EECS 245 Fall 2025
Math for ML

Lecture 25: SVD, PCA

→ Read: Ch 5.4 (under development; final chapter!)

Agenda

1. Recap: SVD
2. Dimensionality reduction
   - → Finding the best line to project onto
   - → Maximizing variance

Announcements

1. HW 11 due Sunday, no slip days!
   Time to work on it in lab
2. Final Exam on Wednesday 12/10
   Logistical details coming soon, but
   - 3 double-sided notes sheets
   - No mock exam session: will release practice probs

$$\underbrace{\begin{bmatrix} 3 & 2 & 5 \\ 2 & 3 & 5 \\ 2 & -2 & 0 \\ 5 & 5 & 10 \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{3\sqrt{2}} & -\frac{1}{\sqrt{3}} & -\frac{2}{3} \\ \frac{1}{\sqrt{6}} & -\frac{1}{3\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & 0 & \frac{1}{3} \\ \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} & 0 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 15 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix}}_{V^T}$$

$$X = U \Sigma V^T$$

singular values $\sigma_i = \sqrt{\lambda_i}$

λ_i eigval X^TX

$X: n \times d \qquad U: n \times n, \quad \Sigma: n \times d, \quad V^T: d \times d$

U → eigenvectors of XX^T

$$\Sigma = \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ 0 & & \ddots \end{bmatrix}$$

V has → eigenvectors of X^TX

symmetric!

