Lecture 19: PCAR

PCA regression, instead of regressing Y onto $X_{N \times P}$

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

(if $q \ll P$, this is basically some "smart" variable selection)

Steps for PCAR

(0) mean center X

$$X_c = \left[X_1 - \text{mean}(X_1) \quad X_2 - \text{mean}(X_2) \quad \cdots \right]$$

(cols of X)

(1) Do PCA: $X_c = UDV^T$

$$Z = X_c V_q$$

(first q cols of V)

(2) regress Y onto Z

typically want to include intercept,

$$D = \left[\underset{|}{1} \quad \underset{|}{Z} \right]$$

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

What about prediction on new data?

$$X^{test} \in \mathbb{R}^{M \times P}$$

For train, $\hat{Y} = D\hat{\beta}^{PCR}$

Careful and do the same processing steps to X^{test}

(0) Center test data

$$X_c^{test} = \left[X_1^{test} - \underbrace{mean(X_1)}_{1} \quad X_2^{test} - \underbrace{mean(X_2)}_{1} \cdots \right]$$

(1) apply PCA

$$Z^{test} = X_c^{test} V_q$$

(2) $\hat{Y}^{test} = Z^{test} \hat{\beta}^{PCR}$

$M \times q \quad q \times 1$

Comparison w/ Ridge

$$\hat{\beta}^{(ridge)} = (X^T X + \lambda I)^{-1} X^T Y$$

$$\hat{Y}_{train}^{ridge} = X\hat{\beta}^{ridge} = X(X^T X + \lambda I)^{-1} X^T Y$$

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

u_j = left singular vecs. of X

σ_i = sing. vals of X

$$= \sum_{j=1}^{p} \Delta_j u_j u_j^T Y \quad \text{where } \Delta_j = \frac{\sigma_j^2}{\sigma_j^2 + \lambda}$$

↖ proj. Y onto PCs and take weighted avg. w/ weights Δ_j

PCR: $\hat{Y}_{train} = Z \hat{\beta}^{PCR}$ $\quad Z = XV_q = U_q D_q$

$$= Z(Z^T Z)^{-1} Z^T Y$$

$$= U_q D_q (D_q U_q^T U_q D_q)^{-1} D_q U_q^T Y$$

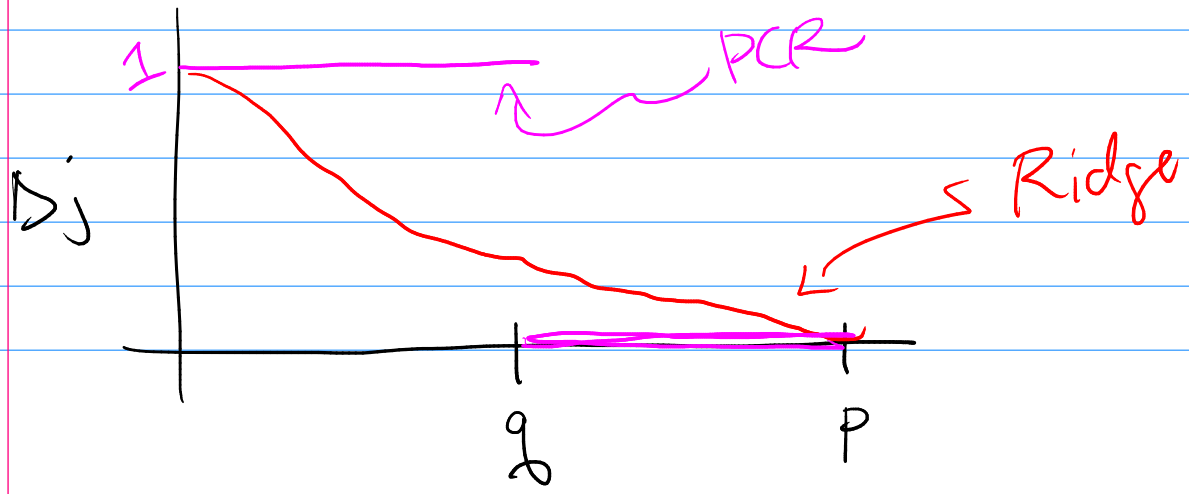
($U_q^T U_q = I$, $D_q D_q = I$)

$$= U_q U_q^T Y$$

$$= \sum_{j=1}^{q} u_j u_j^T Y$$

$$= \sum_{j=1}^{p} \Delta_j^{PCR} u_j u_j^T Y$$

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



If X is full (ol rank ($\text{rank } X = P$)

