Unsupervised Learning

First half: supervised problems

→ a prediction problem:
use covariates (Xs) to predict a response (Y)

→ training data to supervise a $\hat{f}$

finding/summarizing patterns in the data

→ no distinct of Xs vs. Y
all we have is Xs

→ refine data in an interpretable way

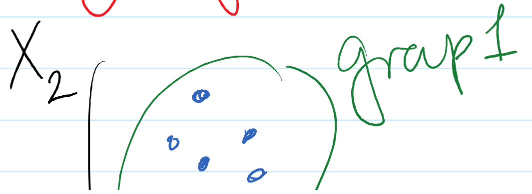
Ex. (1) dimensionality reduction:

represent the data using a smaller number of features

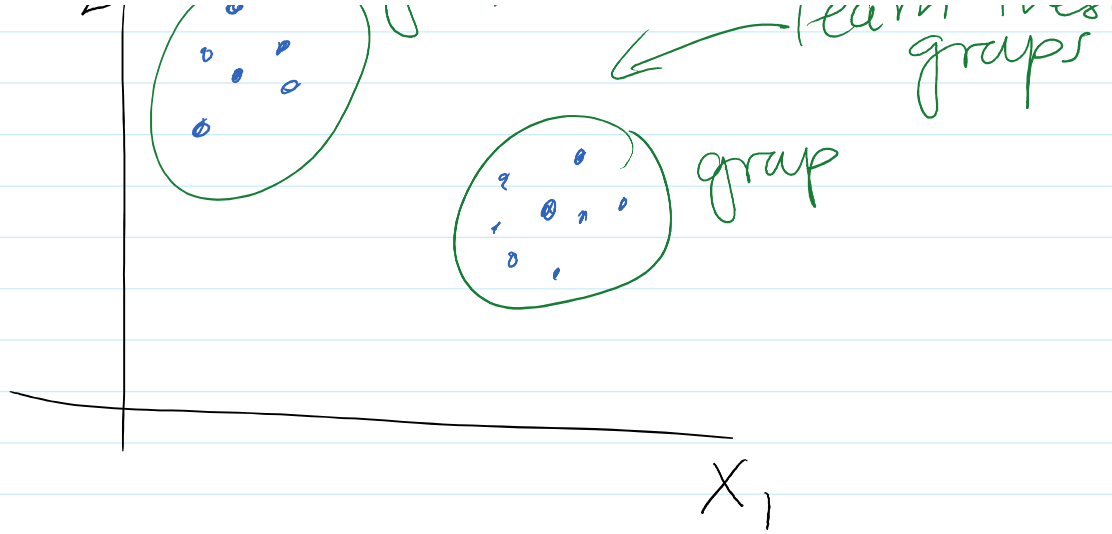
$$P \text{ covariates} \xrightarrow[\text{learning}]{\text{unsupervised}} q \ll P \text{ features}$$

that approx. descr. the data.

(2) Clustering: group data into similar clusters



learn these groups



Today: Principal Components Analysis (PCA)

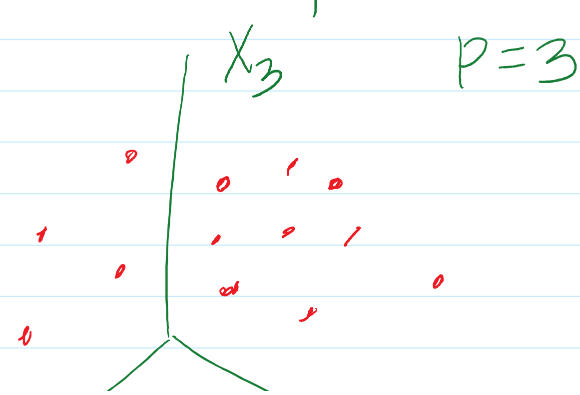
Technique for dim'l reduction.

