선형대수학 in 통계(2)

오정헌

1 | 3

부분최소제곱법

$$x \cdot y = \|x\| \|y\| cos\theta$$

$$x \cdot y = \sum_{i=1}^{n} x_i y_i$$

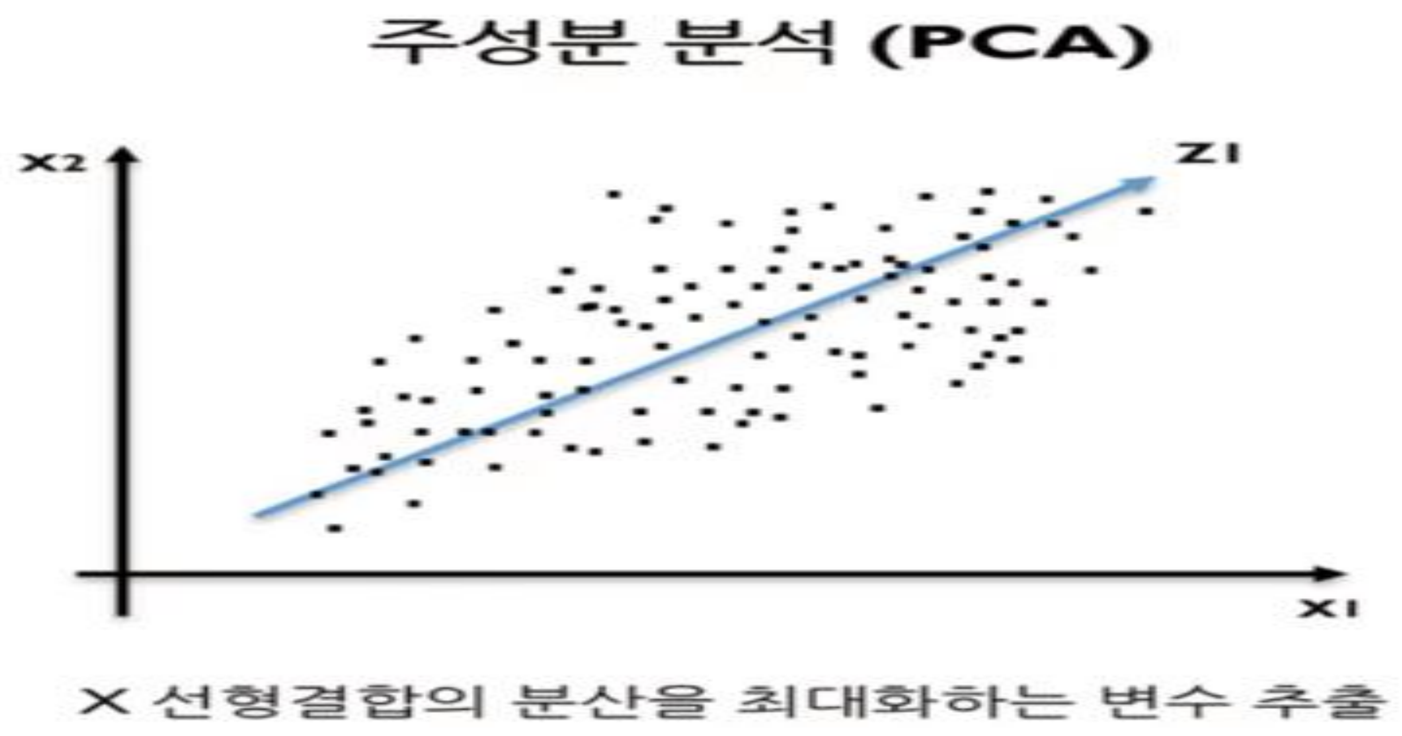
$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \Rightarrow Cov(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1}$$

$$cos\theta = \frac{x \cdot y}{\|x\| \|y\|} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = Cor(X, Y)$$

