9장 정준상관분석

덕성여자대학교 정보통계학과 김 재희

9.1 서론

정준상관분석 (canonical correlation analysis) :

- 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 있으며 정준상관계수가 정준변수간의 상관성을 나타낸다.

▶ 정준변수와 정준상관계수를 구하는 단계

(1) 가장 큰 상관계수를 갖는 한 쌍의 선형결합식을 결정한다.

(2) 첫 번째로 선택된 한 쌍의 선형결합식과는 독립이면서 그 다음으로 큰 상관계수를 갖는 선형결합식을 찾는다.

(3) 이와 같은 방법으로 먼저 찾은 선형결합식들과는 독립이면서 그 다음으로 큰 상관계수를 갖는 선형결합식을 찾는다.

▶ 이렇게 찾은 변수들의 선형결합식: 정준변수(canonical variables)

▶ 정준변수들의 상관계수를 정준상관계수(canonical correlation).

정준상관계수는 두 변수 집단 간의 연관성 정도를 나타낸다.

9.2. 정준변수와 정준상관

9.2.1 정준변수와 정준상관계수의 정의 및 개념

▶ 다중상관계수를 구하는 방법.

(한 개 변수, 여러 개 변수)에 대해 두 변수 집단간의 최대 상관성

$X_1,...,X_p$와 Y_1 의 분산-공분산행렬, 상관행렬로 나타내면 ($Y_1 = Y$로 놓자.)

$$\boldsymbol{S}=\begin{pmatrix}\boldsymbol{S}_{XX} & \boldsymbol{s}_{XY}\\ \boldsymbol{s}_{YX} & s_{YY}\end{pmatrix} \qquad \boldsymbol{R}=\begin{pmatrix}\boldsymbol{R}_{XX} & \boldsymbol{r}_{XY}\\ \boldsymbol{r}_{YX}' & 1\end{pmatrix}$$

여기서 $\boldsymbol{S}_{XY}' = (s_{1Y}, s_{2Y}, ..., s_{pY})$는 X_i와 Y_1의 공분산벡터 $\quad i = 1, \dots, p$

$\boldsymbol{r}_{XY}' = (r_{1Y}, r_{2Y}, ..., r_{pY})$는 X_i와 Y_1의 상관계수벡터.

다중상관계수제곱 :

$$R^2 = \frac{\boldsymbol{s}_{XY}'\boldsymbol{S}_{XX}^{-1}\boldsymbol{s}_{XY}}{s_Y^2} = \boldsymbol{r}_{XY}'\boldsymbol{R}_{XX}^{-1}\boldsymbol{r}_{XY}$$

▶ (여러 개 변수, 여러 개 변수)에 대해 두 변수 집단간의 상관성을 나타내는 방법

각 개체에 대해 두 개의 변수 집단

$$\boldsymbol{X}=(X_1,...,X_p)' \text{ 와 } \quad \boldsymbol{Y}=(Y_1,...,Y_q)', \quad (p \le q)$$

이 관측되었다고 하자.

두 변수 집단으로 구성된 $(p+q)\times 1$ 확률벡터 $\boldsymbol{W}$ 는

$$\boldsymbol{W}=\begin{pmatrix} \boldsymbol{X} \\ --- \\ \boldsymbol{Y} \end{pmatrix}=\begin{pmatrix} X_1 \\ \vdots \\ X_p \\ --- \\ Y_1 \\ \vdots \\ Y_q \end{pmatrix}$$

로 표현되며

W는 다음의 모평균벡터와 모공분산행렬

$$E(W) = \mu_{(p+q)\times 1} = \begin{pmatrix} E(X) \\ --- \\ E(Y) \end{pmatrix} = \begin{pmatrix} \mu_X \\ --- \\ \mu_Y \end{pmatrix},$$

$$\begin{aligned} \Sigma_{(p+q)\times(p+q)} &= E(W-\mu)(W-\mu)' \\ &= \begin{pmatrix} E(X-\mu_X)(X-\mu_X)' & | & E(X-\mu_X)(Y-\mu_Y)' \\ ---------- & | & ---------- \\ E(Y-\mu_Y)(X-\mu_X)' & | & E(Y-\mu_Y)(Y-\mu_Y)' \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \end{aligned}$$

을 가진 모집단으로부터의 확률벡터로 생각할 수 있다.

