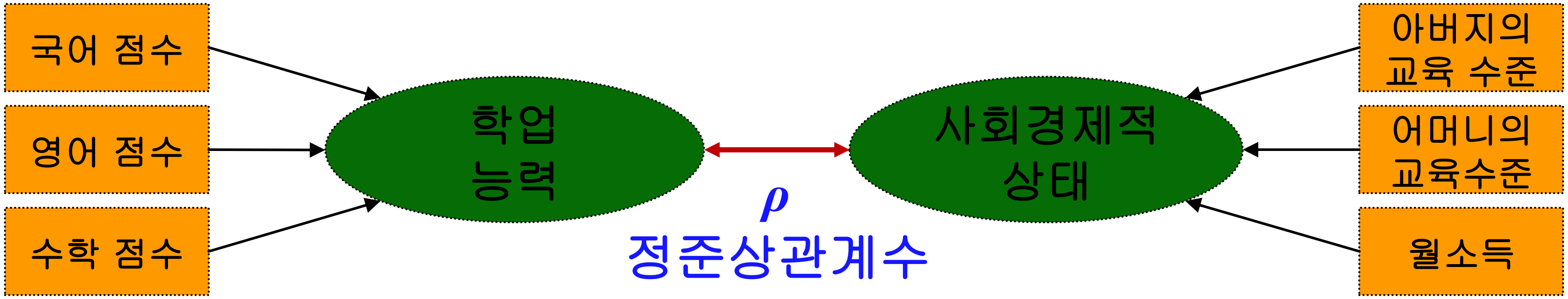
제 4 장
정준상관분석
Canonical Correlation Analysis

4.1 정준상관분석의 개념



$$\begin{cases} v &= \boldsymbol{\alpha}'\mathbf{x} &= \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p \\ w &= \boldsymbol{\beta}'\mathbf{y} &= \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q \end{cases}$$

$$\rho = \max_{\boldsymbol{\alpha},\ \boldsymbol{\beta}} \mathrm{Corr}(v, w)$$

- 정준변량 (Canonical Variate)
- 정준계수 (Canonical Coefficient)

4.1 정준상관분석의 개념

- 두 개의 변수 집단 간의 선형성 상관 관계를 파악하고 양으로 표현하고자 할 때
- Hotelling(1935)에 의해 제안된 방법.

-(수학계산속도와 계산능력), (독해속도와 독해능력) 두 개 변수집단간의 상관관계 계산

- ▶ **단순상관계수 : (한 개 변수, 한 개 변수)에 대한 상관성**
- ▶ **다중상관계수 : (한 개 변수, 여러 개 변수)에 대한 상관성**
- ▶ **정준상관계수 : (여러 개 변수, 여러 개 변수) 에 대한 상관성**

다차원에 놓인 두 변수 집단간의 관계를 저차원의 정준변수 쌍으로 전환하여 관계를 설명할 수 있음. 정준상관계수가 정준변수간의 상관성을 나타냄.

4.2 Example: 업무특성과 만족도 데이터

job	업무특성			만족도		
	x_1	x_2	x_3	y_1	y_2	y_3
A	10	11	70	72	26	9
B	85	22	93	63	76	7
C	83	63	73	96	31	7
D	82	75	97	96	98	6
E	36	77	97	84	94	6
F	28	24	75	66	10	5
G	64	23	75	31	40	9
H	19	15	50	45	14	2
I	33	13	70	42	18	6
J	23	14	90	79	74	4
K	37	13	70	39	12	2
L	23	74	53	54	35	3

- x_1(다양성): 담당하는 업무에 있어서 다양성의 정도(%),
- x_2(피드백): 업무수행에 필요한 피드백(상사 및 동료들의 반응)의 정도(%),
- x_3(자율성): 업무수행에 필요한 자율성의 정도(%),
- y_1(경력): 현재의 업무가 미래의 경력에 도움이 되는지에 대한 만족도(%),
- y_2(관계): 상사 및 관리자와의 관계에 대한 만족도(%),
- y_3(보수): 급여 등 보수적인 측면의 만족도(1~10).

