted by the best line
- magenta lines are individual residuals or individual entries of $\vec{b}\perp$

- The computation approach worked because $\vec{a}_1$ & $\vec{a}_2$ are linearly independent and the plane exists
- To depict it as a 3D image, we are limited to
  - 3 data points or data space = $\mathbb{R}^3$
  - 2 parameters in $y = \beta_1 + \beta_2 x$ or dimension of column space = 2

Next, we will show that the same least-squares framework can fit a nonlinear function

---

OLS example 2: fitting a parabola with 2 parameters

Suppose we want to fit a nonlinear function
through the same set of points that we used for linear regression:

- ($x_1 = 1.1$, $y_1 = -0.2$)
- ($x_2 = 0.5$, $y_2 = -0.3$)
- ($x_3 = 0.1$, $y_3 = 1.2$)

In order to show the 3D image, we are limited to 3 points and 2 parameters,
so we will define the function as a parabola with
a constant term and a quadratic term:

$$y = \beta_1 + \beta_2 x^2$$

and collect the coefficients into the vector $\vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

Substituting the three data points gives the system:

- $y_1 = \beta_1 + \beta_2 x_1^2$
- $y_2 = \beta_1 + \beta_2 x_2^2$
- $y_3 = \beta_1 + \beta_2 x_3^2$

or

- $-0.2 = \beta_1 + 1.21\beta_2$
- $-0.3 = \beta_1 + 0.25\beta_2$
- $1.2 = \beta_1 + 0.01\beta_2$

The same system $A\vec{\beta} = [\, \vec{a}_1 \mid \vec{a}_2 \,]\, \vec{\beta} = \vec{b}$ in matrix form:

$$\begin{bmatrix} 1 & 1.21 \\ 1 & 0.25 \\ 1 & 0.01 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -0.2 \\ -0.3 \\ 1.2 \end{bmatrix}$$

As in the previous example, this overdetermined system has

- two unknowns β_1 and β_2
- 3 equations

The image below shows

- $\vec{a}_1$ as the green arrow corresponding to the constant term in $y = \beta_1 + \beta_2 x^2$

  ($\vec{a}_1 = [1, 1, 1]^T$ as in previous example)

- $\vec{a}_2$ as the green arrow corresponding to the quadratic term in $y = \beta_1 + \beta_2 x^2$

  ($\vec{a}_2$ is different from $\vec{a}_2$ in the prior example)

- the span of linearly independent $\vec{a}_1$ & $\vec{a}_2$ as a plane

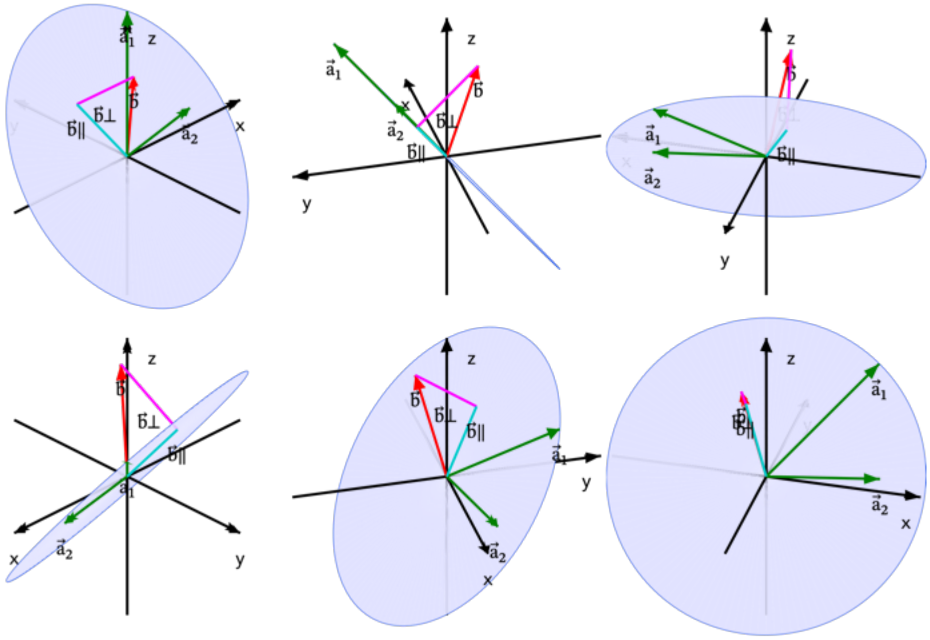
  (plane is different from the plane in the prior example)

- $\vec{b}$ as the red arrow (same as in the prior example)