* 상수계수벡터로서 $p\times 1$ 벡터 a와 $q\times 1$ 벡터 b에 대해 변수들의 선형결합식

$$U=a'\boldsymbol{X}$$
$$V=b'\boldsymbol{Y}$$

는 일변량 확률변수가 된다.

* U와 V의 분산 :

$$Var(U)=a'Cov(\boldsymbol{X})a=a'\Sigma_{XX}a$$
$$Var(V)=b'Cov(\boldsymbol{Y})b=b'\Sigma_{YY}b$$

* U와 V의 상관계수:

$$Corr(U,V)=\frac{a'\Sigma_{XY}b}{\sqrt{a'\Sigma_{XX}a}\ \sqrt{b'\Sigma_{YY}b}} \qquad (9.10)$$

* 식(9.10)의 상관계수를 최대로 하는 상수 벡터 $a,\ b$를 찾고자 한다.

▶ 정준변수를 구하는 과정

1. 첫 번째 정준변수 쌍(first canonical variate pair) (U_1, V_1)은 $Corr(U, V)$를 최대로 하며 $Var(U_1) = Var(V_1) = 1$인 변수들의 선형결합식이다.

2. 두 번째 정준변수 쌍(second canonical variate pair) (U_2, V_2)는 (U_1, V_1)과 서로 독립이면서 $Corr(U, V)$를 최대로 하며 $Var(U_2) = Var(V_2) = 1$인 변수들의 선형결합식이다.

   k번 째 정 준 변 수 쌍 (U_k, V_k)는 $(U_i, V_i),\ i = 1, ..., k-1$과 서 로 독 립 이 면 서 $Corr(U, V)$를 최대로 하며 $Var(U_k) = Var(V_k) = 1$인 변수들의 선형결합식이다.

정리 9.1 $p \le q$ 라고 할 때 $p \times 1$ 확률벡터 $\boldsymbol{X}$와 $q \times 1$ 확률벡터 $\boldsymbol{Y}$가 공분산행렬 $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{XX}}, Cov(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{YY}}, Cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{XY}}$ 가질 때, 상수계수벡터로서 $p \times 1$ 벡터 $\boldsymbol{a}$와 $q \times 1$ 벡터 $\boldsymbol{b}$와의 선형결합식

$$U = \boldsymbol{a}' \boldsymbol{X}$$
$$V = \boldsymbol{b}' \boldsymbol{Y}$$

에 대해 최대 상관계수

$$\max_{\mathrm{a,b}} \mathrm{Corr}(\mathrm{U}, \mathrm{V}) = \rho_1^*$$

를 갖는 첫 번째 정준변수는

$$U_1 = \boldsymbol{e}_1{}' \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1/2} \boldsymbol{X} \text{와 } V_1 = \boldsymbol{f}_1{}' \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1/2} \boldsymbol{Y}$$

로 주어진다. 또한 k번째 정준변수는 $k = 2, 3, \ldots, p$일 때

$$U_k = \boldsymbol{e}_k{}' \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1/2} \boldsymbol{X} \text{ 와 } V_k = \boldsymbol{f}_k{}' \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1/2} \boldsymbol{Y}$$

로 주어지며 $i = 1, 2, \ldots, k-1$번째 정준변수와 서로 독립이면서

$$Corr(U_k, V_k) = \rho_k^*$$

를 최대로 한다.

여기서 $\rho_1^{*2} \geqq \rho_2^2 \geqq \cdots \geqq \rho_p^*$는 $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ 의 고유값이며 해당되는 고유벡터는 $e_1, e_2, \ldots, e_p$이다.

$\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}$의 고유값을 크기 순으로 늘어놓을 때, 고유벡터는 $f_1, f_2, \ldots, f_q$이다.

f_i는 $\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}e_i$ 에 비례한다.

이와 같이 생성된 정준변수에 대해 $(k, l = 1, 2, \ldots, p)$

$$Var(U_k) = Var(V_k) = 1,$$
$$Cov(U_k, U_l) = Corr(U_k, U_l) = 0, \qquad k \neq l$$
$$Cov(V_k, V_l) = Corr(V_k, V_l) = 0, \qquad k \neq l$$
$$Cov(U_k, V_l) = Corr(U_k, V_l) = 0, \qquad k \neq l$$

이 성립한다.