$$Max\, Cov(X, Y) \Leftrightarrow Max < x, y > \Leftrightarrow Max\, cos\theta$$

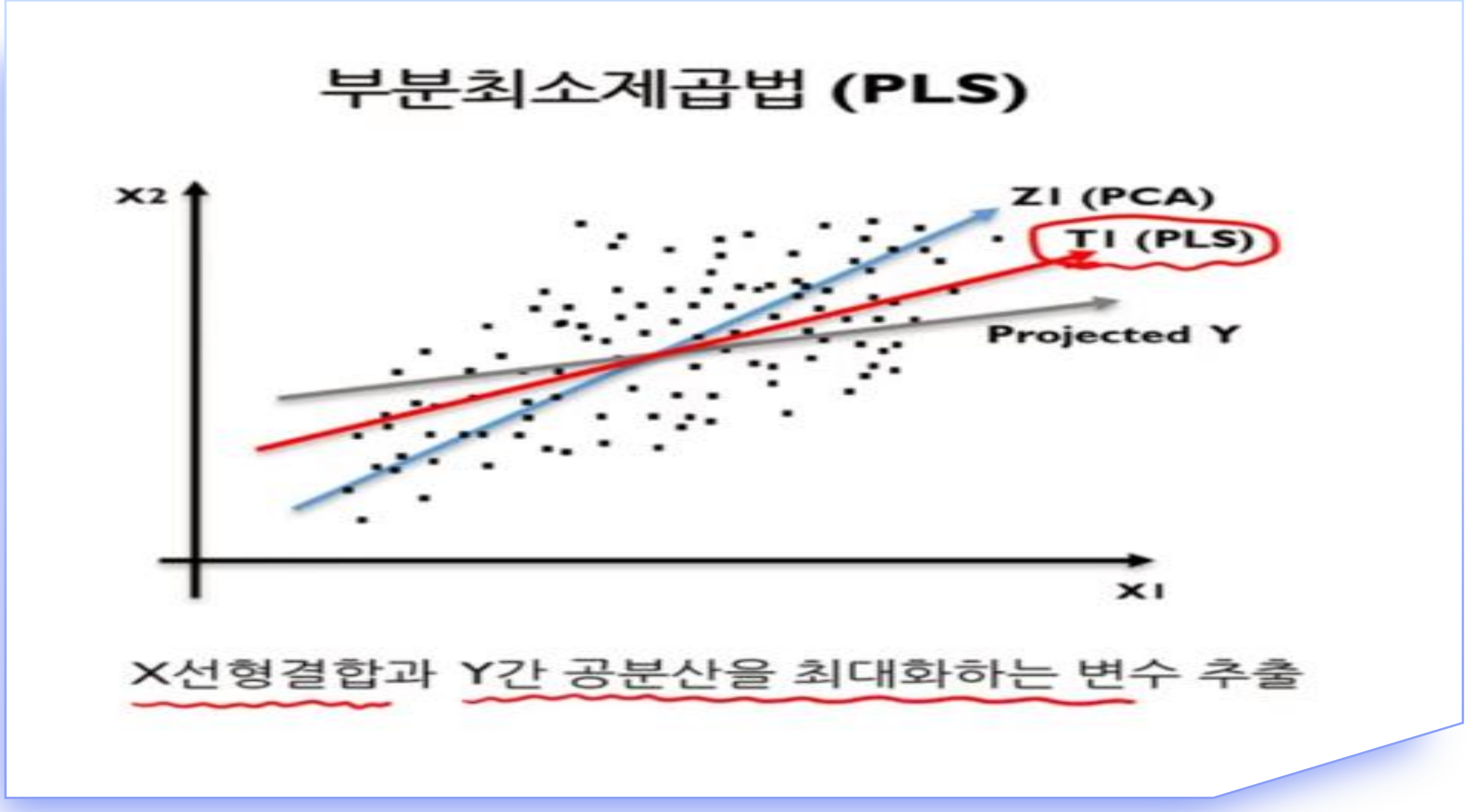
1 | 3

부분최소제곱법



1 | 3

부분최소제곱법



부분최소제곱법 (PLS)
X2
Z1 (PCA)
T1 (PLS)
Projected Y
X1
X선형결합과 Y간 공분산을 최대화하는 변수 추출

1 | 3

부분최소제곱법

새로운 변수 $T = Xw$

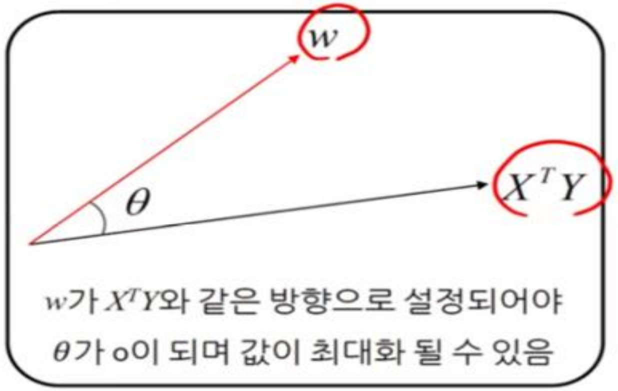
가중치 w는 어떻게 설정하는가?

➡ Maximize $Cov(T, Y) = Cov(Xw, Y)$

➡ Maximize $Xw \circ Y = <Xw, Y> = <w, X^TY>$

$$< w, X^TY > \\ = \|w\| \cdot \|X^TY\| \cos\theta$$

$$\therefore w = X^TY$$



PLS – NIPALS Algorithm

Step 1. 데이터 정규화 (mean centering)

Step 2. 첫 번째 PLS 변수 (t_1) 추출

(1) 첫 번째 X,Y 설정

$$X_1 = X, \qquad Y_1 = Y$$

(2) 공분산이 최대가 되도록 하는 선형조합 가중치 w_1 계산

$$w_1 = \frac{X_1^T Y_1}{\|X_1^T Y_1\|} \longrightarrow \|w_1\| = 1$$

(3) 가중치 w_1을 이용하여 첫 번째 PLS 변수 t_1 추출

$$t_1 = X_1 w_1$$

(4) t_1의 회귀계수 b_1을 계산