Unsupervised setting: data matrix X

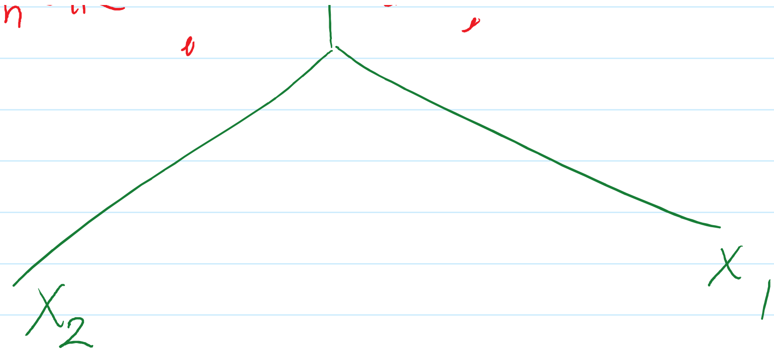
$$\underset{N \times P}{X} = \begin{bmatrix} \text{— obs 1 —} \\ \text{— obs 2 —} \\ \text{— obs 3 —} \\ \vdots \\ \text{— obs } N \text{ —} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \text{Var 1} & \text{Var 2} & \cdots & \text{Var p} \\ | & | & & | \end{bmatrix}$$

Visualize: Feature space / variable space ($\mathbb{R}^P$)

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \end{bmatrix} \quad \text{where } x_n \in \mathbb{R}^P$$



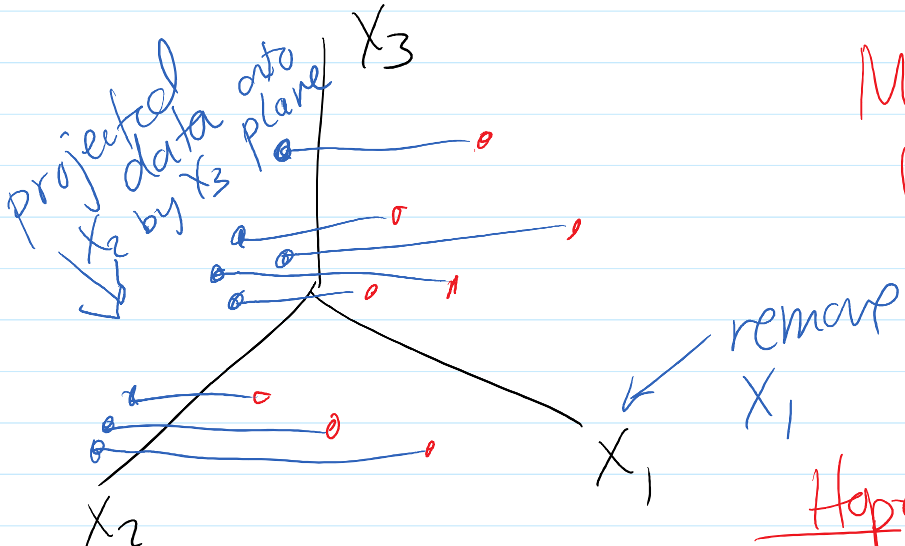
[— x_N —]



think of data as living in some p-dim'l space.

Dimensionality reduction: can I get away w/ fewer dimensions?

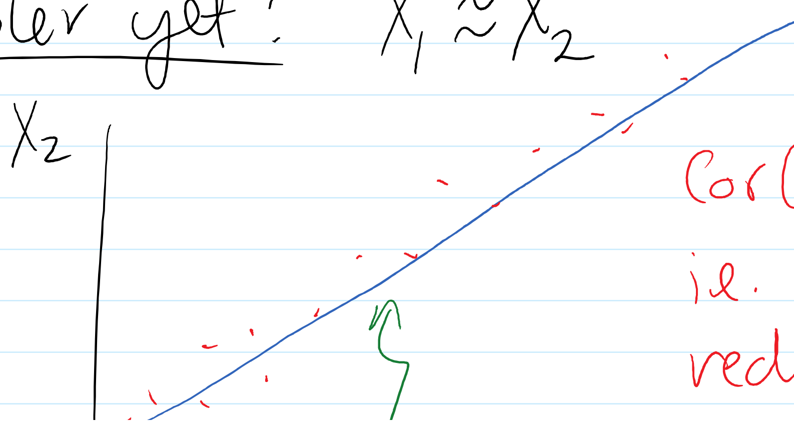
Simple Case: Variable selection



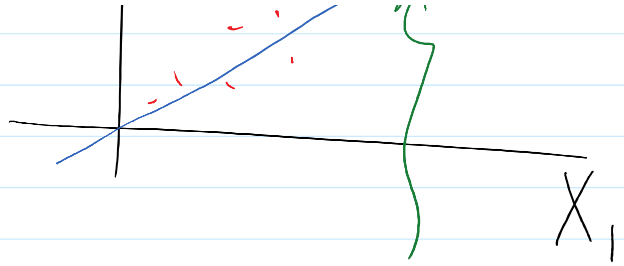
Maybe I don't lose much if I remove X_1 from my data

Hope: made my data simpler w/o losing much info.

