4주차
복습 스터디

전형준

Dimension Reduction

- Feature selection : 기존 설명 변수들 중 소수의 예측 변수만을 선택

ex) Lasso, …

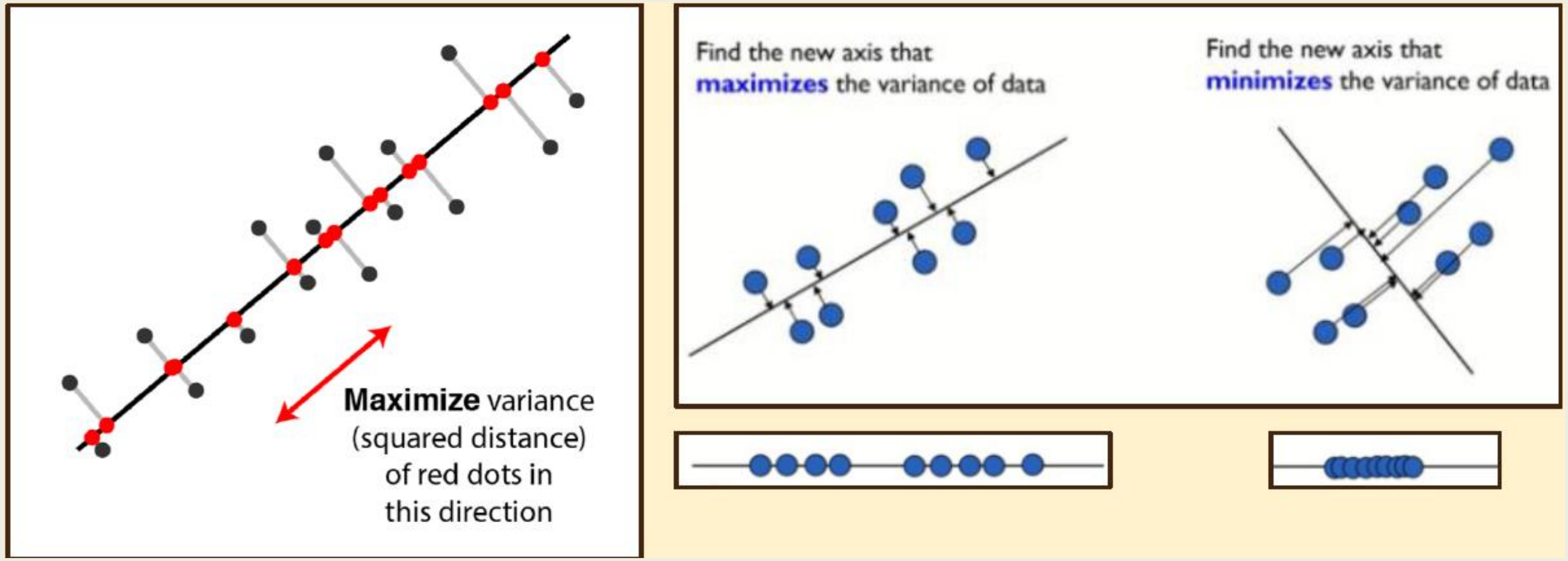
- Feature extraction : 기존 설명 변수들의 변환을 통해 새로운 예측 변수를 추출

ex) PCA (Principal Component Analysis) : unsupervised feature extraction

PLS (Partial Least Squares) : supervised feature extraction

PCA

- D개의 feature를 가진 data를 orthogonal한 q개의 변수로 구성된 데이터(PC)로 요약
- 가정 : 기존 변수들을 선형결합하여 새로운 변수 추출, 데이터들은 centered and scaled
- 원래 데이터의 분산을 최대한 보존하도록, Reconstruction error가 최소가 되도록