R Example

R에서 정준상관분석: cancor() 함수 이용.
* CCA 패키지를 사용할 경우 : cc() 함수 이용.
* yacca 패키지를 사용하는 경우 : cca() 함수 이용.

```
> job<-read.table("job.txt",header=T) #데이터 읽기
> library(CCA)
> head(job)
  job x1 x2 x3 y1 y2 y3
1   A 10 11 70 72 26  9
2   B 85 22 93 63 76  7
3   C 83 63 73 96 31  7
4   D 82 75 97 96 98  6

> dim(job)
[1] 14  7
```

R Example

```
> x<- job[,2:4]  # 업무 특성
> y<- job[,5:7]  # 만족도
> matcor(x,y)
$Xcor
          x1        x2        x3
x1 1.0000000 0.2655074 0.5362404
x2 0.2655074 1.0000000 0.1178639
x3 0.5362404 0.1178639 1.0000000

$Ycor
          y1        y2        y3
y1 1.0000000 0.5538806 0.2244930
y2 0.5538806 1.0000000 0.2108256
y3 0.2244930 0.2108256 1.0000000

$XYcor
          x1          x2        x3        y1        y2          y3
x1 1.0000000  0.26550743 0.5362404 0.3060371 0.4241932  0.39608438
x2 0.2655074  1.00000000 0.1178639 0.5247464 0.5262038 -0.01305114
x3 0.5362404  0.11786387 1.0000000 0.5162016 0.7655578  0.37519987
y1 0.3060371  0.52474641 0.5162016 1.0000000 0.5538806  0.22449295
y2 0.4241932  0.52620376 0.7655578 0.5538806 1.0000000  0.21082561
y3 0.3960844 -0.01305114 0.3751999 0.2244930 0.2108256  1.00000000
```

4.3 정준상관분석의 대수적 의미

$$\begin{cases} v & = \boldsymbol{\alpha}'\mathbf{x} & = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p \\ w & = \boldsymbol{\beta}'\mathbf{y} & = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q \end{cases}$$

- **두 확률변수** $\mathbf{x} = (x_1, x_2, \dots, x_p)', \mathbf{y} = (y_1, y_2, \dots, y_q)'$
- 분산을 다음과 같이 나타내자.

$$Var(\mathbf{x}) = \Sigma_{xx}, \qquad Var(\mathbf{y}) = \Sigma_{yy}, \qquad Cov(\mathbf{x}, \mathbf{y}) = \Sigma_{xy}$$

- **선형 결합의 분산**

$$Var(\alpha'\mathbf{x}) = \alpha'\Sigma_{xx}\alpha, \qquad Var(\beta'\mathbf{y}) = \beta'\Sigma_{yy}\beta,$$

$$Cov(\alpha'\mathbf{x}, \beta'\mathbf{y}) = \alpha'\Sigma_{xy}\beta$$

4.3 정준상관분석의 대수적 의미

$$\begin{cases} v & = \boldsymbol{\alpha}'\mathbf{x} & = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p \\ w & = \boldsymbol{\beta}'\mathbf{y} & = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q \end{cases}$$

$$\rho(v, w) = \frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{xy}\boldsymbol{\beta}}{\sqrt{(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha})(\boldsymbol{\beta}'\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta})}}$$

- $p = q = 1$일 경우 → 단순상관계수 (simple correlation coefficient)

$$v_1 = \alpha_1 x_1,\ w_1 = \beta_1 y_1 \qquad \begin{aligned} \rho_1 &= \max |\mathrm{Corr}(v_1, w_1)| \\ &= \max |\mathrm{Corr}(\alpha_1 x_1,\ \beta_1 y_1)| \\ &= |\mathrm{Corr}(x_1,\ y_1)| \end{aligned}$$