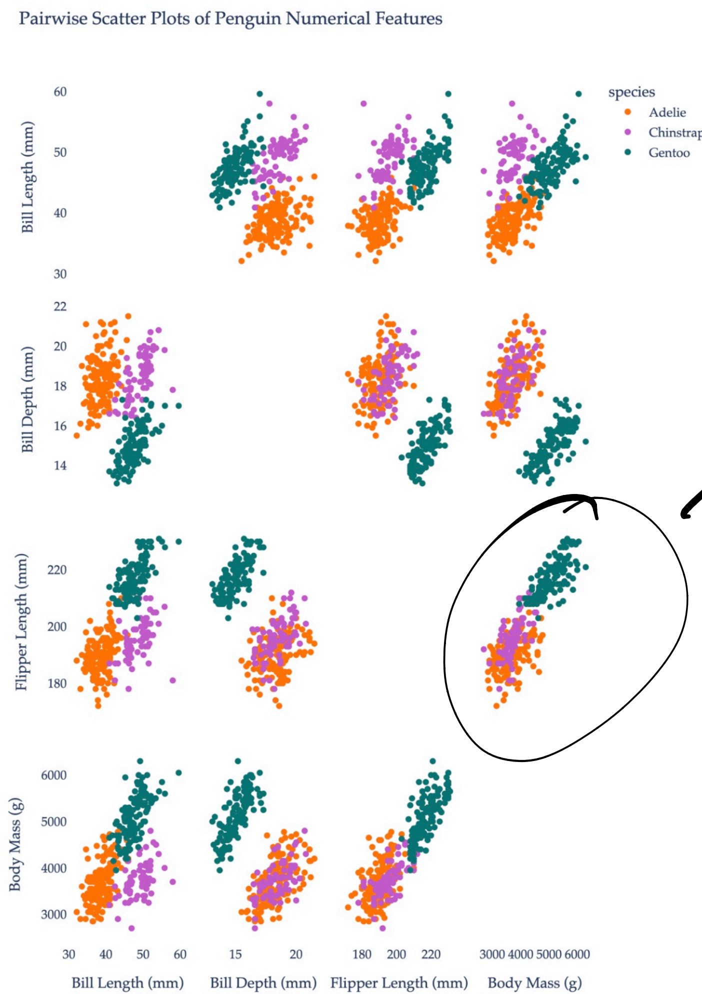
U orthogonal: $U^TU = UU^T = I$

V orthogonal: $V^TV = VV^T = I$

Low-rank approximation

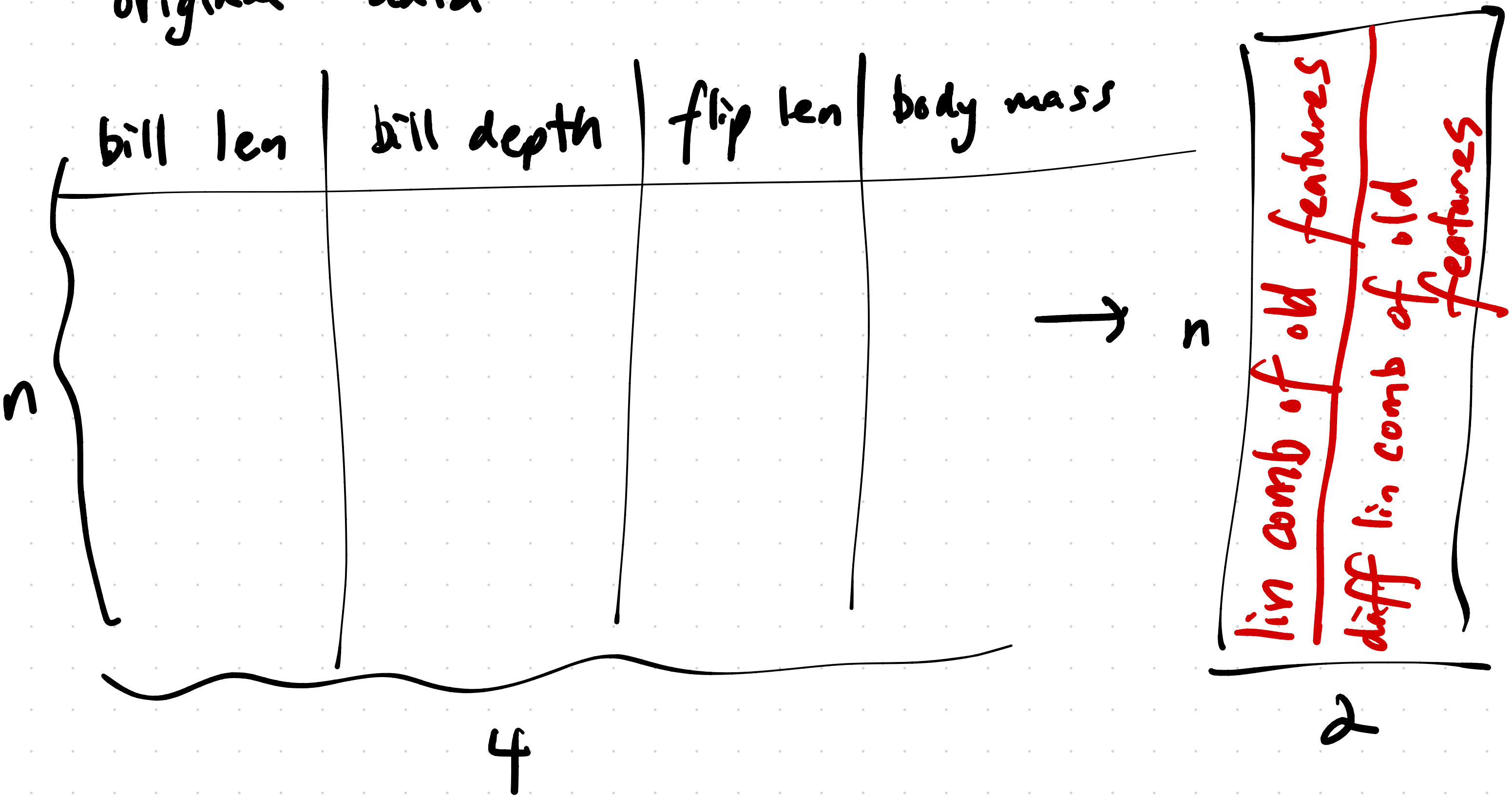
$$\underbrace{\begin{bmatrix} 3 & 2 & 5 \\ 2 & 3 & 5 \\ 2 & -2 & 0 \\ 5 & 5 & 10 \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{3\sqrt{2}} & -\frac{1}{\sqrt{3}} & -\frac{2}{3} \\ \frac{1}{\sqrt{6}} & -\frac{1}{3\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & 0 & \frac{1}{3} \\ \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} & 0 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 15 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix}}_{V^T}$$

$$X = \underbrace{15 \begin{bmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ 0 \\ 2/\sqrt{6} \end{bmatrix} \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{6} & 2/\sqrt{6} \end{bmatrix}}_{\text{rank } 1} + \underbrace{3 \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix}}_{\text{rank } 1}$$



very strongly correlated!