$$Y_1 = t_1 b_1 + F_1, \qquad b_1 = (t_1^T t_1)^{-1} t_1^T Y_1 \quad \text{(by 최소제곱법)}$$

Step 3. 두 번째 PLS 변수 (t_2) 추출

(1) 두 번째 X,Y 설정

※ 앞서 탐색한 t_1이 설명하지 못하는 부분만을 고려하기 위하여, t_1이 기존 X, Y에 대해서 각각 설명하는 부분을 제외함 (Extract the effect of t_1 from both X and Y)

1) 변수 t_1과 회귀계수 b_1을 사용하여, t_1이 기존 Y에 대해서 설명하는 부분을 제외

$$Y_1 = t_1 b_1 + F_1, \qquad b_1 = \left(t_1^T t_1\right)^{-1} t_1^T Y_1 \quad \text{(by 최소제곱법)}$$
$$Y_2 = F_1 = Y_1 - t_1 b_1$$

$* F_1 \rightarrow$ Y_1에 대한 잔차 (t_1이 Y_1에 대해 설명하지 못하는 부분)

2) 변수 t_1이 기존 X에 대해서 설명하는 부분을 제외

$$X_1 = t_1 p_1^T + E_1, \qquad p_1^T = \left(t_1^T t_1\right)^{-1} t_1^T X_1 \quad \text{(by 최소제곱법)}$$
$$X_2 = E_1 = X_1 - t_1 p_1^T$$

$* E_1 \rightarrow$ X_1에 대한 잔차 (t_1이 X_1 에 대해 설명하지 못하는 부분)

1 | 3

부분최소제곱법

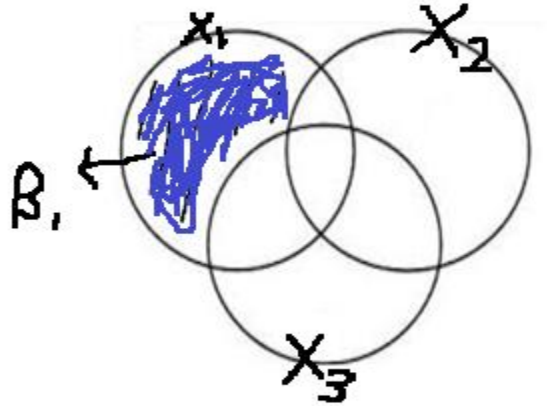
β :?