$\vec{b}$ is again outside the span of $\vec{a}_1$ & $\vec{a}_2$

↓

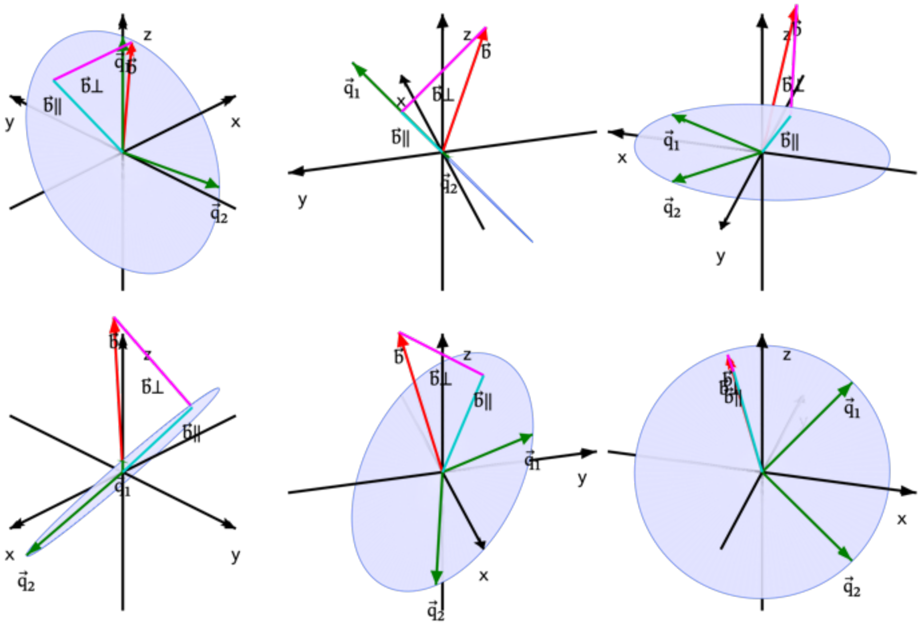
an exact solution does not exist



As before, we orthogonalize the column space of A:

$$Q = [\, \vec{q}_1 \mid \vec{q}_2 \,] = \begin{bmatrix} \approx 0.577 & \approx 0.802 \\ \approx 0.577 & \approx -0.267 \\ \approx 0.577 & \approx -0.535 \end{bmatrix}$$

- $\vec{q}_1$ is the same as normalized $\vec{a}_1$
- the plane spanned by $\vec{q}_1$ & $\vec{q}_2$ is the same as the plane spanned by $\vec{a}_1$ & $\vec{a}_2$
- $\vec{q}_2$ is orthogonal to $\vec{q}_1$ in the same plane
- $\vec{b}$ is unchanged
- $\vec{b}\| = Q\,Q^{T}\,\vec{b}$ or the projection of $\vec{b}$ onto span($\vec{q}_1$, $\vec{q}_2$)



We compute an upper-triangular matrix $R = Q^T A = \begin{bmatrix} \approx 1.732 & \approx 0.849 \\ 0 & \approx 0.898 \end{bmatrix}$

After solving $R\,\vec{\beta} = Q^T\,\vec{b}$, we obtain

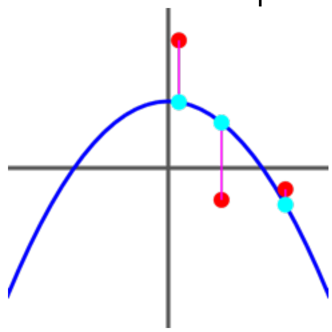
- $\beta_1 \approx 0.627$ (constant term in $y = \beta_1 + \beta_2 x^2$)
- $\beta_2 \approx -0.804$ (quadratic term in $y = \beta_1 + \beta_2 x^2$)

- $\vec{b}$ represents the three actual 'y' values of the dataset
- Each vector in the plane $\text{span}(\vec{a}_1, \vec{a}_2) = \text{span}(\vec{q}_1, \vec{q}_2)$ represents the three predicted values of some parabola $y = \beta_1 + \beta_2 x^2$ evaluated at the three sample points
- The cyan vector $\vec{b}\|$ represents the predicted values for the best fit parabola
- The 3D magenta segment represents the residual vector

$$\vec{b}\perp = \vec{b} - \vec{b}\| = \vec{b} - Q\,Q^T\,\vec{b} = \vec{b} - A\,\vec{\beta}$$

- The sum of squared residuals $|\,\vec{b} - A\,\vec{\beta}\,|^2 \approx 0.886$ is the quantity the algorithm minimizes

Image below shows the best fit parabola $y = \beta_1 + \beta_2 x^2$



- red dots are the 3 input values
- cyan dots are the 3 values predicted by the best parabola
- magenta lines are individual residuals or individual entries of $\vec{b}\perp$