Simpler yet: $X_1 \approx X_2$



$Cor(X_1, X_2)$ high, i.e. X_1 and X_2 are redundant.



redundant.

X_1 and $X_2 \rightsquigarrow \frac{1}{2}(X_1+X_2)$

Goals of PCA:

(1) reduce the data (# of vars) dimension

$X_1, \dots, X_p \xrightarrow{\text{PCA}} Z_1, \dots, Z_q$ where $q \ll p$.

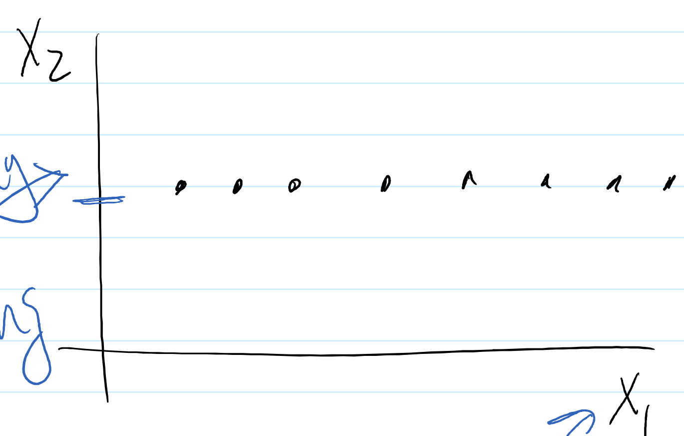
(2) not lose too much info.