original data

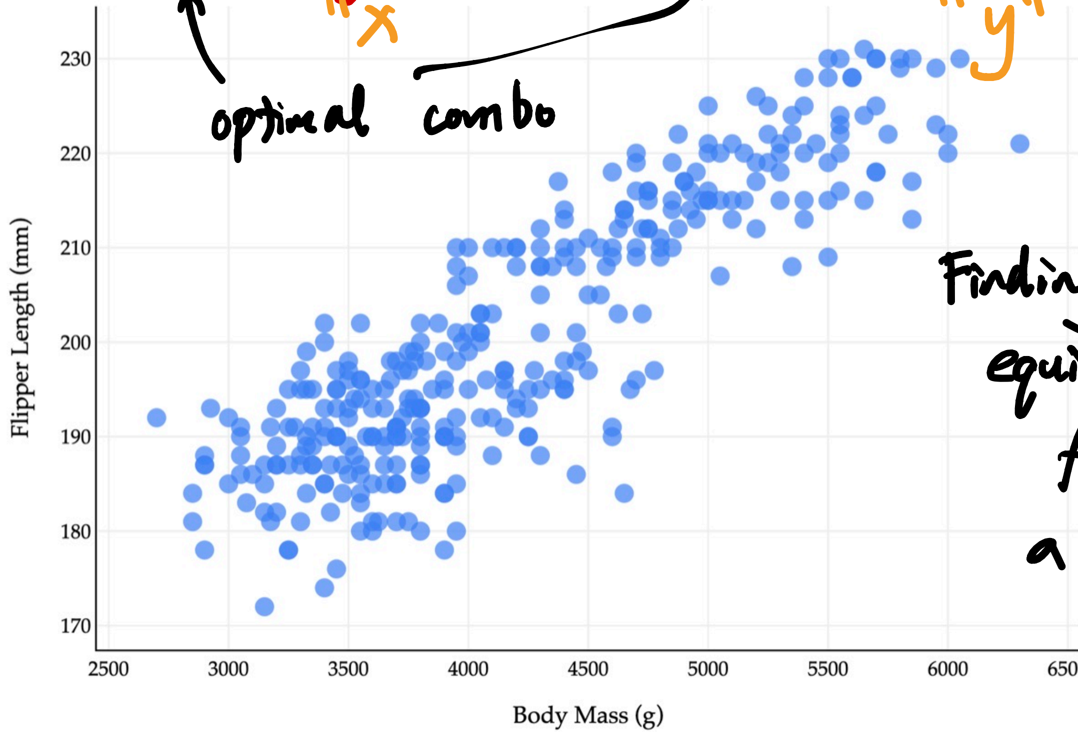


Goal: take dataset in $\mathbb{R}^2$, "project it" to $\mathbb{R}^1$

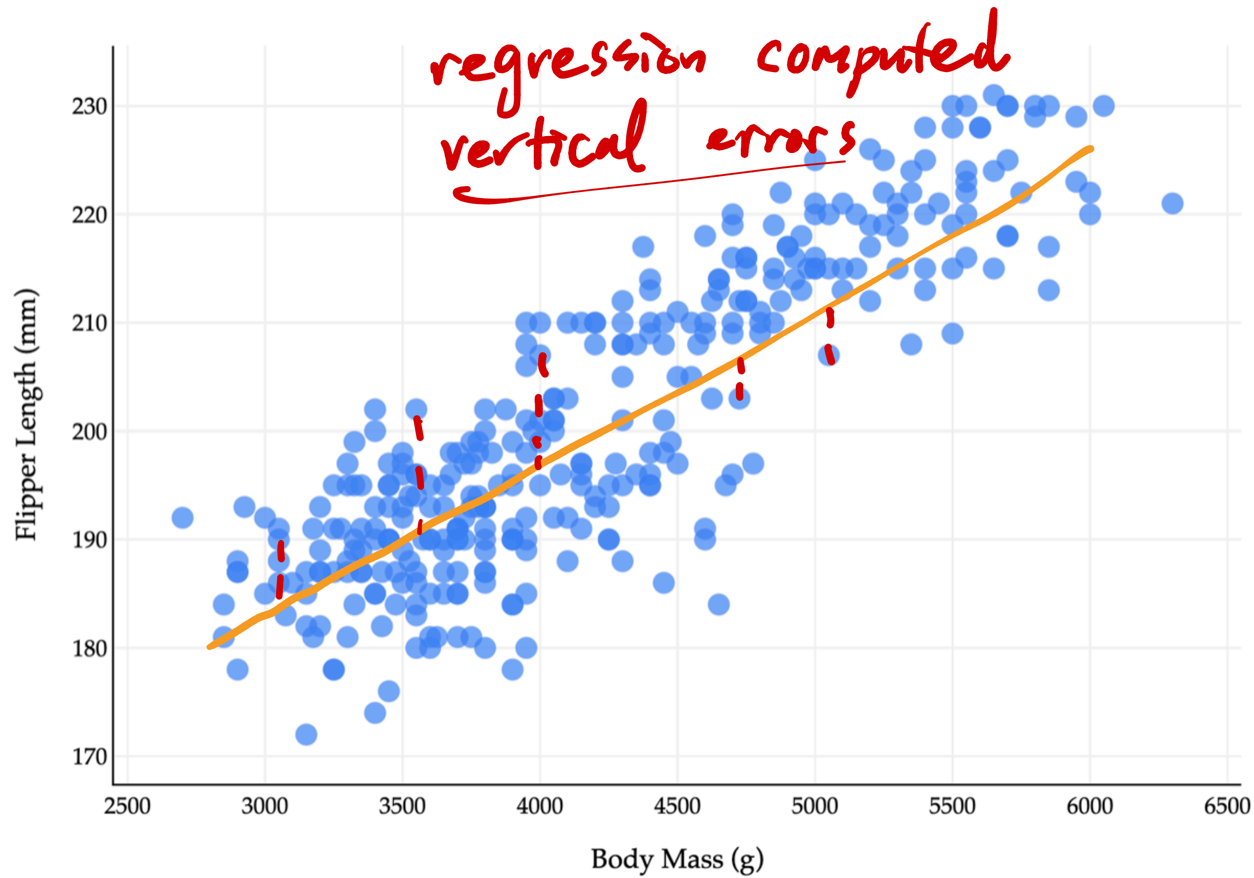
new feature$_i$ = a (body mass)$_i$ + b (flipper length)$_i$

"x" "y"