then as $\lambda \to 0$

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

similarly, as $q \to P$

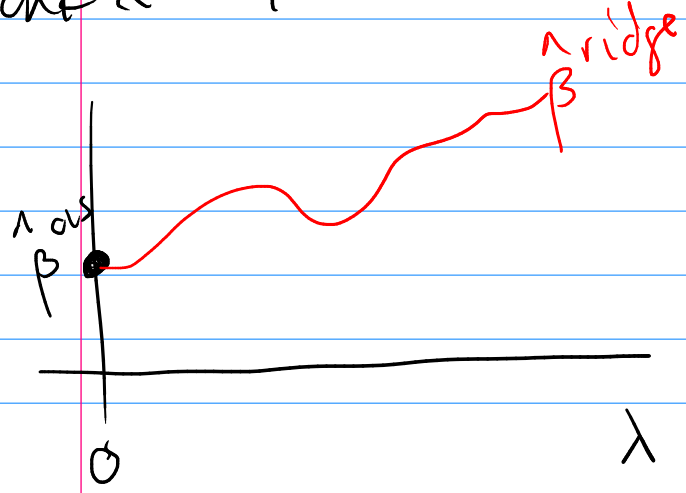
$$\hat{\beta}^{PCR} \longrightarrow \hat{\beta}^{OLS}$$

Hower, if $\text{rank } X < P$. then $\hat{\beta}^{OLS}$ dosn't exist

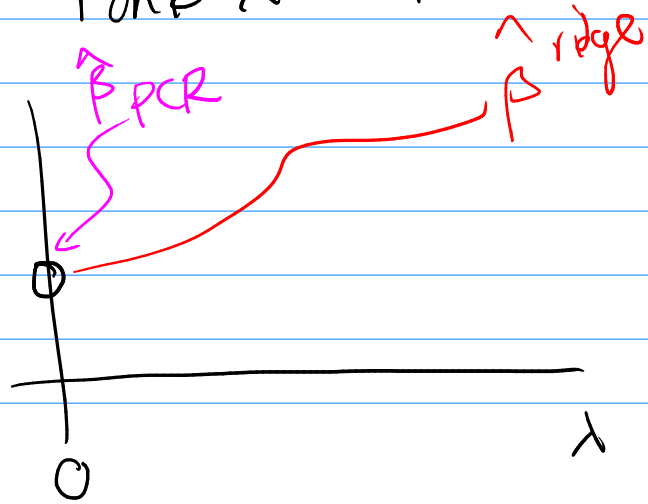
So $\hat{\beta}^{ridge}$ dosn't exist at $\lambda = 0$,

however, $\lim_{\lambda \to 0} \hat{\beta}^{ridge} = \hat{\beta}^{PCR}$ w/ $q = \text{rank } X$.

$\text{rank } X = P$



$\text{rank } X < P$



Back to unsupervised learning

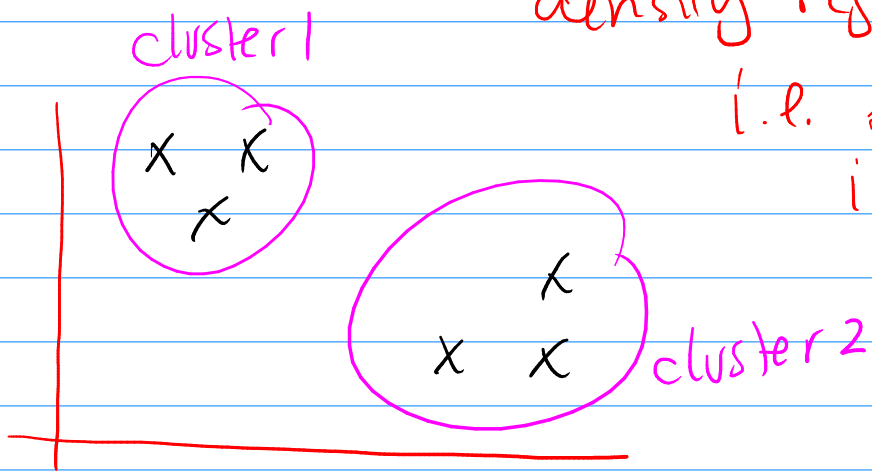
- Supervised problems: interested in $p(x,y)$

unsupervised problems: interested in $p(x)$

PCA: find a subspace where $p(x)$ is concentrated

clustering: want to find high density regions of $p(x)$

i.e. find "clusters" in our data



Goal: automatically find clusters

To do clustering we need some measure of either similarity or dissimilarity among observations

If I have N observations I need a matrix of $(N \times N)$

Called the dissim. mtx where

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

[Can create dissim w/ dec. trans. of a sim. metric]

Most/all clustering algos only need D, not a data mtx X

Properties of D

1. $D_{ii} = 0$ (diag. elements are zero)
2. $D_{ii'} \geq 0$ (non-neg)
3. $D = D^T$ (symmetric)

Ex. If I have a data mtx $X_{N \times P}$ I can calc. "attribute-based" dissims

$$D_{ii'} = d(x_i, x_{i'}) = \sum_{j=1}^{P} d_j(x_i, x_{i'})$$

↑ meas. for var j

caveat:
careful w/ scale

Numeric

Ex. $d_j(x_i, x_{i'}) = (x_{ij} - x_{i'j})^2$

↑ euclidean

Ex. Categorical

$$d_j(x_i, x_{i'}) = \mathbb{1}(x_{ij} = x_{i'j})$$