- $p = 1,\ q \geq 2$일 경우 → 다중상관계수 (multiple correlation coefficient)

$$v_1 = \alpha_1 x_1,\ w_1 = \boldsymbol{\beta}\mathbf{y} \qquad \begin{aligned} \rho_1 &= \max |\mathrm{Corr}(v_1, w_1)| \\ &= \max |\mathrm{Corr}(\alpha_1 x_1,\ \boldsymbol{\beta}\mathbf{y})| \\ &= \max |\mathrm{Corr}(x_1,\ \boldsymbol{\beta}\mathbf{y})| \end{aligned}$$

4.3.1 정준상관계수와 정준변수

$$\rho(v,w) = \frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{xy}\boldsymbol{\beta}}{\sqrt{(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha})(\boldsymbol{\beta}'\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta})}}$$

위의 상관계수를 최대로 하는 정준계수벡터 α 와 β 를 찾는다.

▶ 정준변수를 구하는 과정

1. 첫 번째 정준변수 쌍(first canonical variate pair) $(\boldsymbol{v_1}, \boldsymbol{w_1})$은 $|\boldsymbol{\rho}(\boldsymbol{v}, \boldsymbol{w})|$를 최대로 하며 $\boldsymbol{Var}(\boldsymbol{v_1}) = \boldsymbol{Var}(\boldsymbol{w_1})$=1 인 변수들의 선형결합식이다.

2. 두 번째 정준변수 쌍(second canonical variate pair) $(\boldsymbol{v_2}, \boldsymbol{w_2})$는 $(\boldsymbol{v_1}, \boldsymbol{w_1})$과 독립이면서 $|\boldsymbol{\rho}(\boldsymbol{v}, \boldsymbol{w})|$를 최대로 하며 $\boldsymbol{Var}(\boldsymbol{v_2}) = \boldsymbol{Var}(\boldsymbol{w_2})$=1 인 변수들의 선형결합식이다.

3. 위와 같은 과정을 반복한다. (정준변수의 개수 = s = min(p,q))

4.3.1 정준상관계수와 정준변수

▶ 정준계수벡터를 구하는 과정

$|\boldsymbol{\rho}(\boldsymbol{v},\boldsymbol{w})|$를 최대로 하는 정준계수벡터 α_1 과 β_1 은 다음과 같은 두 행렬의 첫 고유벡터가 된다.

$$\boldsymbol{A} = \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \quad \boldsymbol{B} = \Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

두 행렬 $\boldsymbol{A}$와 $\boldsymbol{B}$의 첫 고유값은 둘 다 같다. 이 고유값을 ρ_1^2 라고 하면

$$\boldsymbol{Cor}(\boldsymbol{v}_1 = \alpha_1'\mathbf{x}, \boldsymbol{w}_1 = \beta_1'\mathbf{y}) = \rho_1$$

$|\boldsymbol{\rho}(\boldsymbol{v},\boldsymbol{w})|$를 최대로 하는 정준계수벡터 α_2 과 β_2는 $\boldsymbol{A}$와 $\boldsymbol{B}$ 의 두 번째 고유벡터.

$$\boldsymbol{Cor}(\boldsymbol{v}_2 = \alpha_2'\mathbf{x}, \boldsymbol{w}_2 = \beta_2'\mathbf{y}) = \rho_2$$

4.3.2 표준화된 정준계수

정준계수벡터 $\boldsymbol{a}_k$에 대해, $Var(X_i) = \sigma_{ii},\ \ i = 1,2,\ldots,p$

$$\begin{aligned}\boldsymbol{a}_k{}'(\boldsymbol{X}-\boldsymbol{\mu}) &= a_{k1}(X_1-\mu_1)+a_{k2}(X_2-\mu_2)+\cdots+a_{kp}(X_p-\mu_p)\\ &= a_{k1}\sqrt{\sigma_{11}}\frac{(X_1-\mu_1)}{\sqrt{\sigma_{11}}}+\cdots+a_{kp}\sqrt{\sigma_{pp}}\frac{(X_p-\mu_p)}{\sqrt{\sigma_{pp}}}\\ &= c_{k1}Z_1+c_{k2}Z_2+\cdots+c_{kp}Z_p\end{aligned}$$