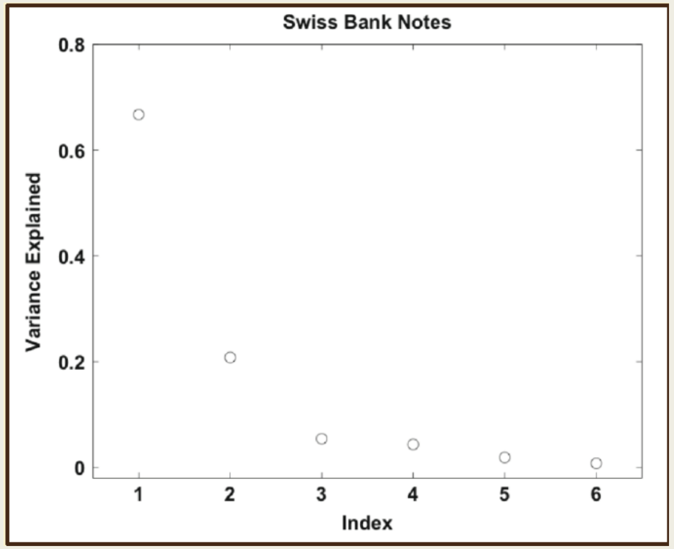
PCA

- How? $\sum(Covariance\ martix)$의 eigenvector가 가중치, eigenvalue가 분산값
- $\Sigma = \Gamma\Lambda\Gamma^T \Rightarrow Y = \Gamma^T(X - \mu)$
- In sample : $\mathcal{Y} = (\mathcal{X} - 1_n\bar{x}^T)\mathcal{G}$
- Proof : HW!

PCA

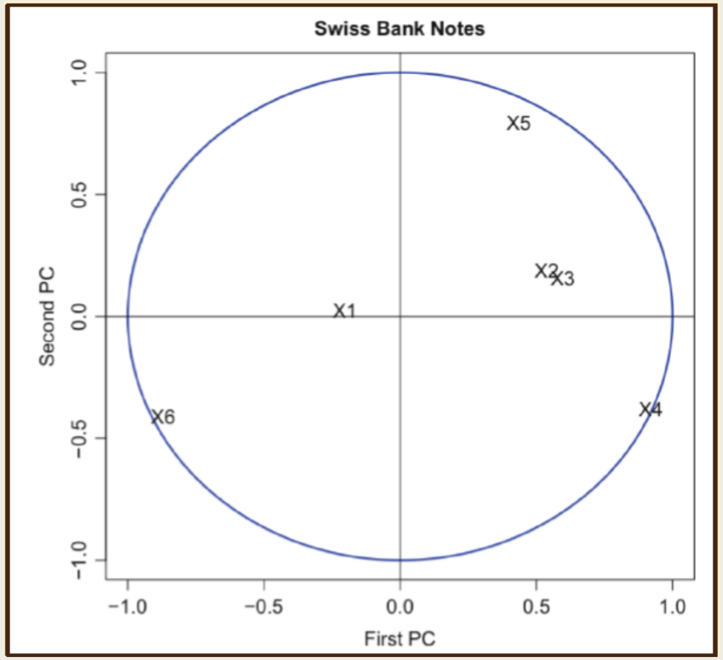
- How to select q?
- Method 1 : 고유값 감소율이 유의미하게 낮아지는 Elbow point 선택
- Method 2 : 일정 수준 이상의 분산 비를 보존하는 최소의 주성분 선택 (보통 70%)

$$\psi_q = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{q} \mathrm{Var}(Y_j)}{\sum_{j=1}^{p} \mathrm{Var}(Y_j)}.$$