Two models for same dataset

① Suppose we want to fit a line

$y = \beta_1 + \beta_2 x$

through the following point set:

- $(x_1 = 1, y_1 = 1)$
- $(x_2 = -1, y_2 = 1)$
- $(x_3 = 0, y_3 = -2)$

This gives us the following system:

$[\, \vec{a}_1 \mid \vec{a}_2 \,]\, \vec{\beta} = \vec{b}$

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Notice that $\vec{a}_1 \cdot \vec{b} = 0$ and $\vec{a}_2 \cdot \vec{b} = 0$

↓

red vector $\vec{b}$ shown on previous pages is orthogonal to the plane

↓

$\vec{b}|| = \vec{0}$

$$Q = [\ \vec{q}_1 \mid \vec{q}_2\ ] = \begin{bmatrix} \approx 0.577 & \approx 0.707 \\ \approx 0.577 & \approx -0.707 \\ \approx 0.577 & 0 \end{bmatrix}$$

$\vec{q}_1$ & $\vec{q}_2$ are within the span of $\vec{a}_1$ & $\vec{a}_2$

↓

$Q^T \vec{b} = \vec{0}$

$$R = Q^T A = \begin{bmatrix} \approx 1.732 & 0 \\ 0 & \approx 1.414 \end{bmatrix}$$

solving $R\,\vec{\beta} = Q^T \vec{b}$

$$\begin{bmatrix} \approx 1.732 & 0 \\ 0 & \approx 1.414 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- $\beta_1 = 0$ (in $y = \beta_1 + \beta_2 x$)
- $\beta_2 = 0$ (in $y = \beta_1 + \beta_2 x$)

↓

dataset cannot be represented by a linear model

(note that in this example, $\vec{a}_1 \cdot \vec{a}_2 = 0$, which is not a necessary condition)

② Next, we will fit a parabola $y = \beta_1 + \beta_2 x^2$

through the same point set

As we saw on the previous page,

- $\vec{b}$ will remain the same
- $\vec{a}_1$ will remain the same
- $\vec{a}_2$ will change direction & the plane will change orientation

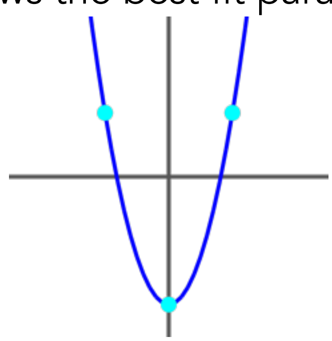
$$Q = [\ \vec{q}_1 \mid \vec{q}_2\ ] = \begin{bmatrix} \approx 0.577 & \approx 0.408 \\ \approx 0.577 & \approx 0.408 \\ \approx 0.577 & \approx -0.816 \end{bmatrix}$$

solving $R\,\vec{\beta} = Q^T\,\vec{b}$

$$\begin{bmatrix} \approx 1.732 & \approx 1.155 \\ 0 & \approx 0.816 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \approx 2.449 \end{bmatrix}$$

- $\beta_1 = -2$ (in $y = \beta_1 + \beta_2 x^2$)
- $\beta_2 = 3$ (in $y = \beta_1 + \beta_2 x^2$)

Image below shows the best fit parabola $y = \beta_1 + \beta_2 x^2$



Notice that the parabola fits the dataset precisely

that means that $\vec{b}\perp = \vec{0}$

↓

$\vec{b} \in \text{col}(A)$

You can also see why a straight line could not fit the same point set

In this example, replacing x with x^2 changed the plane from one orthogonal to $\vec{b}$ to one containing $\vec{b}$

Higher-dimensional set

Next, we will fit a sum of sine and cosine waves:

$y = \beta_1 + \beta_2 \sin(x) + \beta_3 \cos(x) + \beta_4 \sin(2x) + \beta_5 \cos(2x)$

into the following dataset

- ($x_1 = -1.88$, $y_1 = 0.62$)
- ($x_2 = -1.41$, $y_2 = -0.17$)
- ($x_3 = -1.07$, $y_3 = 1.34$)
- ($x_4 = -0.73$, $y_4 = 0.21$)
- ($x_5 = -0.31$, $y_5 = -1.08$)
- ($x_6 = 0.12$, $y_6 = 0.57$)
- ($x_7 = 0.44$, $y_7 = -0.64$)
- ($x_8 = 0.91$, $y_8 = 1.11$)
- ($x_9 = 1.36$, $y_9 = 0.18$)
- ($x_{10} = 1.94$, $y_{10} = -0.86$)