Ex) $e(x_1|1,x_2,x_3)$ is orthogonal to span$\{1,x_2,x_3\}$.

$$\text{span}\{1,x_1,x_2,x_3\} = \text{span}\{1,x_2,x_3\} \bigoplus \text{span}\{e(x_1|1,x_2,x_3)\}$$

$$\Rightarrow \hat{y}(1,x_1,x_2,x_3) = \hat{y}(1,x_2,x_3) + \hat{y}(e(x_1|1,x_2,x_3)).$$

$$= \hat{y}(1,x_2,x_3) + \hat{\beta}_1 e(x_1|1,x_2,x_3)$$



1 | 3

부분최소제곱법

Step 3 (Continue). 두 번째 PLS 변수 (t_2) 추출

(2) 공분산이 최대가 되도록 하는 선형조합 가중치 w_2 계산

$$w_2 = \frac{X_2^T Y_2}{\|X_2^T Y_2\|} \rightarrow \|w_2\| = 1$$

(3) 가중치 w_2를 이용하여 두 번째 PLS 변수 t_2 추출

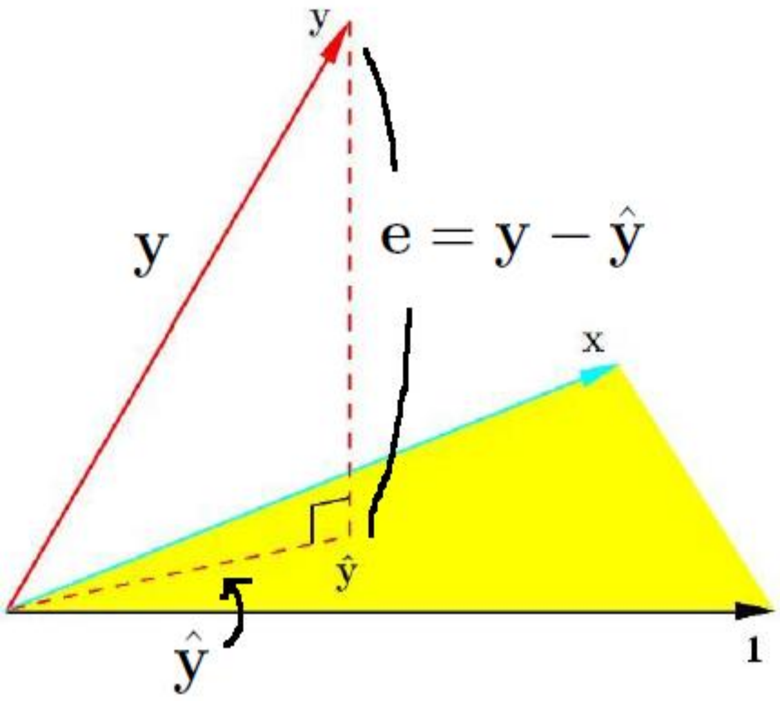
$$t_2 = X_2 w_2$$

(4) t_2의 회귀계수 b_2를 계산

$$Y_2 = t_2 b_2 + F_2, \qquad b_2 = (t_2^T t_2)^{-1} t_2^T Y_2$$

2 | 3

조건부평균



Projection : written by new basis

y -> $sp\{1,x\}$

=> $\hat{y} = b_0 + b_1 x$

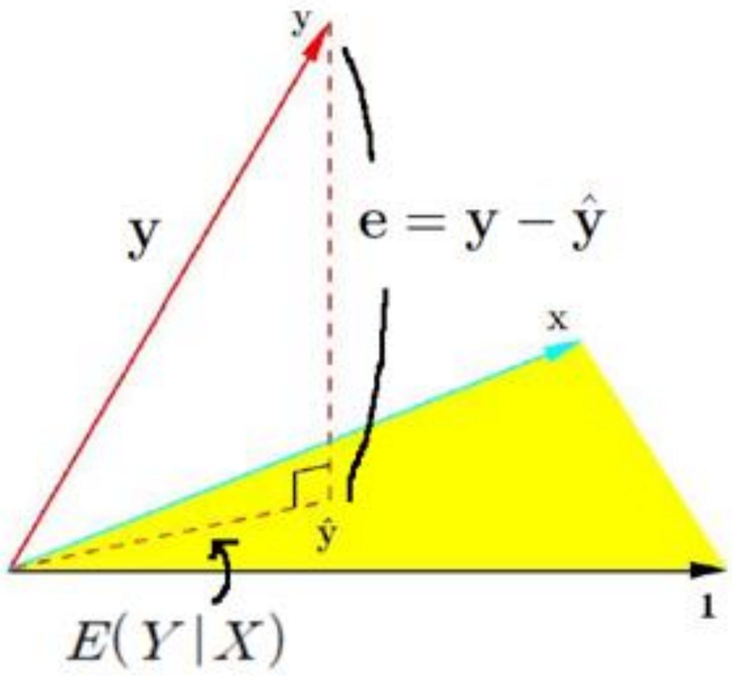
$= E(Y|X)$

∴ Projection = 조건부 평균

$E(E(Y|X)|X) = E(Y|X)$

2 | 3

조건부평균



Projection : Minimizes $\|e\|^2$

➡ Best Approximation!!

1. 평균이 같으면 좋겠다!

$$E(Y) = E(E(Y|X)) = E(\hat{Y})$$

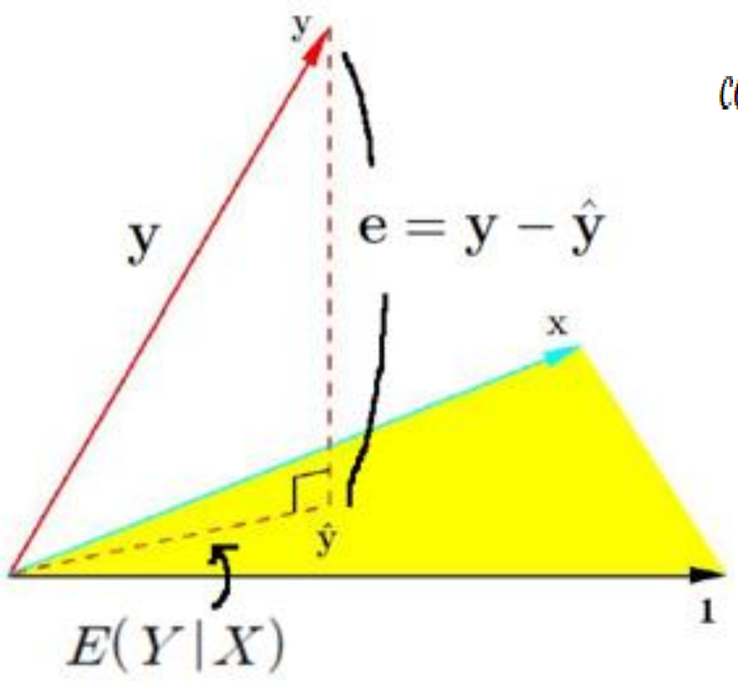
2. 분산이 작았으면 좋겠다!