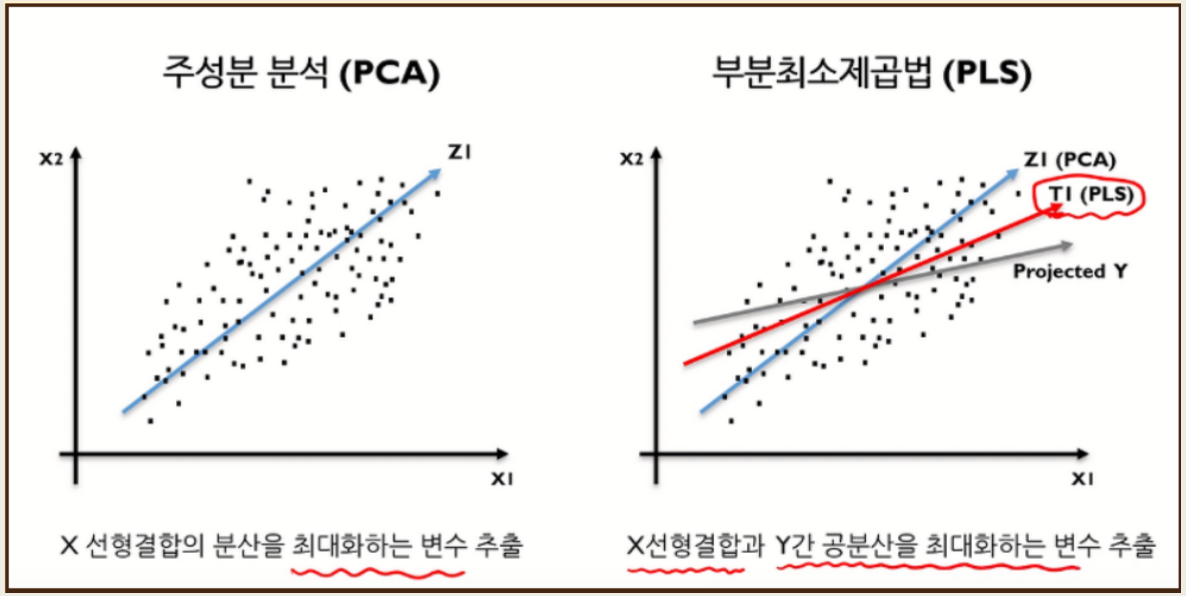
Interpretation of PC

- $\rho_{X_i Y_j} = \frac{Cov(X,Y)}{Var(X)Var(Y)} = \frac{\gamma_{ij}\lambda_j}{(\sigma_{X_i X_i}\lambda_j)^{1/2}} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{X_i X_i}}\right)^{1/2}$
- $Cov(X,Y) = E(XY^T) = E(XX^T\Gamma) = E(XX^T)\Gamma = \Sigma\Gamma = \Gamma\Lambda\Gamma^T\Gamma = \Gamma\Lambda$
- In sample, $r_{X_i Y_j} = g_{ij}\left(\frac{l_j}{s_{X_i X_i}}\right)^{1/2}$