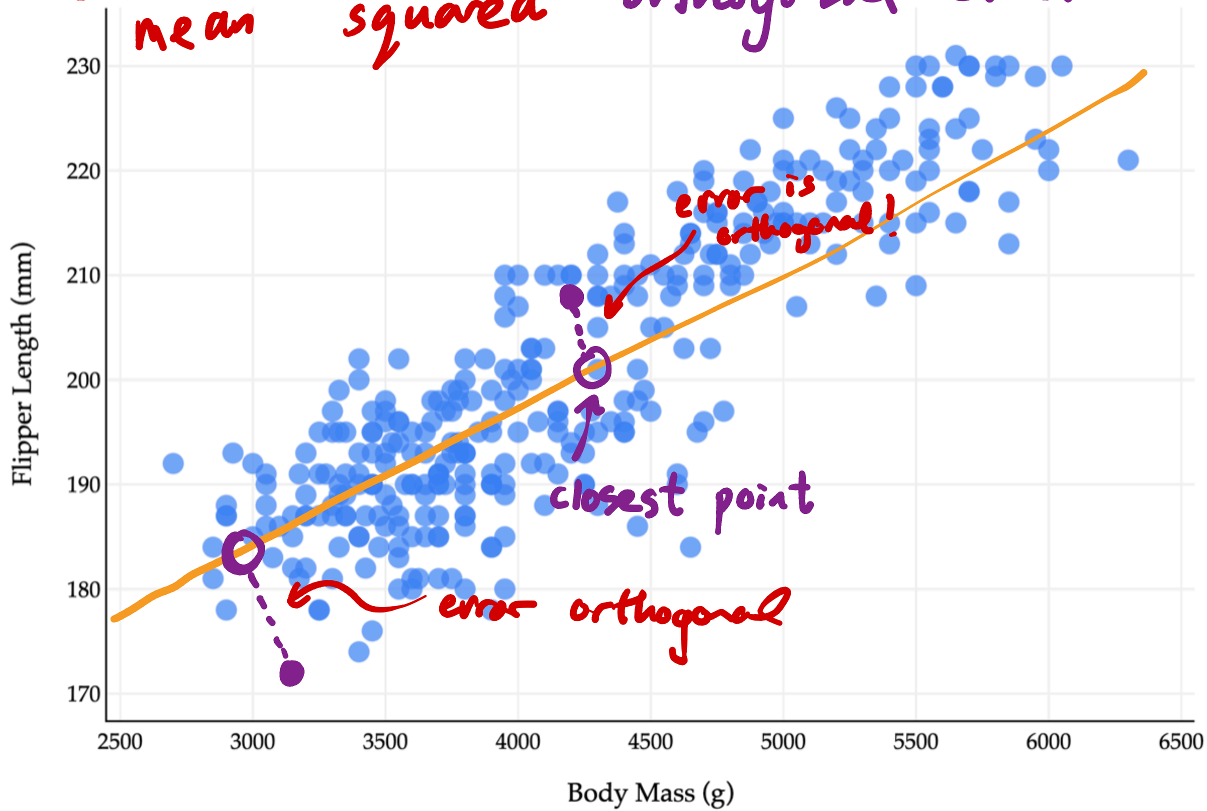
optimal combo



Finding a and b equivalent to finding a best line

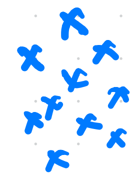

regression computed
vertical errors
230
220
210
200
190
180
170
Flipper Length (mm)
2500
3000
3500
4000
4500
5000
5500
6000
6500
Body Mass (g)

searching for the line with the smallest possible mean squared orthogonal error



span of one vector $\vec{v} \in \mathbb{R}^d$ is a line through origin

data usually far from origin

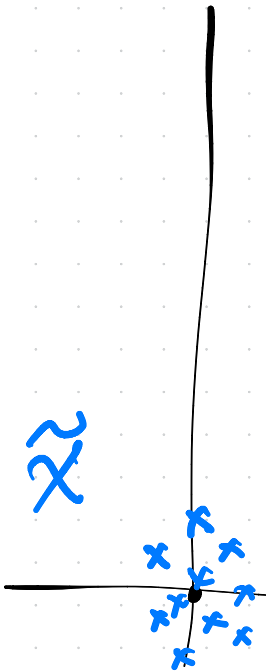


X

center

from each col, subtract the mean of that col

$\tilde{X}$



now, data is centered at $(0,0)$, so a line through the origin is good!

$(0,0)$

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(d)} \\ x_2^{(1)} & x_2^{(2)} & \ddots & \\ & & & x_n^{(d)} \end{bmatrix}$$

μ_j is the mean of col j

"mean centering"

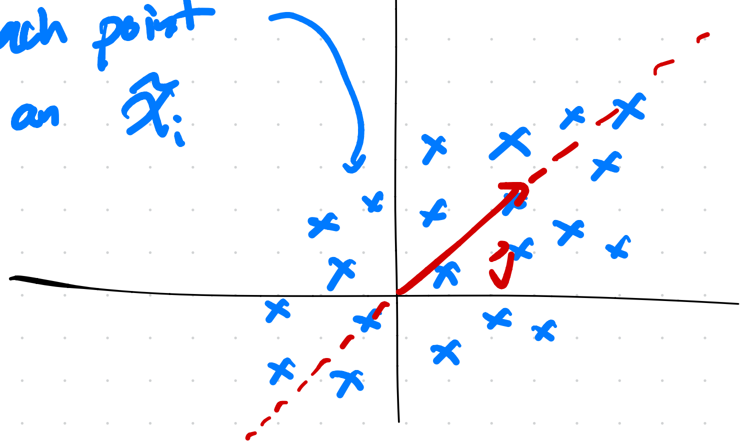
$$\downarrow$$

cols of $\tilde{X}$ each have mean 0

$$\tilde{X} = \begin{bmatrix} x_1^{(1)} - \mu_1 & x_1^{(2)} - \mu_2 & \\ x_2^{(1)} - \mu_1 & \vdots & \cdots \\ \vdots & & \\ x_n^{(1)} - \mu_1 & x_n^{(2)} - \mu_2 & \end{bmatrix}$$

