Principal Components Regression (PCR)

Instead of regressing Y onto $X_{N \times P}$,

regress Y onto $Z_{N \times q}$

PCs

Steps for PCR

(0) mean center X

$$X_c = \begin{bmatrix} X_1 - \hat{\mu}_1 & X_2 - \hat{\mu}_2 & \cdots & X_p - \hat{\mu}_p \end{bmatrix}$$

$$\hat{\mu}_i = \text{mean}(X_i)$$

(1) do PCA on $X_c = UDV^T$

$$Z = X_c V_{1:q}$$

(2) regress Y onto Z
(typically include intercept)

$$D = \begin{bmatrix} | & | \\ \mathbf{1} & Z \\ | & | \end{bmatrix}$$

$$\hat{\beta}^{(PCR)} = (D^T D)^{-1} D^T Y \in \mathbb{R}^{q+1}$$

What about new data?

Let X^{test} be $M \times P$

For training: $\hat{Y} = D \hat{\beta}^{(PCR)}$

Need to calc. D^{test} by apply some procedure to X^{test}

(0) mean center X^{test}

$$X_c^{test} = \left[X_1^{test} - \hat{\mu}_1 \quad \cdots \quad X_p^{test} - \hat{\mu}_p \right]$$

$$X_c = [X_1 - \mu_1 \quad \cdots \quad X_p - \mu_p]$$

from training data

① apply PCA:

$$Z^{test} = X_c^{test} V_{1:q}$$

from training

② $$D^{test} = \begin{bmatrix} | & | \\ 1 & Z^{test} \\ | & | \end{bmatrix}$$

③ $$\hat{Y}^{test} = D^{test} \hat{\beta}_q^{(PCR)}$$

from train.

Compare PCR w/ ridge regression

$$\hat{Y}^{(ridge)} = X \hat{\beta}^{(ridge)}$$

$$\vdots$$

up scaling cols of Z

$$= \sum_{j=1}^{p} \underbrace{\left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)}_{\Delta_j} u_j u_j^T Y$$

Can also show:

$$\hat{Y}^{(PCR)} = Z\hat{\beta}^{(PCR)} = \cdots = \sum_{j=1}^{p} \Delta_j u_j u_j^T Y$$