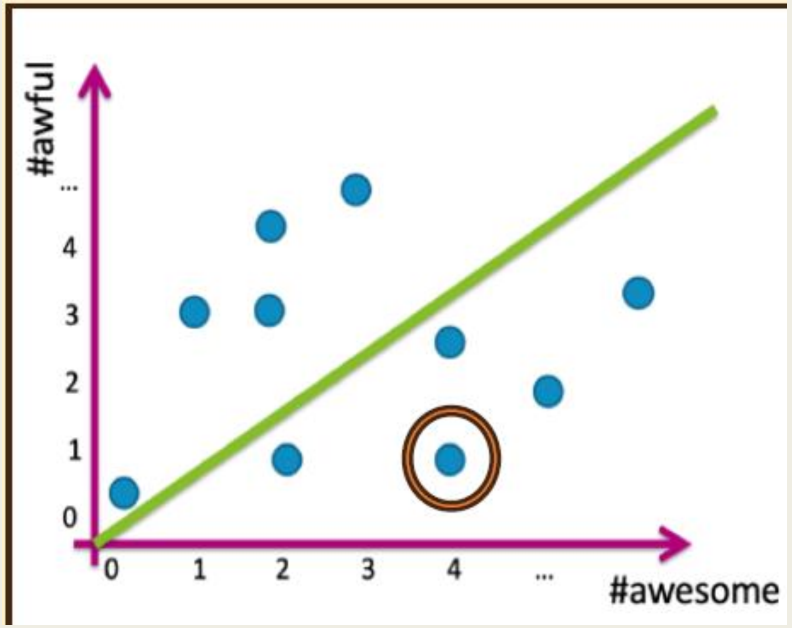
PLS

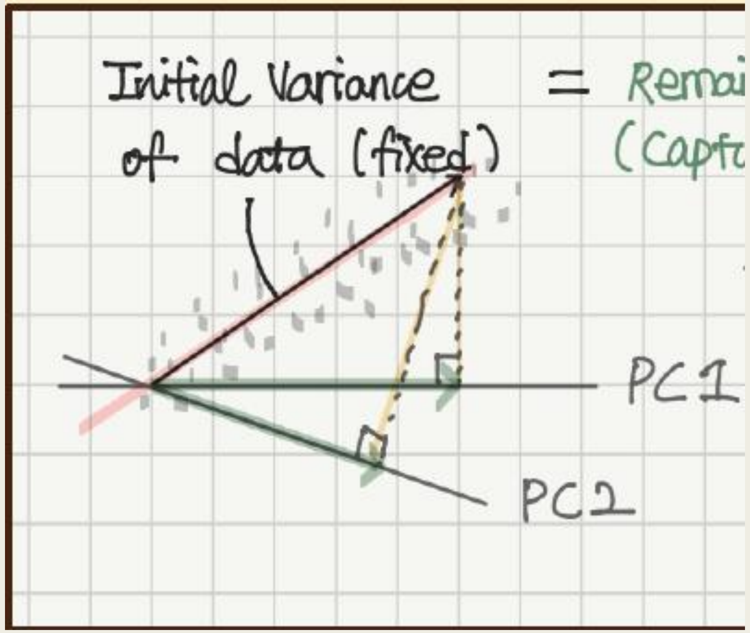
- PCA에서는 반영하지 못했던 Y와의 상관관계를 반영
- $t = Xw \ \Rightarrow Cov(t, Y) = Cov(Xw, Y) = E[(Xw \ - E[Xw])(Y - E[Y])] = E(Xw \cdot Y)$
- $E(Xw \cdot Y) = \frac{1}{n}\sum_1^n (Xw)_i \cdot Y_i = \frac{1}{n}(Xw)^T Y = \frac{1}{n} w^T (X^T Y)$
- $\therefore w = \frac{X^T Y}{\|X^T Y\|}$



PCA in terms of Reconstruction error

- $\|\hat{x} - x\|_2^2$를 최소화하는 PC를 사용
- 결국 Remaining Variance를 최대화하는 앞의 방법과 근본적으로 같음



