Principal Components (Analysis) Regression

Unsupervised Techniques:

(1) Dimensionality Reduction

$$X_1, \dots, X_p \rightsquigarrow Z_1, \dots, Z_q \text{ where } q \ll p$$

thereby we go from p dims to $q \ll p$ dims.

(2) PCA summarize the Xs into Zs that are Linear Combs of Xs:

$$Z_j = \sum_{i=1}^{p} \omega_{ij} X_i = X\omega_j$$

i.e. $Z = XW$

$$\begin{bmatrix} | & & | \\ Z_1 & \cdots & Z_q \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ X_1 & \cdots & X_p \\ | & & | \end{bmatrix} \begin{bmatrix} | & & | \\ W_1 & \cdots & W_q \\ | & & | \end{bmatrix}$$

$N \times q \qquad N \times P \qquad P \times q$

(3) Capture as much "info" in the orig. Xs in the new Zs

"info" = Variance

So $\omega_1 = \max Var(Z_1)$ $\quad [Z_1 = X\omega_1]$

$\|w\|=1$

otherwise

$Var(1000\, Xw) = 1000^2\, Var(Xw)$

$$w_2 = \max_{\substack{\|w\|=1 \\ cor(z_1, z_2)=0}} Var(z_2)$$

don't capture same info (parsimoniously reducing dim)

$$w_3 = \max_{\substack{\|w\|=1 \\ cor(z_1, z_3) = cor(z_2, z_3) = 0}} Var(z_3)$$

etc. up to $r = rank(X)$

<u>Soln</u>: $W = V_{1:q}$ where $X = UDV^T$

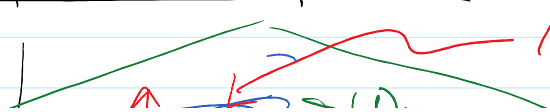
and $Z = X V_{1:q}$.

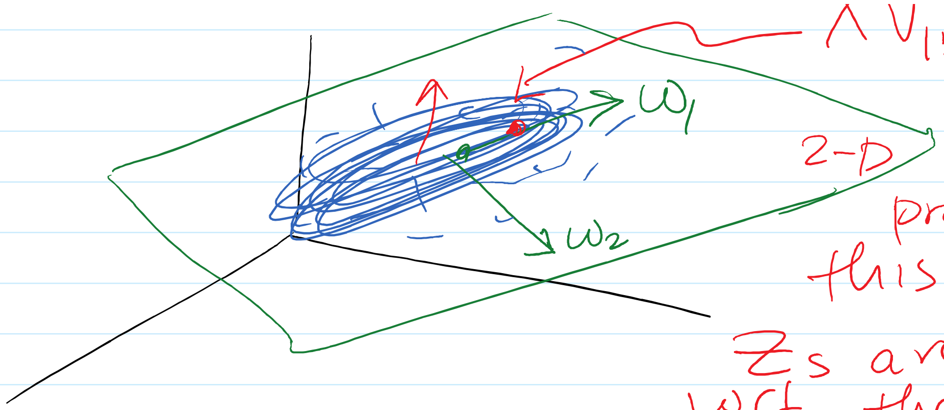
<u>Typically,</u> we ① mean center X_j's

Sometimes ② re-scale X_j's by their s.d.

<u>One more interpretation of PCA:</u>

$XV_{1:q} = Z$





$\wedge V_{1:q} = Z$

proj. data onto this subspace

Zs are coords wrt. the basis of this space

Instead look at the projected data in the original basis of the space:

$$X_q = X \underbrace{P_W} = \overbrace{X V_{1:q}}^{\text{coords}} V_{1:q}^T$$

proj. onto subspace

alt. If $X = U \Sigma V^T$

$$X_q = U_{1:q} D_q V_{1:q}^T = \text{truncated SVD}$$

$q \times q$ upper sub mtx

notice I have up to $r = \text{rank}(X)$ possible PCs (Z_r)

If $q = r$ then $X_q = X$

minimize info lost:

Eckhart-Young Theorem:

$$\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$$

$$X_q = \arg\min_{B:\,\mathrm{rank}(B)=q} \|X - B\|_F$$

when $q < r = \mathrm{rank}(X)$

use X_q over X how much I lose

PCR: Instead of regressing Y onto $X_{N\times p}$ lets regress Y onto $Z_{N\times q}$ $q \ll p$.

(basically variable selection (?))

Steps for PCR:

① Assume that Xs have been mean centered

mean $(X_j) = 0$

(optionally: re-scale by s.dis of X_j)

② Let $X = UDV^T$ = PCA basically

since $V_{1:q} = W$

then

$Z = XV_{1:q}$

$XV = UD$

$U_{1:q}$ basically Zs

regress Y onto Z,

want an intercept.

[often $Z = \begin{bmatrix} 1 & XV_{1:q} \\ 1 & \end{bmatrix}$] want an intercept.

i.e. $\hat{\gamma}^{(PCR)} = (Z^T Z)^{-1} Z^T Y \in \mathbb{R}^q$ [or $\mathbb{R}^{q+1}$]

③ <u>Prediction:</u>

(training data) $\hat{Y}_{train} = Z\hat{\gamma}$.