equiv.: maximize the amount of info retained

min lose $\iff$ max retain

Central dogma: Variance = interestingness or info

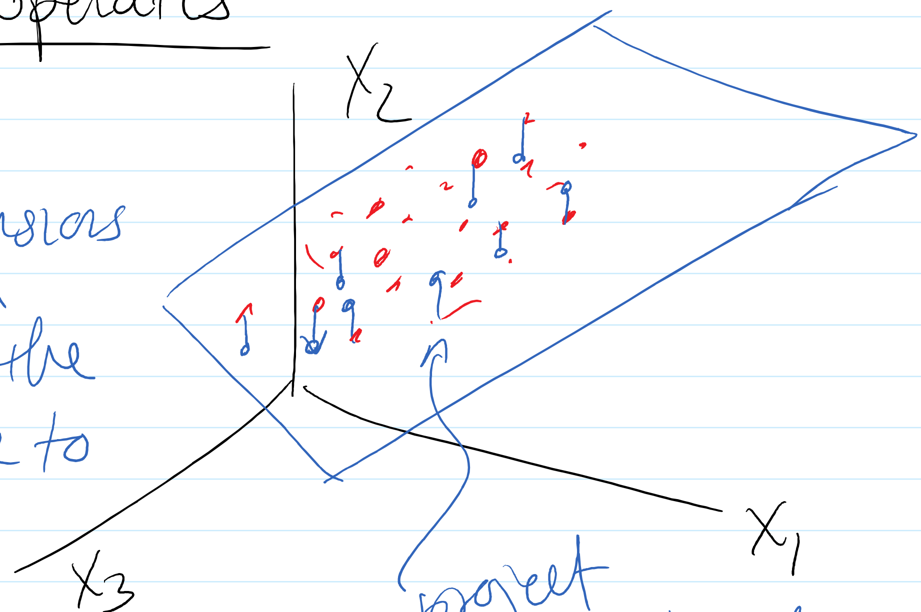
tells us basically nothing interesting



tells us something about the data

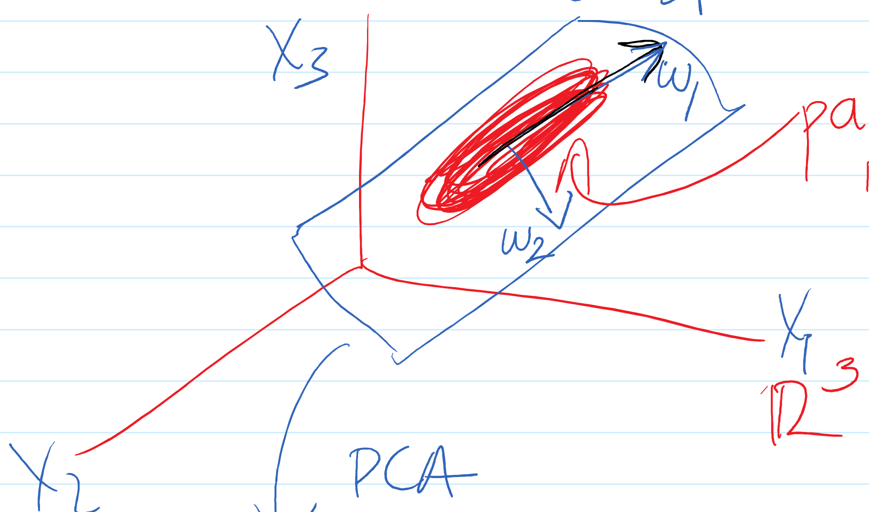
How PCA operates

Instead of removing dimensions of our data aligned w/ the axes, reduce to any good subspace

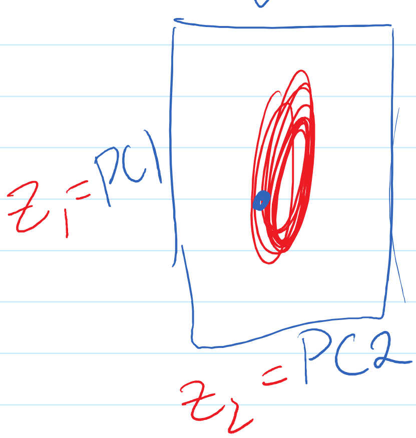


project data onto this subspace.

Idea!



pancake in 2-D



work w/ these PCs instead

PCA:

PCA:

Take orig. $X_1, \dots, X_p \xrightarrow{\text{create}} Z_1, \dots, Z_q$

If w_i is the i^{th} basis vector for this subspace then

$$\boxed{\underset{N\times q}{Z} = \underset{N\times p}{X} W}$$

change of basis mtx

where $W = \begin{bmatrix} | & | & & | \\ w_1 & w_2 & \cdots & w_q \\ | & | & & | \end{bmatrix}$

alt.

ith PC

$$Z_i = Xw_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{ip}X_p$$

Alternative Formulation:

Let Z_i = LCs of my $X_1, \dots, X_p$

where Z_1 maximizes variance

Z_2 max. variance s.t. $\mathrm{cor}(Z_1, Z_2) = 0$

Z_3 max. var. s.t., it is ...

uncorr. w/ z_1 and z_2

⋮

z_q max. var. s.t. uncorr w/ rest.

<u>Mathematically</u>

$X_{N \times p}$ we want to find $W = \begin{bmatrix} | & & | \\ w_1 & \cdots & w_q \\ | & & | \end{bmatrix}$

So that we maximize PC$_i$

$$\mathrm{Var}(z_i) = \mathrm{Var}(X w_i)$$ weights

subject to the constraint that $\mathrm{Cor}(z_i, z_j) = 0$ for all $j < i$.

<u>Constraint:</u> W is orthogonal

equiv. w_i's are unit vectors.

<u>Review:</u> $x = (x_1, \ldots, x_N)$ (some variable/col of X)

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{1}{N} \mathbb{1}^T x = \frac{1}{N} x^T \mathbb{1}$$

where $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^N$

WLOG: assume $\bar{X} = 0$ (if not we'll center our data)

PCA assume mean-centered data.

$$\text{Var}(x) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 = \frac{1}{N}\sum_n x_n^2 = \frac{1}{N}\|x\|^2$$
$$= \frac{1}{N}x^T x$$

If $\bar{X}=0$ then $\|\cdot\|^2 \approx \text{Var}$

If $y \in \mathbb{R}^N$ and $\bar{y} = 0$ then

$$\text{Cov}(x,y) = \frac{1}{N}\sum_n x_n y_n = \frac{1}{N}x^T y$$

If $\bar{X} = \bar{y} = 0$ then inner product $\approx$ Covariance

So if $x^T y = 0 \Leftrightarrow \text{Cov}(x,y) = 0 \Leftrightarrow \text{Cor}(x,y) = 0$