- 표준화 변수의 정준상관변수 계수는 원래 변수 X_i로 구한 것에 sd(X_i) 를 곱한 형태.
- 정준변수는 인공적으로 만들어 낸 변수이므로 인자나 주성분과 같은 절대적 의미를 부여하기는 힘들며 관심있는 변수집단에 대해 연관성을 알고자 할 때 주로 이용할 수 있다.
- 변수를 표준화 하더라도 정준상관계수는 변하지 않으므로 단위의 표준화와 해석을 위해서는 표준화 변수들에 대한 정준상관분석을 권장한다.

4.3.2 표준화된 정준계수 Example in R

- 타이어 광고효과에 관한 조사로부터 얻은 데이터. (n=252, p=2, q=3)

- **X1:** **TV** 상업광고에서 주장하는 내용에 대해 신뢰하는 정도
- **X2:** 신상품에 대한 사후 관심 정도
- **Y1:** 해당 상품에 대한 관심 정도
- **Y2:** 지난 번에 구매한 타이어의 상표에 대한 관심 정도
- **Y3:** 신상품에 대한 사전(광고 접하기 전) 관심 정도

4.3.2 표준화된 정준계수 (타이어 Example in R)

- 타이어 광고효과에 관한 조사로부터 얻은 데이터. (n=252, p=2, q=3)
- 표준화 됨. (advert.RData)

```
> load("advert.RData")
> cor(XY)
            x1        x2         y1          y2        y3
x1  1.0000000 0.6107893 0.16769710 -0.02274270 0.2832269
x2  0.6107893 1.0000000 0.31998459  0.16247188 0.5452918
y1  0.1676971 0.3199846 1.00000000  0.08601442 0.1360072
y2 -0.0227427 0.1624719 0.08601442  1.00000000 0.1707519
y3  0.2832269 0.5452918 0.13600722  0.17075187 1.0000000
```

- x 와 y 의 상관관계는 0.545 가 가장 크고 나머지는 대부분 작음.

```
> result<-cc(XY[,1:2],XY[,3:5])
```

4.3.2 표준화된 정준계수 Example in R

```
> result$cor     # ρ(v,w) 값들.
[1] 0.6073073 0.1359451
```

- 첫번째 정준변수가 두 집단의 연관관계를 대부분 설명하며 두번째 정준변수는 별로 설명하지 못하고 있음.

```
> result$xcoef   # α1, α2값들.
         [,1]       [,2]
x1  0.1875053  1.3192867
x2 -1.1240528 -0.6374495
```

$\hat{\alpha}_1$ $\hat{\alpha}_2$

$$\begin{cases} v &= \boldsymbol{\alpha}'\mathbf{x} &= \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p \\ w &= \boldsymbol{\beta}'\mathbf{y} &= \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q \end{cases}$$

```
> result$ycoef   # β1, β2 값들.
         [,1]       [,2]
y1 -0.4234695  0.1325906
y2 -0.1178380 -0.9562355
y3 -0.8017381  0.2531249
```

$\hat{\beta}_2$

$\hat{\beta}_1$

4.3.3 정준점수 (Canonical Score)

- 정준변수: 두 변수 집단간의 연관관계에 의해 요약된 인공변수 (v_1, w_1) 또는 (v_2, w_2). 전체 변수의 차원
- 정준점수: 각 개체(sample)의 특징 또는 프로파일을 정준변수의 차원에서 고려했을 때 얻어지는 점수. 각 샘플의 차원.