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

$$Var(Y) \geq Var(E(Y|X))$$

2 | 3

조건부평균



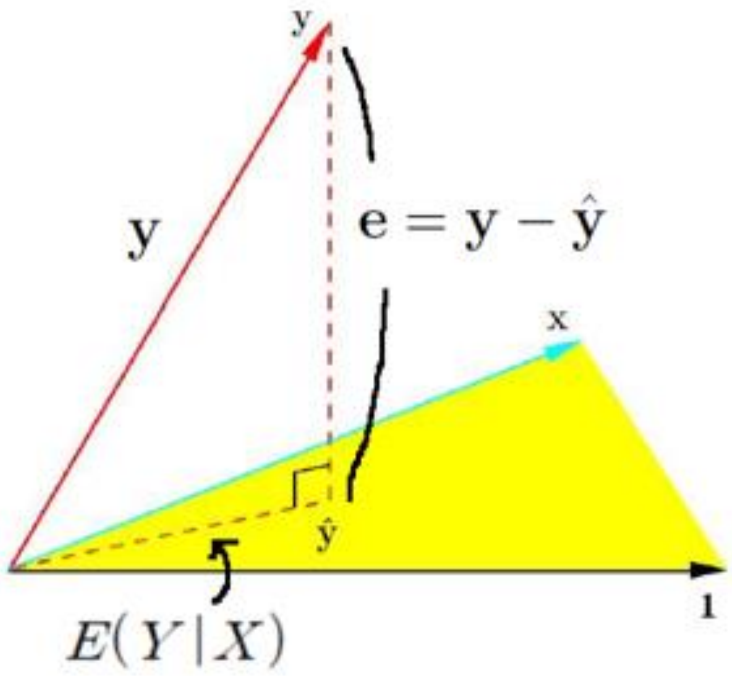
$$\cos\theta = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = Cor(X,Y)$$

$$||x|| = \sqrt{\sum(x_i - \bar{x})^2}$$

$$||x|| \propto \sum(x_i - \bar{x})^2 \Rightarrow ||x|| \propto Var(x)$$

2 | 3

조건부평균



Projection : Minimizes $\|e\|^2$

➡ Best Approximation!!

1. 평균이 같으면 좋겠다!

$$E(Y) = E(E(Y \mid X)) = E(\hat{Y})$$

2. 분산이 작았으면 좋겠다!

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

$$Var(Y) \geq Var(E(Y|X))$$

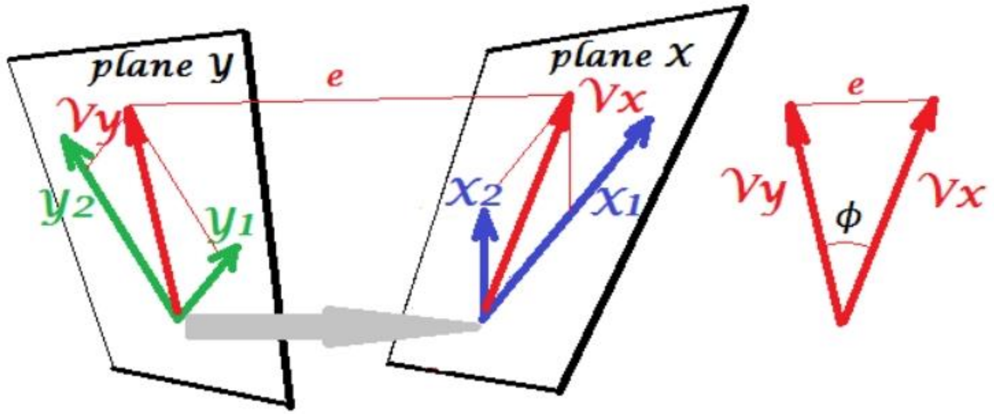
$$\|Y\|^2 \geq \|\hat{Y}\|^2$$

3 | 3

정준상관분석

정준상관분석 (Canonical Correlation Analysis)