$$\Delta_j = \begin{cases} 1, & j \le q \\ 0 & j > q \end{cases}$$

OLS: $\Delta_j = 1$

ridge: $\Delta_j = \dfrac{\sigma_j^2}{\sigma_j^2+\lambda}$

PCR: $\Delta_j = \mathbb{1}(j \le q)$

Consider X to be full rank

Then as $\lambda \to 0$ we have

Then as $\lambda \to 0$ we have

$$\hat\beta^{(ridge)} \longrightarrow \hat\beta^{(OLS)}$$

Similarly, as $q \to P$ then

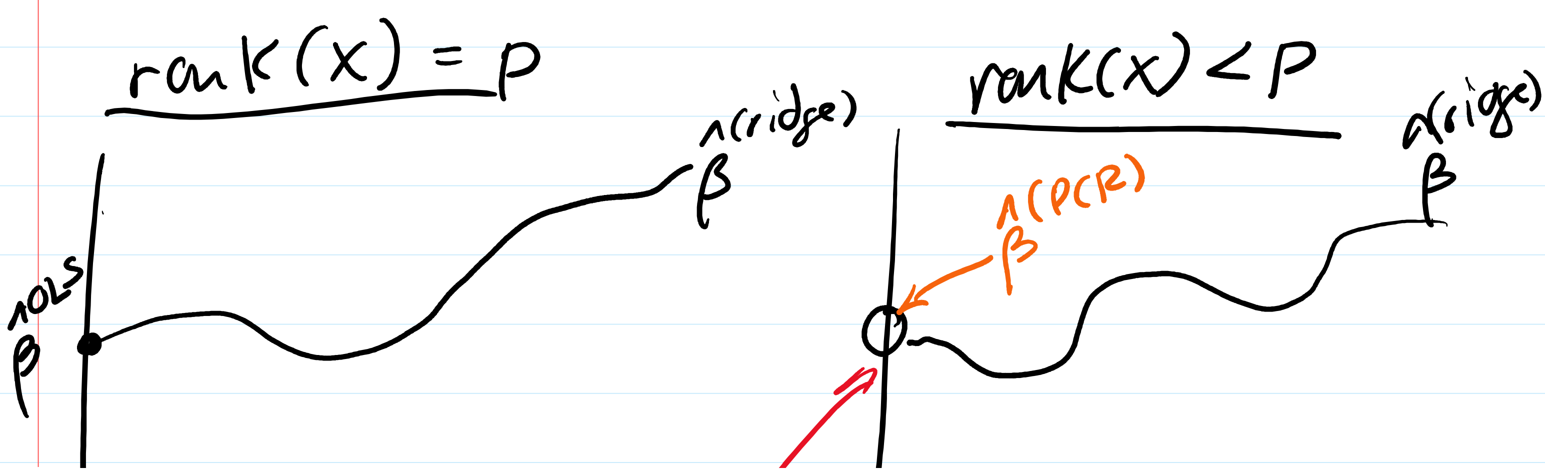
$$\hat\beta^{(PCR)} \longrightarrow \hat\beta^{(OLS)}$$

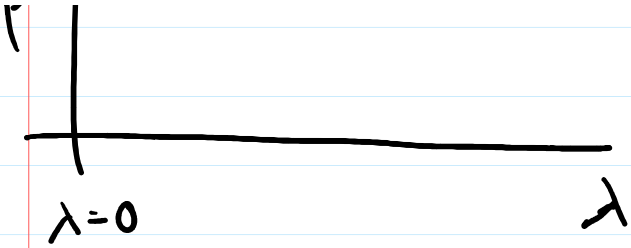
<u>If $\operatorname{rank}(X) < P$</u>

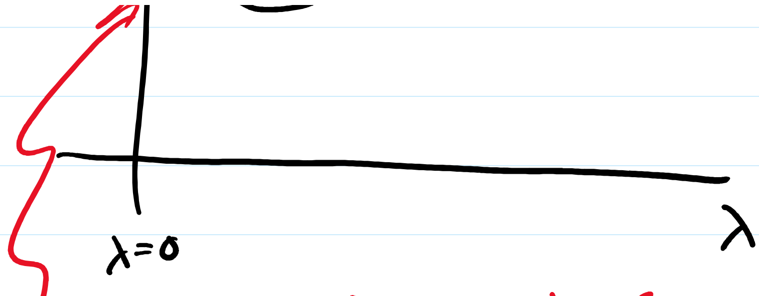
Then OLS doesn't exist.

(Equiv. ridge w/ $\lambda = 0$ doesn't exist)

However $\lim_{\lambda \to 0} \hat\beta^{(ridge)} = \hat\beta^{(PCR)}$ w/ $q = \operatorname{rank}(X)$