This time, we will be projecting $\vec{b}$ in $\mathbb{R}^{10}$ onto the 5-dimensional col space(A)

$$A = \begin{bmatrix} 1 & \sin(x_1) & \cos(x_1) & \sin(2x_1) & \cos(2x_1) \\ 1 & \sin(x_2) & \cos(x_2) & \sin(2x_2) & \cos(2x_2) \\ 1 & \sin(x_3) & \cos(x_3) & \sin(2x_3) & \cos(2x_3) \\ 1 & \sin(x_4) & \cos(x_4) & \sin(2x_4) & \cos(2x_4) \\ 1 & \sin(x_5) & \cos(x_5) & \sin(2x_5) & \cos(2x_5) \\ 1 & \sin(x_6) & \cos(x_6) & \sin(2x_6) & \cos(2x_6) \\ 1 & \sin(x_7) & \cos(x_7) & \sin(2x_7) & \cos(2x_7) \\ 1 & \sin(x_8) & \cos(x_8) & \sin(2x_8) & \cos(2x_8) \\ 1 & \sin(x_9) & \cos(x_9) & \sin(2x_9) & \cos(2x_9) \\ 1 & \sin(x_{10}) & \cos(x_{10}) & \sin(2x_{10}) & \cos(2x_{10}) \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix}$$

$$\begin{bmatrix} 1 & \approx-0.953 & \approx-0.304 & \approx0.58 & \approx-0.815 \\ 1 & \approx-0.987 & \approx0.16 & \approx-0.316 & \approx-0.949 \\ 1 & \approx-0.877 & \approx0.48 & \approx-0.842 & \approx-0.539 \\ 1 & \approx-0.667 & \approx0.745 & \approx-0.994 & \approx0.111 \\ 1 & \approx-0.305 & \approx0.952 & \approx-0.581 & \approx0.814 \\ 1 & \approx0.12 & \approx0.993 & \approx0.238 & \approx0.971 \\ 1 & \approx0.426 & \approx0.905 & \approx0.771 & \approx0.637 \\ 1 & \approx0.79 & \approx0.614 & \approx0.969 & \approx-0.247 \\ 1 & \approx0.978 & \approx0.209 & \approx0.409 & \approx-0.912 \\ 1 & \approx0.933 & \approx-0.361 & \approx-0.673 & \approx-0.74 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0.62 \\ -0.17 \\ 1.34 \\ 0.21 \\ -1.08 \\ 0.57 \\ -0.64 \\ 1.11 \\ 0.18 \\ -0.86 \end{bmatrix}$$

We solve it in the same fashion by

• orthogonalizing col space(A)

• solving the upper-triangular system $R \vec{\beta} = Q^T \vec{b}$:

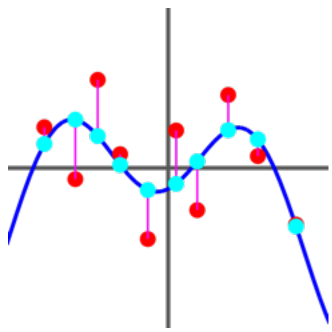
• $\beta_1 \approx -0.596$ (constant term)

• $\beta_2 \approx -0.316$ (sin(x) term)

• $\beta_3 \approx 1.272$ (cos(x) term)

• $\beta_4 \approx 0.408$ (sin(2x) term)

• $\beta_5 \approx -0.996$ (cos(2x) term)



Summary: models with one input variable

① Which models correspond to linear least-squares problems

• Valid models:

parameters appear only as linear coefficients:

$$y = \beta_1 f_1(x) + \beta_2 f_2(x) + \ldots + \beta_n f_n(x)$$

Invalid models:

- parameters raised to power: $y = \beta_1 + \beta_2^2 x$
- parameters used as exponents: $y = \beta_1 + e^{\beta_2}$
- parameters used as function arguments: $y = \beta_1 + \sin(\beta_2 x)$
- products of two or more parameters: $y = \beta_1 + \beta_2 \beta_3$

Valid model examples

Model	Basis functions $f_i(x)$	Utility
Constant	1	Data fluctuates around a constant level
Line	1, x	Output varies linearly with one variable
Parabola	1, x^2	Curved trend with one turning point
Polynomial	1, x, x^2, x^3, …	Smooth curve of unknown shape
Fourier function sum of sine and cosine waves	1, sin(x), cos(x), sin(2x), cos(2x)	Periodic behavior

② General form of OLS system based on general model

$$y = \beta_1 f_1(x) + \beta_2 f_2(x) + \ldots + \beta_n f_n(x)$$

shown for m = 5 points and n = 3 parameters

$A \vec{\beta} = \vec{b}$ or

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & f_3(x_1) \\ f_1(x_2) & f_2(x_2) & f_3(x_2) \\ f_1(x_3) & f_2(x_3) & f_3(x_3) \\ f_1(x_4) & f_2(x_4) & f_3(x_4) \\ f_1(x_5) & f_2(x_5) & f_3(x_5) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