$$\hat{v}_{i1} = \hat{\alpha}_1 x_{i1} + \hat{\alpha}_2 x_{i2} \quad \text{그리고} \quad \hat{w}_{i1} = \hat{\beta}_1 y_{i1} + \hat{\beta}_2 y_{i2} + \hat{\beta}_3 y_{i3}$$

- 정준점수는 주성분점수 또는 인자점수와 비슷한 개념
- 후속적인 통계분석에서 어떤 개념상의 의미를 부여할 수 있는 차원 축약된 새로운 변수 혹은 지표로 이용될 수 있음.

```
> ls(result$score) # score에 저장되어 있는 6개의 결과들.
[1] "corr.X.xscores" "corr.X.yscores" "corr.Y.xscores" "corr.Y.yscores"
"xscores"        "yscores"
> ? cc # 여기서 score 부분을 보면 comput() 가 쓰였음을 알 수 있다.
> ? comput #  score를 구성하는 6개의 output에 대한 설명을 보여줌.
```

4.3.3 정준점수 (Canonical Score) 타이어 Example

- 정준점수: 각 개체(sample)의 특징 또는 프로파일을 정준변수의 차원에서 고려했을 때 얻어지는 점수. 각 샘플의 차원.

$$\hat{v}_{i1} = \hat{\alpha}_{11}x_{i1} + \hat{\alpha}_{12}x_{i2} \qquad \hat{w}_{i1} = \hat{\beta}_{11}y_{i1} + \hat{\beta}_{12}y_{i2} + \hat{\beta}_{13}y_{i3}$$

$$\hat{v}_{i2} = \hat{\alpha}_{21}x_{i1} + \hat{\alpha}_{22}x_{i2} \qquad \hat{w}_{i2} = \hat{\beta}_{21}y_{i1} + \hat{\beta}_{22}y_{i2} + \hat{\beta}_{23}y_{i3}$$

```
> head(result$score$xscores)
            [,1]       [,2]
[1,]  0.056470264 -1.4524366
[2,]  0.006835419  0.4620357
[3,]  1.377336750  0.8452755
[4,]  1.908960705  0.4521501
```

$\hat{v}_{11}$	$\hat{v}_{12}$
$\hat{v}_{21}$	$\hat{v}_{22}$
$\hat{v}_{31}$	$\hat{v}_{32}$
$\hat{v}_{41}$	$\hat{v}_{42}$

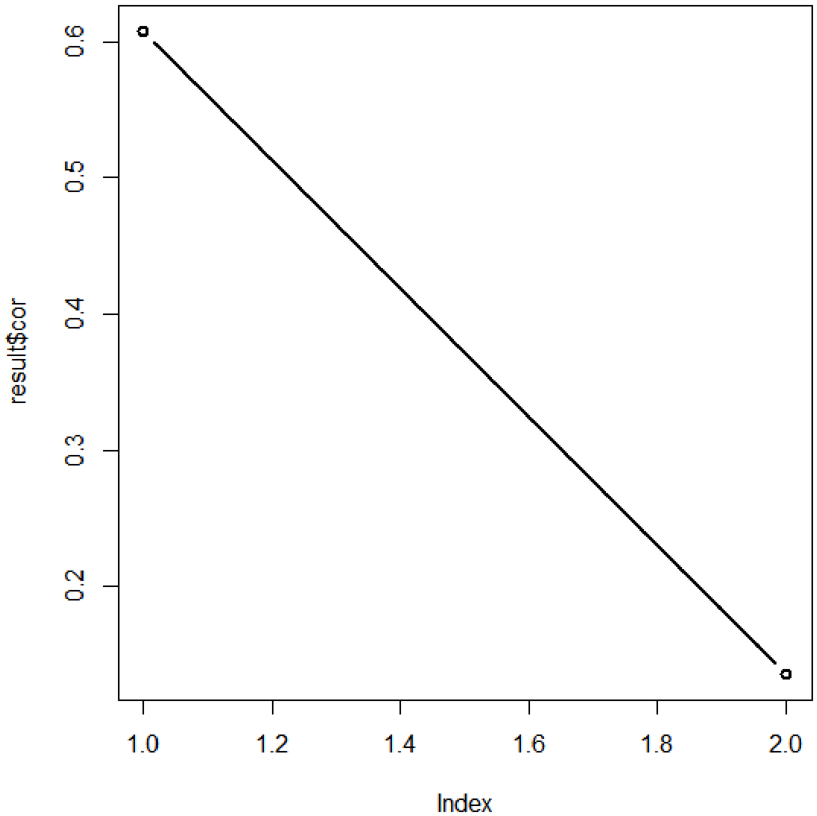
```
> head(result$score$yscores)
          [,1]       [,2]
[1,]  0.9389963 -1.9115303
[2,]  0.4017414 -0.1099658
[3,]  0.1433949  0.8901011
[4,]  0.6286915 -0.1328721
```