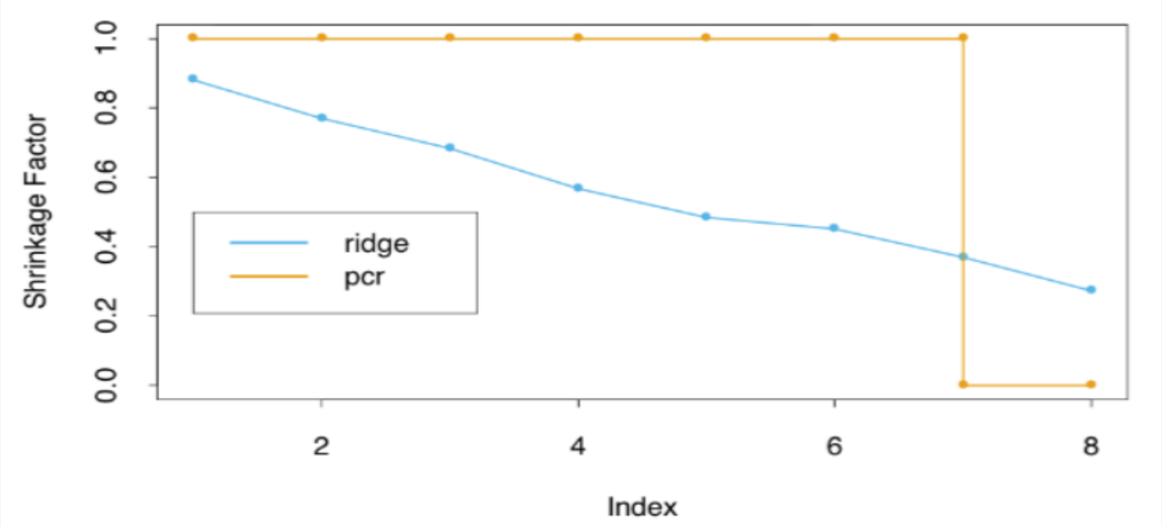


PCA in terms of Reconstruction error

- $f(\lambda) = \mu + V_q\lambda \Rightarrow min_{\mu,\{\lambda_i\},V_q} \sum_{i=1}^{n} \left\| x_i - \mu - V_q\lambda \right\|^2$
- $\hat{\mu} = \bar{x}, \hat{\lambda} = V_q^T(x_i - \bar{x}) \Rightarrow min_{V_q} \sum_{i=1}^{n} \left\| (x_i - \bar{x}) - V_q V_q^T(x_i - \bar{x}) \right\|^2$
- $V_q : first\ q\ columns\ of\ V\ where\ X = UDV^T$
- Columns of UD : principal components of X
- In terms of Unsupervised Learning : $\mathrm{L(v)} = \frac{1}{n} \sum_{i=1}^{n} \| x_i - decode(encode(x_i; V); V) \|_2^2$

PCR

- $\hat{y}_{PCR} = \bar{y} + X\hat{\beta}_{PCR}$ *where* $\hat{\beta}_{PCR} = \sum_{j=1}^{q} \hat{\theta}_j v_j$
- $\hat{Y}_{ridge} = \sum_{j=1}^{d} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T Y$
- $\hat{Y}_{PCR} = \sum_{j=1}^{q} u_j 1 u_j^T Y + \sum_{j=q+1}^{d} u_j 0 u_j^T Y$



When should / should not I use PCA?

☑ 변수를 줄이고는 싶은데, predictor들의 구조도 모르고 뭘 drop해야 할지도 잘 모르겠을 때..

☑ You want to ensure your variables are independent of one another.

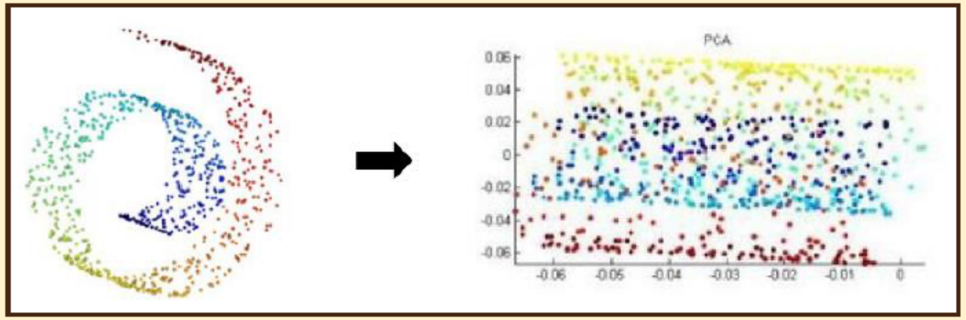
✕ You are not comfortable making your independent variables less interpretable.

(PCA 쓰려면 설명력은 포기해야...)

✕ PCA assumes there is a lower dimensional linear subspace that represents the data well.

(Doesn't work well with non-linear manifold) → solutions include t-SNE, Isomap, etc.

(PC들은 기존 변수들의 선형결합이기 때문에 비선형 data에서는 잘 작동하지 않을수도 있음.)



HW