Simplify PCA:

$$\text{Var}(Xw_i) = \text{Var}(z_i)$$

$$(*)\quad w_1 = \underset{\|w\|=1}{\text{argmax}}\ \|Xw\|^2$$

$$w_2 = \underset{\|w\|=1}{\text{argmax}}\ \|Xw\|^2$$

$$w^T w_1 = 0$$

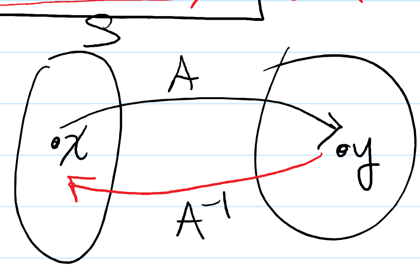
$$w_3 = \underset{\substack{\|w\|=1 \\ w^T w_1 = 0 \\ w^T w_2 = 0}}{\operatorname{argmax}} \|Xw\|^2$$

$$\vdots$$

$$w_\ell = \cdots$$

Aside: if $\boxed{A \text{ is an invertible matrix}}$ and

$$\boxed{x^* = \underset{x \in S}{\operatorname{argmax}}\, f(x)}$$



$$y^* = \underset{y \in A^{-1}S}{\operatorname{argmax}}\, f(Ay)$$

find $\boxed{x^* = Ay^*}$

Ex. U orthogonal mtx $\Leftrightarrow U^{-1} = U^T$

$$\underset{\|z\|=1}{\operatorname{argmax}}\, f(z) = U^T \underset{\|U^T z\|=1}{\operatorname{argmax}}\, f(Uz)$$

Ex. $$\underset{z^T a = 0}{\operatorname{argmax}}\, f(z) = U^T \underset{z^T U^T a = 0}{\operatorname{argmax}}\, f(Uz)$$

PC1 $$w_1 = \operatorname{argmax} \|Xw\|^2$$

PC1

$$\omega_1 = \underset{\|w\|=1}{\arg\max} \|Xw\|^2$$

Let $X = UDV^T$ $\qquad \|Ua\| = \|a\|$ (showed prev.)

$$= \underset{\|w\|=1}{\arg\max} \|UDV^Tw\|^2$$

$$= \underset{\|w\|=1}{\arg\max} \|DV^Tw\|^2$$

$$= V^T \underset{\|V^Tw\|=1}{\arg\max} \|DV^TVw\|^2$$

$$D = \left[\begin{array}{cccc|c} \sigma_1 & & & & \\ & \sigma_2 & & & 0 \\ & & \ddots & & \\ & & & \sigma_r & \\ \hline & & 0 & & 0 \end{array}\right]$$

$$= V^T \underset{\|w\|=1}{\arg\max} \underbrace{\|Dw\|^2}_{\left(\cancel{w_1}\,\sigma_1 + \cancel{w_2\sigma_2 + \cdots w_r\sigma_r}\right)^2}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots$

set $w_1 = 1$

others $= 0$

$$= V^T \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = V_1 \leftarrow \text{first col of } V.$$

PC2

$$\omega_2 = \underset{\substack{\|w\|=1 \\ V_1^Tw=0}}{\arg\max} \|Xw\|^2$$

$$= \underset{\|w\|=1}{\arg\max} \|\cancel{U}DV^Tw\|^2$$

$$= \underset{\substack{\|w\|=1 \\ v_1^T w = 0}}{\operatorname{argmax}} \|\cancel{U} D V^T w\|^2$$

$$v_1^T w = 0$$

$$\downarrow$$

$$v_1^T V^T w$$

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T w = 0$$

$$= V^T \underset{\substack{\|w\|=1 \\ v_1^T V^T w = 0}}{\operatorname{argmax}} \|Dw\|^2$$

simplify to constraint that first comp. of $w = 0$

$$= V^T \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = V_2 \leftarrow \text{col 2 of } V.$$

Keep playing this game.

$$W_{N \times q} = V_{1:q} \leftarrow \text{first } q \text{ cols of } V$$

when $X = UDV^T$.

<u>Punchline:</u> <u>PCA procedure:</u>

1. mean center cols of X (call it X_c)
2. let $X_c = UDV^T$

3. W = first q cols of $V = V_{1:q}$

4. $Z_{N \times q} = XW = XV_{1:q}$