$$\tilde{X} = \begin{bmatrix} - \; \tilde{\vec{x}}^{1T} \; - \\ - \; \tilde{\vec{x}}^{2T} \; - \\ \vdots \\ - \; \tilde{\vec{x}}^{nT} \; - \end{bmatrix}$$

each point is an $\tilde{\vec{x}}_i$



assume $\vec{v}$ unit vector

want $\vec{v}$ to have the smallest poss mean sq. orthogonal error

if $\tilde{\vec{x}}_i$ is a point, the projection onto $\vec{v}$ is

$$\vec{p}_i = \left(\frac{\tilde{\vec{x}}_i \cdot \vec{v}}{\vec{v} \cdot \vec{v}} \right) \vec{v}$$

$$= (\tilde{\vec{x}}_i \cdot \vec{v}) \vec{v}$$

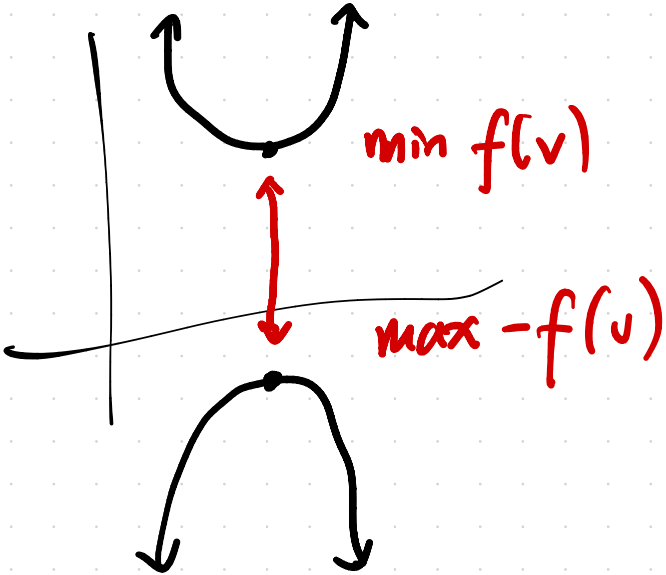
new feature values

mean sq orthogonal error: want to minimize

$$J(\vec{v}) = \frac{1}{n}\sum_{i=1}^{n} \| \underbrace{\tilde{x}_i - (\tilde{x}_i \cdot \vec{v})\vec{v}}_{\text{error vec}} \|^2$$

see 5.4

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} \|\tilde{x}_i\|^2}_{\text{constant w.r.t. } \vec{v}\text{! irrelevant}} - \frac{1}{n}\sum_{i=1}^{n} (\tilde{x}_i \cdot \vec{v})^2$$


min f(v)
max -f(v)

minimizing $J(\vec{v}) = \frac{1}{n} \| \tilde{x}_i - (\tilde{x}_i \cdot \vec{v})\vec{v} \|^2$

equivalent to

maximizing $\frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i \cdot \vec{v})^2$

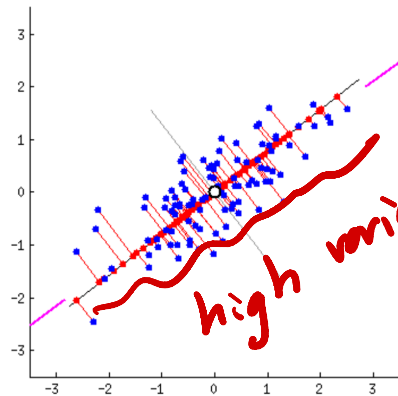
maximizing variance equivalent to min mean squared orthogonal error

= Variance of $\tilde{x}_i \cdot \vec{v}$'s!