- columns(A) = basis functions
- rows(A) = data points

Assuming $m > n$ & A has full-column rank:

- $\vec{b} \in \text{col}(A)$ → exact solution
- $\vec{b} \notin \text{col}(A)$ → least-squares projection onto col(A)
- $\vec{b} \perp \text{col}(A)$ → projection is zero → model cannot explain the data

④ Note that one of the f functions is usually 1
which corresponds to the constant term in
$y = \beta_1 f_1(x) + \beta_2 f_2(x) + \ldots + \beta_n f_n(x)$
This creates a column $[1, 1, \ldots, 1]^T$ in matrix A

$\vec{b}_\perp$ is orthogonal to every column of A

↓

$$[1, 1, \ldots, 1]^T \vec{b}_\perp = \sum (b_\perp)_i = 0$$

⇔

the sum of all entries of $b_\perp$ (or all residuals) is 0

⇔

model is centered on the data

Model with 2 input variables

Suppose we want to fit the best plane through the points:

- (x_1 = 0.8, y_1 = 1.2, z_1 = 5.602)
- (x_2 = 2, y_2 = 0.7, z_2 = 1.538)
- (x_3 = 1.5, y_3 = 2.4, z_3 = 0.68)
- (x_4 = 3.1, y_4 = 1.8, z_4 = 5.523)
- (x_5 = 2.6, y_5 = 3.3, z_5 = 5.913)

We define the plane as

$$z = \beta_1 + \beta_2 x + \beta_3 y$$

Solving $A\vec{\beta} = \vec{b}$ or

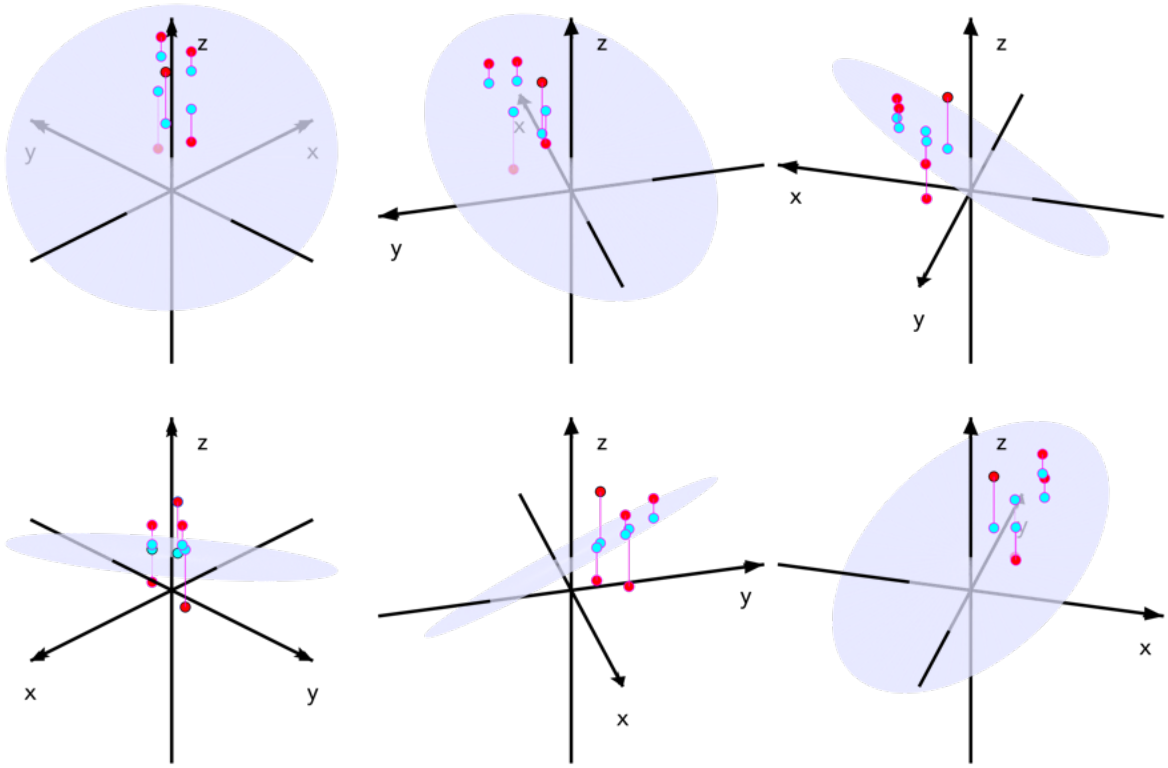
$$\begin{bmatrix} 1 & 0.8 & 1.2 \\ 1 & 2 & 0.7 \\ 1 & 1.5 & 2.4 \\ 1 & 3.1 & 1.8 \\ 1 & 2.6 & 3.3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 5.602 \\ 1.538 \\ 0.68 \\ 5.523 \\ 5.913 \end{bmatrix}$$