$\hat{w}_{11}$	$\hat{w}_{12}$
$\hat{w}_{21}$	$\hat{w}_{22}$
$\hat{w}_{31}$	$\hat{w}_{32}$
$\hat{w}_{41}$	$\hat{w}_{42}$

4.3.3 정준점수 (Canonical Score) Example

- 알맞은 정준변수의 개수 선택

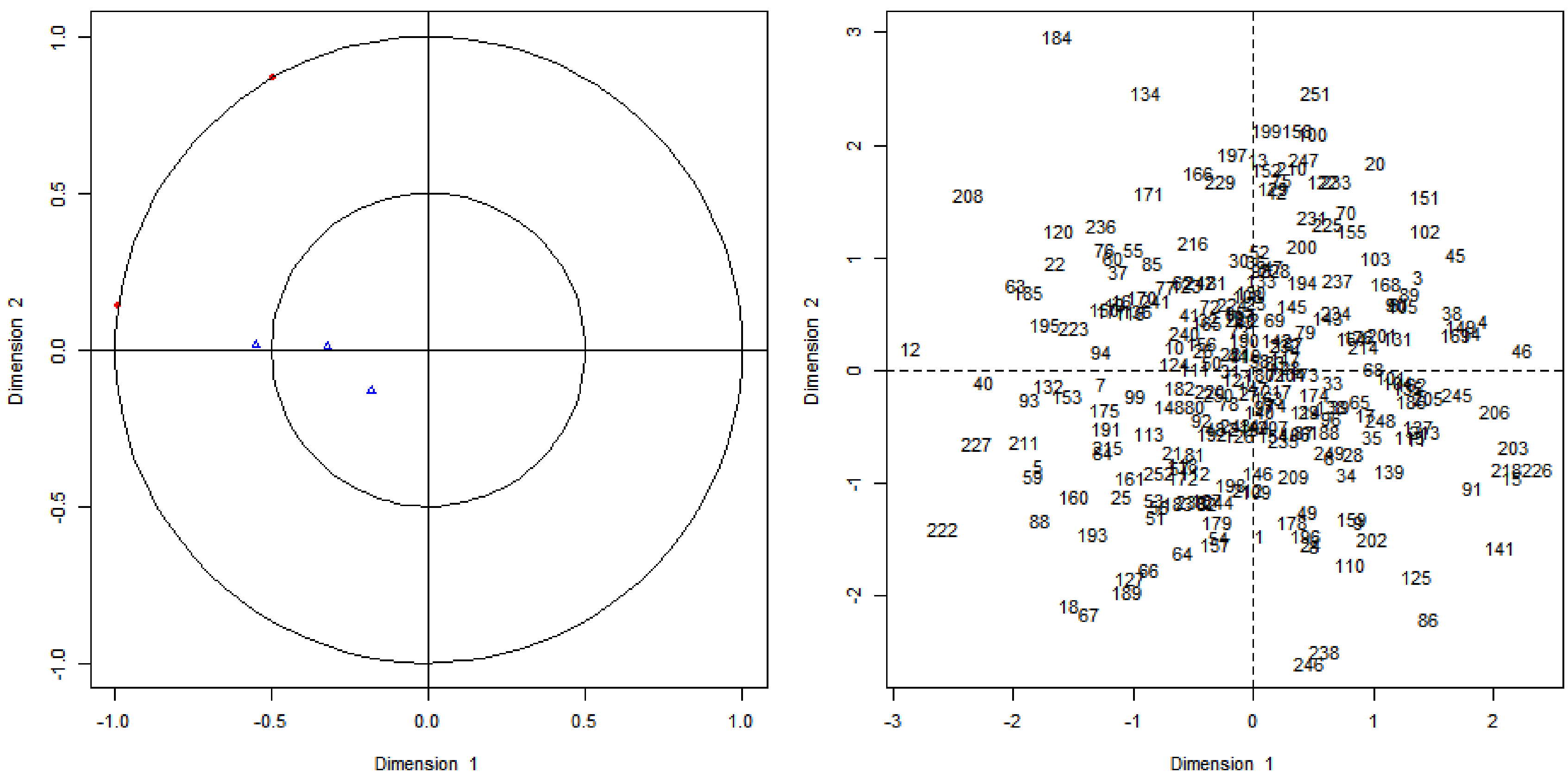
```
> plot(result$cor,type='b') # Scree plot 과 비슷하게 이용. 여기서 1개 선택.
```