9.2.2 표준화 변수에 대한 정준변수와 정준상관계수

정준계수벡터 a_k에 대해, $Var(X_i)=\sigma_{ii},\ i=1,2,...,p$

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

즉, 표준화 변수의 정준상관변수 계수는 원래 변수 X_i 로부터 구한 것에 $\sqrt{\sigma_{ii}}$를 곱한 형태.

a_k'가 k번째 정준변수 U_k의 계수벡터이면

표준화변수 $\boldsymbol{Z}$의 계수벡터는 $a_k'\boldsymbol{D}_{XX}^{1/2}$ 여기서 $\boldsymbol{D}_{XX}=diag\{\sigma_{11},\sigma_{22},...,\sigma_{pp}\}$

마찬가지로, $b_k'\boldsymbol{D}_{YY}^{-1/2}$는 두 번째 변수 집단 $\boldsymbol{Y}$의 표준화 변수에 대한 정준계수벡터 ($\boldsymbol{D}_{YY}$는 변수들의 분산으로 구성된 대각행렬)

표준화 변수들에 대한 정준상관계수는 변하지 않는다.

≪예제 9.1≫ (표준화 변수들의 정준변수와 정준상관계수의 계산)

표준화 변수들로 구성된 두 개의 변수 집단 $\boldsymbol{Z}^{(1)} = [Z_1^{(1)}, Z_2^{(1)}]'$, $\boldsymbol{Z}^{(2)} = [Z_1^{(2)}, Z_2^{(2)}]'$ 에 대한 정준상관계수를 구하고자 한다.

$\boldsymbol{Z} = [\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}]'$가 공분산행렬

$$Cov(\boldsymbol{Z}) = \begin{pmatrix} \boldsymbol{\rho}_{11} & | & \boldsymbol{\rho}_{12} \\ -- & -- & -- \\ \boldsymbol{\rho}_{21} & | & \boldsymbol{\rho}_{22} \end{pmatrix} = \left(\begin{array}{cc|cc} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ \hline 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{array}\right)$$

을 갖는다고 하자.

$$\boldsymbol{\rho}_{11}^{-1/2} = \begin{pmatrix} 1.0681 & -0.2229 \\ -0.2229 & 1.0681 \end{pmatrix}$$

$$\boldsymbol{\rho}_{22}^{-1} = \begin{pmatrix} 1.0417 & -0.2083 \\ -0.2083 & 1.0417 \end{pmatrix}$$

이므로

$$\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2} = \begin{pmatrix} 0.4371 & 0.2178 \\ 0.2178 & 0.1096 \end{pmatrix}$$

$\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}$의 고유값 ρ_1^{*2}, ρ_2^{*2}는 다음의 특성방정식

$$\begin{aligned} 0 &= \begin{vmatrix} 0.4371-\lambda & 0.2178 \\ 0.2178 & 0.1096-\lambda \end{vmatrix} \\ &= (0.4371-\lambda)(0.1096-\lambda) - (2.178)^2 = \lambda^2 - 0.5467\lambda + 0.0005 \end{aligned}$$

으로부터 $\rho_1^{*2} = 0.5458$와 $\rho_2^{*2} = 0.0009$로 구해진다.

고유값 $\rho_1^{*2} = 0.5458$에 해당하는 고유벡터 $\boldsymbol{e}_1' = [0.8947, 0.4466]$로부터

$$\boldsymbol{a}_1 = \boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{e}_1 = \begin{pmatrix} 0.8561 \\ 0.2776 \end{pmatrix}$$

을 얻는다.

정리 9.1로부터 $\boldsymbol{f}_i \propto \boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{e}_i$ 이고

$$\boldsymbol{b}_1 = \boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{f}_1$$

이므로

$$b_1 \propto \rho_{22}^{-1}\rho_{21}a_1 = \begin{pmatrix} 0.3959 & 0.2292 \\ 0.5209 & 0.3542 \end{pmatrix}\begin{pmatrix} 0.8561 \\ 0.2776 \end{pmatrix} = \begin{pmatrix} 0.4206 \\ 0.5443 \end{pmatrix}$$

이 된다.

$$Var(V_1) = Var(b_1{}'Z^{(2)}) = b_1{}'\rho_{22}b_1 = 1$$

이 되도록 b_1을 구한다.