- $\beta_1 \approx 1.792$ (constant term)
- $\beta_2 \approx 0.601$ ('x' term)
- $\beta_3 \approx 0.456$ ('y' term)

- $\vec{a}_1, \vec{a}_2, \vec{a}_3, \vec{b}, \vec{b}\perp, \vec{b}\| \in \mathbb{R}^5$ and cannot be drawn
- dimension $\big(\text{col}(A)\big) = 3$

The image below shows the plane fitted to the dataset

- red dots are the 5 input values
- cyan dots are the 5 values predicted by the best plane
- magenta lines are individual residuals or individual entries of $\vec{b}\perp$



Non-linear model with 2 input variables

Example of fitting the best elliptic paraboloid through the points:

- $(x_1 = -1.8, y_1 = -1.2, z_1 = 4.35)$
- $(x_2 = -1.4, y_2 = 0.9, z_2 = 0.35)$
- $(x_3 = -0.7, y_3 = -1.5, z_3 = 1.1)$
- $(x_4 = -0.2, y_4 = 1.3, z_4 = -2.1)$
- $(x_5 = 0.6, y_5 = -0.8, z_5 = 0.65)$
- $(x_6 = 1.1, y_6 = 0.4, z_6 = -0.85)$
- $(x_7 = 1.6, y_7 = -1.1, z_7 = 3.85)$

- $(x_8 = 2, y_8 = 1.2, z_8 = 2.45)$

We define the paraboloid as

$z = \beta_1 + \beta_2 x^2 + \beta_3 y^2$

Solving $A \vec{\beta} = \vec{b}$

$$\begin{bmatrix} 1 & 3.24 & 1.44 \\ 1 & 1.96 & 0.81 \\ 1 & 0.49 & 2.25 \\ 1 & 0.04 & 1.69 \\ 1 & 0.36 & 0.64 \\ 1 & 1.21 & 0.16 \\ 1 & 2.56 & 1.21 \\ 1 & 4 & 1.44 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 4.35 \\ 0.35 \\ 1.1 \\ -2.1 \\ 0.65 \\ -0.85 \\ 3.85 \\ 2.45 \end{bmatrix}$$

- $\beta_1 \approx -1.643$ (constant term)
- $\beta_2 \approx 1.163$ (x^2 term)
- $\beta_3 \approx 0.708$ (y^2 term)

The target paraboloid has its vertex at

- $x_0 = 0$ (since the model does not have a linear x-term)
- $y_0 = 0$ (since the model does not have a linear y-term)
- $z_0 = \beta_1$

the solution produced β_2 & $\beta_3 > 0$

↓

the surface is an upward-opening elliptic paraboloid, which we can parametrize as

$z = \beta_1 + \beta_2 x^2 + \beta_3 y^2$

by introducing variables t and θ:

- θ: controls angle in the x-y plane
- t: controls distance from the vertex

- $x = t \dfrac{\cos(\theta)}{\sqrt{\beta_2}}$

- $y = t \dfrac{\sin(\theta)}{\sqrt{\beta_3}}$

- $z = \beta_1 + t^2$

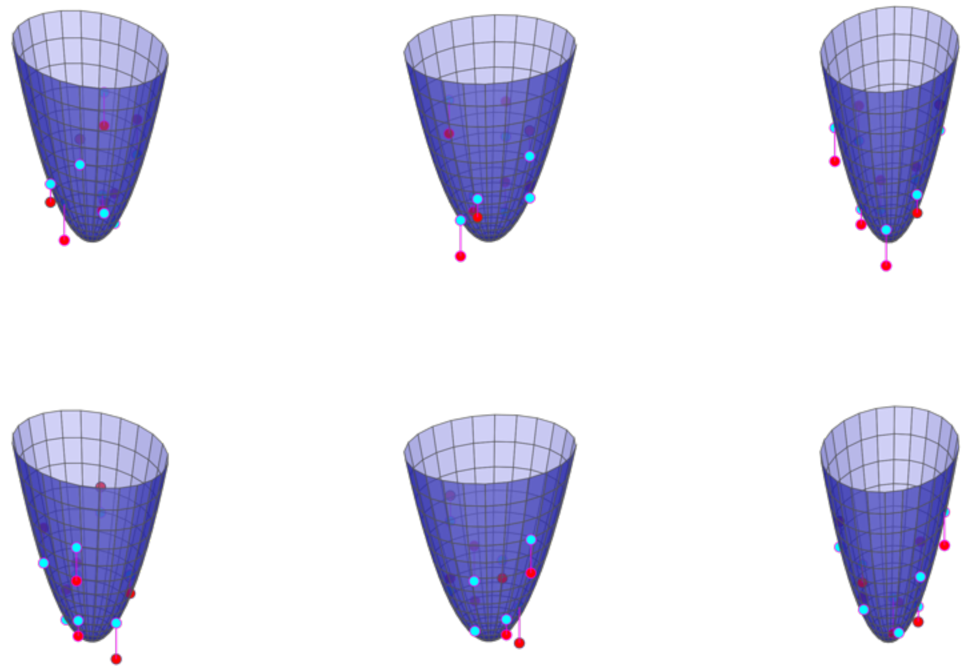
- $z = \beta_1 + \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} \beta_2 & 0 \\ 0 & \beta_3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