4.3.3 그래프 이용 Example in R

- 정준변수와 정준점수를 그래프를 통하여 나타내기.

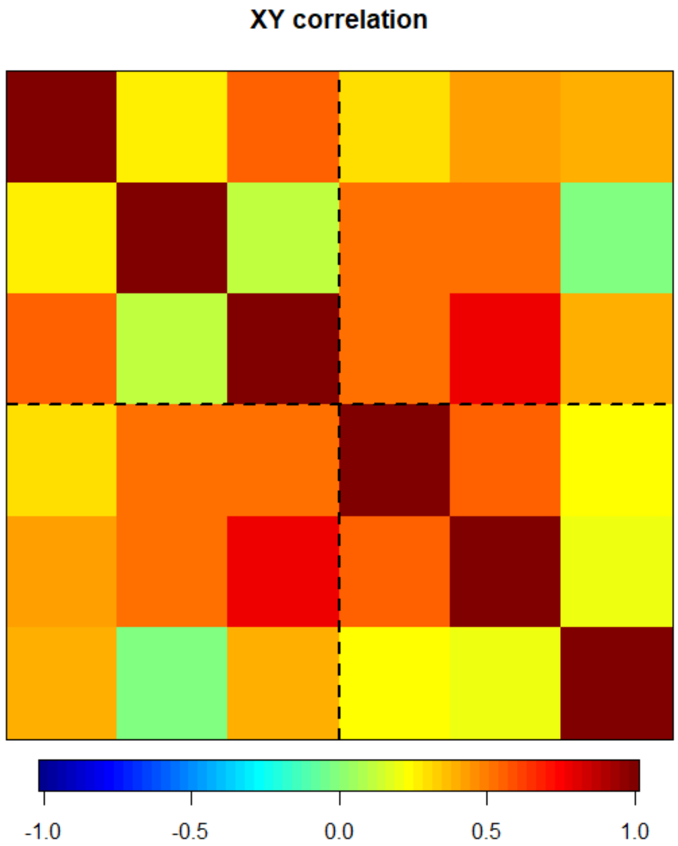
```
> plt.cc(result)
```



4.3.3 그래프 이용 Example in R

- 두 행렬간의 상관성 이미지 그림 1.

```
> aaa<- matcor(x,y); img.matcor(aaa,type=1)
```



4.3.3 그래프 이용 Example in R

- 두 행렬간의 상관성 이미지 그림 2.

```
> aaa<- matcor(x,y); img.matcor(aaa,type=2)
```