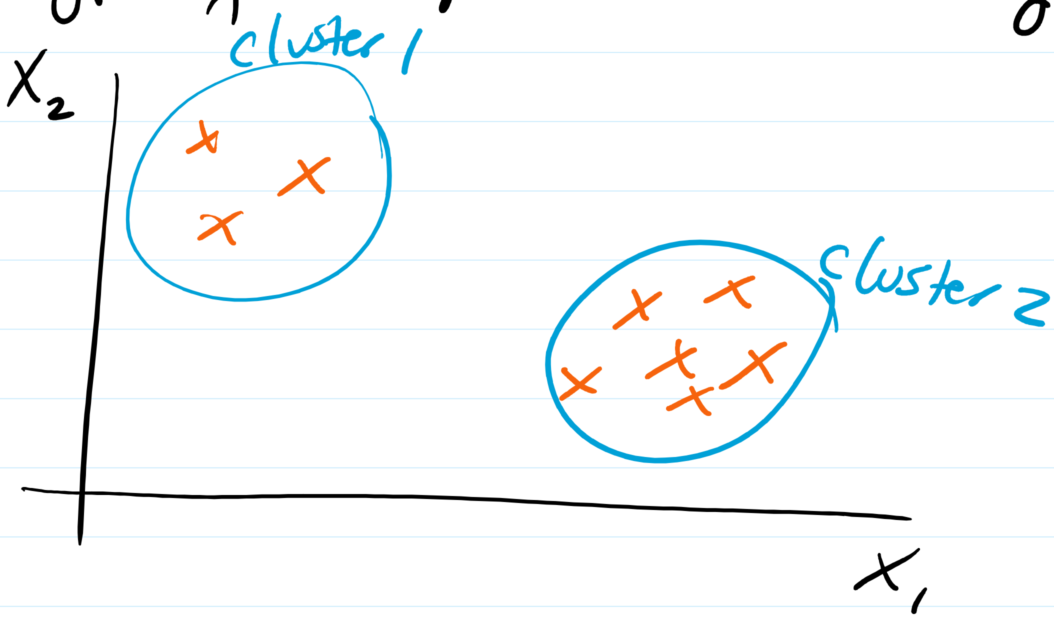


$\lim_{\lambda \to 0} \hat{\beta}^{(ridge)}$ = ridgeless regression

---

Back to unsupervised

Second type of unsupervised: clustering



To find clusters I need to define how pts are similar/dissim to each other

Dissimilarity Measure

Dissimilarity Measure

If I have N obs then I can define a dissim mtx

$$D \qquad (N \times N)$$

where

$D_{ii'}$ = dissim meas betwn obs. i and i'.

For many clustering methods — don't need X but only D.

Properties of D:

1. $D_{ii} = 0$
2. $D_{ii'} \geq 0$
3. $D = D^T$

K-means clustering

Assume that each data pt belongs to one of K clusters (or groups):

Assume that each data pt belongs to one of K clusters (or groups):

$$G_1, \dots, G_K$$

<u>want</u>: assign each point i to some cluster G_k

<u>how</u>: want to make assignments so that we minimize some measure of not being well clustered (loss)

classic way to measure this is

$$W = \text{total within-cluster dissimilarity}$$

$$= \sum_{k=1}^{K} \sum_{i,i' \in G_k} D_{ii'}$$

small if clustering is good
large // bad

T = total dissim $= \sum_{i,i'} D_{ii'}$

B = total between cluster dissim

$$= \sum_{k \neq k'} \sum_{i \in G_k} \sum_{i' \in G_{k'}} D_{ii'} .$$

Can show: $T = B + W$

So, to find $G_1, \dots, G_K$

(1) minimize W

or (2) maximize B

Ideally: try all possible cluster assignments

practically: not possible — too computationally demanding

Assume all data is numeric: $x_i \in \mathbb{R}^p$

Assume all data is numeric : $x_n \in \mathbb{R}^p$

define

$$D_{ii'} = \|x_i - x_{i'}\|_2^2$$

If I do this then

$$W = \sum_{k=1}^{K} 2N_k \overbrace{\sum_{i \in G_k} \|x_i - \bar{x}_k\|^2}^{\text{dist of pt to mean of cluster}}$$

$N_k = |G_k|$

mean of pts in cluster k

Goal: try to minimize this W

Lloyd's Algorithm

⓪ Initialization Step

make some random intializations of cluster means:

$$\underbrace{\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_K^{(0)}}$$

For $t = 1, 2, 3, \dots$

at the t^{th} iteration:

① Assignment Step:

assign i (or x_i) to cluster G_k w/ the closest mean to x_i

$$G_k^{(t)} = \{x_i : \|x_i - \mu_k^{(t)}\| \le \|x_i - \hat{\mu}_{k'}^{(t)}\| \ \forall k'\}$$

② Update Step: re-calculate μs given current assignments

$$\hat{\mu}_k^{(t+1)} = \underset{i \in G_k^{(t)}}{\text{mean}}\ x_i$$