- $\vec{a}_1, \vec{a}_2, \vec{a}_3, \vec{b}, \vec{b}\perp, \vec{b}\| \in \mathbb{R}^8$
and cannot be drawn

- dimension $\big(\text{col}(A)\big) = 3$

The image below shows the paraboloid fitted to the dataset

- red dots are the 8 input values
- cyan dots are the 8 values predicted by the best paraboloid
- magenta lines are individual residuals or individual entries of $\vec{b}\perp$



The table below summarizes the shapes we showed so far

Shape	Model	Number of parameters	Number of inputs	Basis functions (columns of A or model subspace)
Line	$y = \beta_1 + \beta_2 x$	2	1	1, x
Parabola	$y = \beta_1 + \beta_2 x^2$	2	1	1, x^2
Periodic curve	$y = \beta_1 + \beta_2 \sin(x) + \beta_3 \cos(x) + \beta_4 \sin(2x) + \beta_5 \cos(2x)$	5	1	1, sin(x), cos(x), sin(2x), cos(2x)
Plane	$z = \beta_1 + \beta_2 x + \beta_3 y$	3	2	1, x, y
Paraboloid	$z = \beta_1 + \beta_2 x^2 + \beta_3 y^2$	3	2	1, x^2, y^2

note that a parabola with only 2 parameters was chosen to display

- data space as $\mathbb{R}^3$
- 2-dimensional parameter space as a subset of $\mathbb{R}^3$

Summary: models with > 1 input variable

General form of an OLS model with one input variable was shown on the previous page:

$$y = \beta_1 f_1(x) + \beta_2 f_2(x) + \ldots + \beta_n f_n(x)$$

Same principle extended to 2 variables:

$$z = u_1 f_1(x) + u_2 f_2(x) + \ldots + u_k f_k(x) +$$
$$v_1 g_1(y) + v_2 g_2(y) + \ldots + v_m g_m(y) +$$
$$w_1 h_1(x, y) + w_2 h_2(x, y) + \ldots + w_n h_n(x, y)$$

- line 1: functions of input variable x
- line 2: functions of input variable y
- line 3: functions of both input variables (interaction terms)

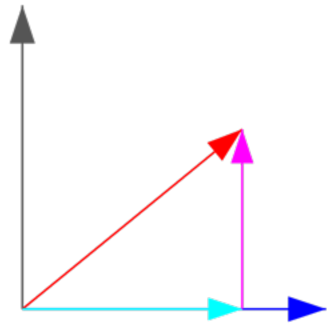
Below is one possible model with

- 2 unknowns u
- 2 unknowns v
- 2 unknowns w
- 8 data points

$$\text{Vector of unknowns } \vec{\beta} = \begin{bmatrix} u_1 \\ u_2 \\ v_1 \\ v_2 \\ w_1 \\ w_2 \end{bmatrix} \qquad \text{right side } \vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \\ b_8 \end{bmatrix}$$

(the coefficients (u, v, w) are collected into a single parameter vector $\vec{\beta}$)

$$A = \begin{bmatrix} f_1(x_1) & f_2(x_1) & g_1(y_1) & g_2(y_1) & h_1(x_1, y_1) & h_2(x_1, y_1) \\ f_1(x_2) & f_2(x_2) & g_1(y_2) & g_2(y_2) & h_1(x_2, y_2) & h_2(x_2, y_2) \\ f_1(x_3) & f_2(x_3) & g_1(y_3) & g_2(y_3) & h_1(x_3, y_3) & h_2(x_3, y_3) \\ f_1(x_4) & f_2(x_4) & g_1(y_4) & g_2(y_4) & h_1(x_4, y_4) & h_2(x_4, y_4) \\ f_1(x_5) & f_2(x_5) & g_1(y_5) & g_2(y_5) & h_1(x_5, y_5) & h_2(x_5, y_5) \\ f_1(x_6) & f_2(x_6) & g_1(y_6) & g_2(y_6) & h_1(x_6, y_6) & h_2(x_6, y_6) \\ f_1(x_7) & f_2(x_7) & g_1(y_7) & g_2(y_7) & h_1(x_7, y_7) & h_2(x_7, y_7) \\ f_1(x_8) & f_2(x_8) & g_1(y_8) & g_2(y_8) & h_1(x_8, y_8) & h_2(x_8, y_8) \end{bmatrix}$$