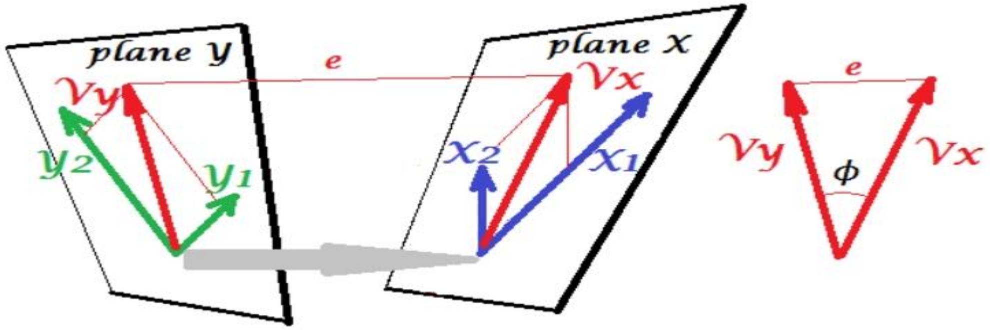
: 두 변수 집단간의 관계를 저차원의 정준변수를 통하여 관계를 설명



- (단순)상관관계 : (변수 1개, 변수 1개) 에 대한 상관성
- 다중상관관계 : (변수 여러 개, 변수 1개)에 대한 상관성
- 정준상관관계 : (변수 여러 개, 변수 여러 개)에 대한 상관성

3 | 3

정준상관분석



$$v_x = a_1 x_1 + a_2 x_2 = a^T X$$

$$v_y = b_1 y_1 + b_2 y_2 = b^T Y$$

$$\rho(a, b) = cor(v_x, v_y) = cor(a^T X, b^T Y) = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}}$$

$$K = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$$

$$K = U \Lambda V^T$$

$$a_i = \Sigma_{XX}^{-1/2} u_i$$

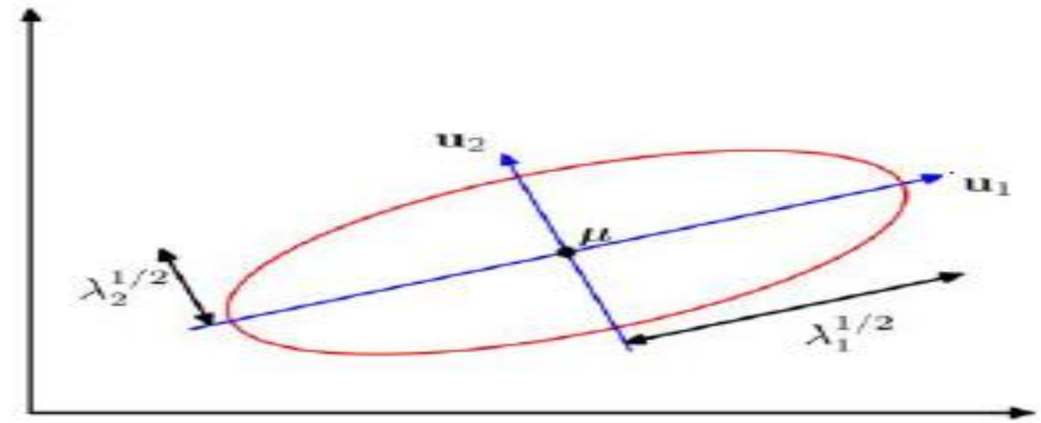
$$b_i = \Sigma_{YY}^{-1/2} v_i$$

3 | 3

정준상관분석

$$\rho(a,b) = cor(v_x, v_y) = cor(a^T X, b^T Y) = \frac{a^T \Sigma_{XY}\, b}{\sqrt{a^T \Sigma_{XX}\, a}\sqrt{b^T \Sigma_{YY}\, b}}$$

$$K = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2} \qquad K = U \Lambda V^T \quad \Rightarrow \quad a^T K K^T a \qquad b^T K^T K b \quad \Rightarrow \quad a_i = \Sigma_{XX}^{-1/2} u_i \qquad b_i = \Sigma_{YY}^{-1/2} v_i$$



$$Cov(a_i X, a_j X) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$Cov(b_i Y, b_j Y) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$Cov(a_i X, b_j Y) = \begin{cases} \lambda_i & i = j \\ 0 & i \neq j \end{cases}$$

THANK
YOU