(new data) $\hat{Y}_{test} = \ldots ?$

⓪ $A_{N \times p} = [a_1 \cdots a_p]$ ← new data

① $\tilde{A} = [1 \;\; a_1 - \bar{X}_1 \cdots a_p - \bar{X}_p]$ ← mean center [opt: re-scale]

② $\tilde{A} V_{1:q}$ ← proj. into subspace

③ $\boxed{\hat{Y}_{new} = \tilde{A} \underbrace{V_{1:q}\hat{\gamma}}}$

$\underset{N \times P}{X} = \underset{N \times N}{U} D \underset{P \times P}{V^T}$

<u>Notice:</u> $\hat{\beta}^{(PCR)} \in \mathbb{R}^P$ since $V_{1:q}$ is $P \times q$

so $\hat{\gamma} \in \mathbb{R}^q$

So $\hat{\gamma} \in \mathbb{R}^q$

$$\hat{Y} = \underset{N \times p}{\tilde{A}} \; \underbrace{\hat{\beta}^{(PCR)}}_{\mathbb{R}^p \text{ (or } \mathbb{R}^{p+1})}$$

Formalize comparison to Ridge Regression

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

$X = UDV^T$ ← train data, $D = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \sigma_r \\ & & 0 \end{bmatrix}$

New data $A \in \mathbb{R}^{N \times p}$

$$\hat{Y} = A\hat{\beta}^{(ridge)} = \underbrace{AVD^{-1}}_{\mathring{A}} \overbrace{D^{-1}D}^{I} V^T \hat{\beta}^{(ridge)}$$

$\mathring{A}$ = proj. A onto V-basis and re-scale by SVs

= A under different basis

$$DV^T\hat{\beta}^{(ridge)} = DV^T(X^TX + \lambda I)^{-1} X^T Y$$
$$= DV^T(VD^2V^T + \lambda I)^{-1} VDU^T Y$$
$$= \underbrace{D(D^2 + \lambda I)^{-1} D}_{\text{Diag}\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right) = \Delta} U^T Y$$

$$\hat{Y} = \mathring{A} D V^T \hat{\beta}^{(ridge)} = \mathring{A} \Delta U^T Y$$

$$= \sum_{j=1}^{p} \mathring{A}_j \underbrace{\left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right)}_{\text{weight } \Delta_i} \underbrace{u_j^T Y}$$

$$= \sum_{j=1} \dots \underbrace{(\sigma_j^2+\lambda)^{-1}}_{\Delta_j} \dots$$

coordinate of A in PC basis

cor btwn PCs and Y

Same game for PCR

$$A\hat{\beta}^{(PCR)} = AUD^{-1}DV^{T}\hat{\beta}^{(PCR)}$$

$$= \dot{A}DV^{T}V_{1:q}\,\hat{\gamma} \qquad Z = XV_{1:q} \approx U_{1:q}D_q$$

$$= \dot{A}DV^{T}V_{1:q}(Z^TZ)^{-1}Z^TY \qquad X^TX = VD^2V^T$$

$$= \dot{A}DV^{T}V_{1:q}(V_{1:q}^{T}X^TXV_{1:q})^{-1}V_{1:q}^{T}X^TY$$

$$= \dot{A}D\overbrace{V^{T}V_{1:q}}^{I}(D_q^2)^{-1}D_qU_{1:q}^{T}Y$$

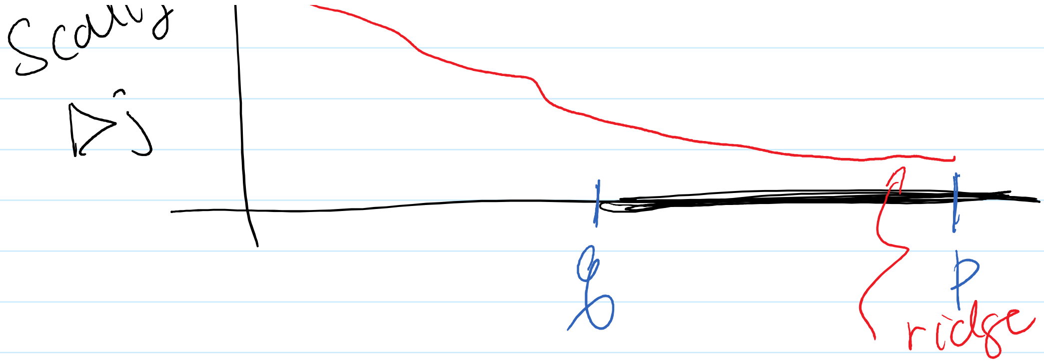
$$= \dot{A}\,\cancel{D(D_q)^{-2}D_q}\,U_{1:q}^{T}Y$$

coord of A wrt PCs

$$= \sum_{j=1}^{q} \dot{A}_j u_j^TY \quad \boxed{= \sum_{j=1}^{p} \dot{A}_j \Delta_j u_j^TY}$$

$$\Delta_j = \begin{cases} 1 & \text{if } j \le q \\ 0 & \text{else.} \end{cases}$$

PCAR:



If X is full rank (X^TX invertible)

as $\lambda \to 0$

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

Similarly as $q \to p = \text{rank}(X)$

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

<u>However</u>: If X isn't full rank, however

as $\lambda \to 0$ $\hat{\beta}^{(ridge)} \longrightarrow \hat{\beta}^{(PCR)}$

w/ $q = \text{rank}(X)$.