= Variance of my new feature!

ilar values of w_1 and w_2.

ook here very carefully -- here is what these projections look like for different lines (red re projections of the blue dots):



high variance

the shorter the orthogonal errors are (on average), the more spread out the points are!

id before, PCA will find the "best" line according to two different criteria of what is the . First, the variation of values along this line should be maximal. Pay attention to how the

see animation in notes

$\vec{v}^*$ minimizes $\frac{1}{n}\sum_{i=1}^{n} \| \tilde{x}_i - (\tilde{x}_i \cdot \vec{v})\vec{v} \|^2$

$\vec{v}^*$ maximizes $\frac{1}{n}\sum_{i=1}^{n} (\tilde{x}_i \cdot \vec{v})^2$

$$= \frac{1}{n} \| \tilde{X} \vec{v} \|^2$$

$\tilde{x}_i$ is row i of $\tilde{X}$

need to account for fact that $\vec{v}$ is unit vector

this is a constraint

one solution:

maximize

$$f(\vec{v}) = \frac{\|\tilde{X}\vec{v}\|^2}{\|\vec{v}\|^2}$$

effectively restricts $\vec{v}$ to unit vectors

$$= \frac{\vec{v}^T \tilde{X}^T \tilde{X} \vec{v}}{\vec{v}^T \vec{v}}$$

$$\nabla f(\vec{v}) = \frac{2}{\vec{v}^T\vec{v}} \left(\tilde{X}^T \tilde{X} \vec{v} - f(\vec{v})\vec{v} \right) = \vec{0}$$

$$\nabla f(\vec{v}) = \frac{2}{\vec{v}^T\vec{v}} \underbrace{\left(\tilde{X}^T\tilde{X}\vec{v} - f(\vec{v})\vec{v}\right)}_{} = \vec{0}$$

for this to be $\vec{0}$,

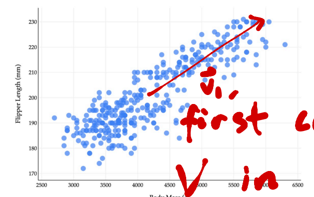
$\vec{v}$ must be eigvec of $\tilde{X}^T\tilde{X}$

i.e. singular vec of $\tilde{X}$ in

$$\tilde{X} = U\Sigma V^T!$$

I'm not suggesting that we just select some of the original features and drop the others; rather, I'm proposing that we find 1 or 2 new features that are **linear combinations** of the original features. If everything works out correctly, we may be able to **reduce the number of features** we need to deal with, without losing too much information, and without the interpretability issues that come with multicollinearity.

To illustrate what I mean by constructing a new feature, let's suppose we're starting with a 2-dimensional dataset, which we'd like to reduce to 1 dimension. Consider the scatter plot below of `flipper_length_mm` vs. `body_mass_g` (with colors removed).



$\vec{v}_1$, first col of V in $X = U\Sigma V^T$

Above, each penguin is represented by two numbers, which we can think of as a vector

$$\vec{x}_i = \begin{bmatrix} \text{body mass}_i \\ \text{flipper length}_i \end{bmatrix}$$

The fact that each point is a vector is not immediately intuitive given the plot above: I want you to imagine drawing arrows from the origin to each point.

Now, suppose these vectors $\vec{x}_i$ are in the rows of a matrix X, i.e.

$$X = \begin{bmatrix} \text{body mass}_1 & \text{flipper length}_1 \\ \text{body mass}_2 & \text{flipper length}_2 \\ \vdots & \vdots \\ \text{body mass}_n & \text{flipper length}_n \end{bmatrix}$$

Our goal is to construct a single new feature, called a **principal component**, by taking a linear combination of both existing features.

$$\text{new feature}_i = a \cdot \text{body mass}_i + b \cdot \text{flipper length}_i$$

This is equivalent to approximating this 2-dimensional dataset with a single line. (The

key point: the "best direction" is the singular vector $\vec{v}_1$ with the largest sing. value, σ_1!

side quest

$$f(\vec{v}) = \frac{\vec{v}^T A \vec{v}}{\vec{v}^T \vec{v}}$$

A symmetric

$$\nabla f(\vec{v}) = \frac{2}{\vec{v}^T \vec{v}} \left(A\vec{v} - f(\vec{v})\,\vec{v} \right) = \vec{0}$$

$$A\vec{v} - f(\vec{v})\,\vec{v} = \vec{0}$$

critical points are eigvecs of A!

$$A\vec{v} = f(\vec{v})\,\vec{v}$$

to max f, pick $\vec{v}$ with biggest λ!