Suppose A is an m×n matrix with rank = n

• dataset lives in $\mathbb{R}^m$

• while $\mathbb{R}^m$ cannot be visualized directly, it can be visualized as two spanning and orthogonal subspaces or two sets of vectors:

col space(A) ⊕ ℓ-null(A)

• col space(A) is shown in blue and has the dimension of rank(A) = n

• ℓ-null(A) is shown in gray and has a dimension of m-n

• $\vec{b}$ is shown in red and decomposes as $\vec{b}_{\parallel} + \vec{b}_{\perp}$

• $\vec{b}_{\parallel} \in$ col space(A) is shown in cyan

• $\vec{b}_{\perp} \in$ ℓ-null(A) is shown in magenta

• A $\vec{\beta} = \vec{b}$ is solved by projecting $\vec{b}$ onto col space(A)

• A regression model defines col(A) or a subspace of the data space

• Choosing a model means choosing the subspace or the columns of A

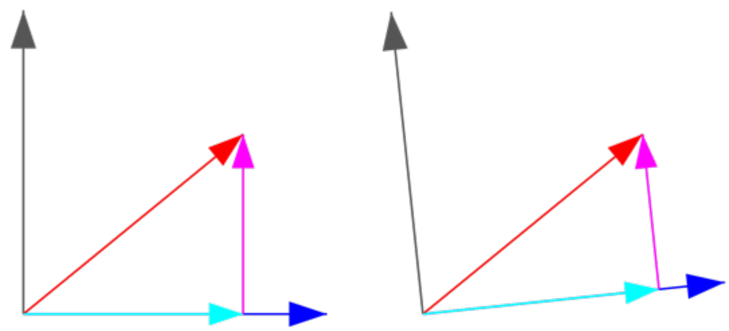
• Fitting the model means finding the best linear combination of those columns

• Least squares algorithm replaces $\vec{b}$ with $\vec{b}_{\parallel}$ or its projection onto the model subspace col(A)

Two ways to change a regression model

① Change the directions spanned by col space(A) or
rotate the model subspace inside the data space as shown above:

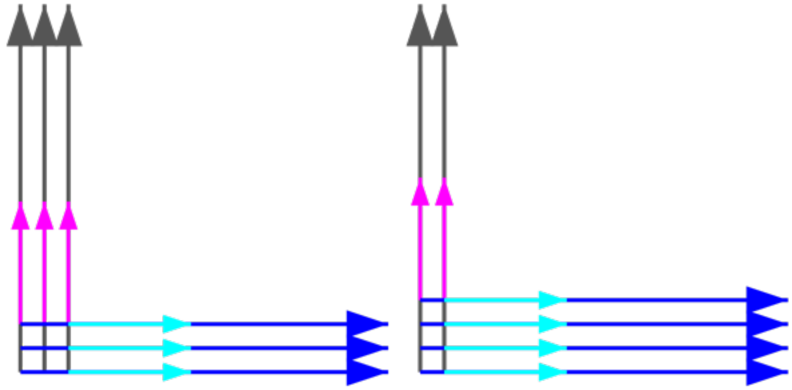
- keeps the dimension the same
- changes the projection of $\vec{b}$ onto col(A)
- changes the decomposition of $\vec{b}$ into $\vec{b}\perp + \vec{b}\|$



- ‣ basis of col(A): combined into one blue arrow
- ‣ basis of ℓ-null(A): combined into one gray arrow
- ‣ $\vec{b}$: red
- ‣ $\vec{b}\perp$: magenta
- ‣ $\vec{b}\|$: cyan

② Add new parameters or linearly independent columns of A:

- increases dimension of col space(A)
- decreases dimension of ℓ-null space(A)
- gives the projection more directions available
- can only decrease the residual norm



‣ basis of col(A): individual blue arrows

‣ basis of ℓ-null space(A): individual gray arrows

‣ projections of $\vec{b}$ onto ℓ-null space(A) vectors: magenta

‣ projections of $\vec{b}$ onto col space(A) vectors: cyan

This leads to the concept of underfitting and overfitting:

• Underfitting:
the model subspace is too small
↓
too much of b remains in ℓ-null(A)

• Overfitting:
the model subspace is too large
↓
the model reflects both true and accidental variation in the data

Geometric trade-off:
choose a model subspace large enough to capture the structure in b,
but not so large that it begins to fit accidental variation