벡터 $(0.4026,\ 0.5443)'$ 에 대해

$$b_1{}'\rho_{22}b_1 = (0.4206,\ 0.5443)\begin{pmatrix} 1.0 & 02 \\ 02 & 1.0 \end{pmatrix}\begin{pmatrix} 0.4026 \\ 0.5443 \end{pmatrix} = 0.5460$$

이므로 $\sqrt{0.5460} = 0.7389$로 나누어주면

$$b_1 = \frac{1}{0.7389}\begin{pmatrix} 0.4026 \\ 0.5443 \end{pmatrix} = \begin{pmatrix} 0.5448 \\ 0.7366 \end{pmatrix}$$

으로 얻어진다.

그러므로 첫 번째 정준변수 쌍은

$$U_1 = \boldsymbol{a}_1{}' Z^{(1)} = 0.86 Z_1^{(1)} + 0.28 Z_2^{(1)}$$
$$V_1 = \boldsymbol{b}_1{}' Z^{(2)} = 0.54 Z_1^{(2)} + 0.73 Z_2^{(2)}$$

로 구해지며 이렇게 구한 정준변수의 상관계수인 정준상관계수는

$$\rho_1^* = \sqrt{\rho_1^{*2}} = \sqrt{0.5458} = 0.74$$

가 되며 $Z^{(1)}$과 $Z^{(2)}$의 선형결합식 중 최대 상관계수가 된다.

두 번째 정준상관계수는 $\rho_2^* = \sqrt{0.0009} = 0.03$로 매우 작은 값이며 두 변수 집단간의 연관성에 관해 거의 정보를 갖지 않음을 알 수 있다.

$\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}$의 고유벡터를 이용하여 정준계수벡터 $\boldsymbol{a}_k = \Sigma_{XX}^{-1/2}\boldsymbol{e}_k$ 와 $\boldsymbol{b}_k = \Sigma_{YY}^{-1/2}\boldsymbol{f}_k$를 구할 때

계산의 편리상 $|\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \rho^{*2}\boldsymbol{I}| = 0$으부터 계산할 수도 있다.

▶ 행렬의 곱에 대한 고유값의 성질을 이용하여 보자.

$A = \Sigma_{YY}^{-1}\Sigma_{YX}$, $B = \Sigma_{XX}^{-1}\Sigma_{XY}$라 놓으면 $BA = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$와

$AB = \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ 으로부터 고유값을 구하면 고유벡터는 다르지만 같은 고유값 $\rho_1^2, \rho_2^2, \ldots, \rho_s^2$ 을 얻게 될 것이며 여기서 $s = \min(p,q)$이다.

다음의 특성방정식(characteristic equation)

$$|\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \rho^{*2} I| = 0,$$
$$|\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \rho^{*2} I| = 0$$

의 해로 고유값을 구하고 다음의 고유벡터방정식

$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} a = \rho^{*2} a,$$
$$\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} b = \rho^{*2} b$$

으로부터 고유벡터를 구한 후 정준계수벡터는 a, b를 계산하면 된다. 일반적으로 행렬 $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$과 $\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$는 대칭이 아니다.

9.3 표본정준변수와 표본정준상관계수

n 개의 확률표본으로부터 관측된 $(p+q)\times 1$ 확률벡터 $\boldsymbol{W}$는

$$\boldsymbol{X}=(X_1,\ldots,X_p)'\text{와 }\ \boldsymbol{Y}=(Y_1,\ldots,Y_q)',\quad (p\le q)$$

로 구성되어 관측되었다고 하자.

두 변수 집단으로 구성된 $(p+q)\times 1$ 확률벡터 $\boldsymbol{W}$는

$$\boldsymbol{W}=\begin{pmatrix}\boldsymbol{X}\\ \hline \boldsymbol{Y}\end{pmatrix}=\left(\begin{array}{c}X_1\\ \vdots\\ X_p\\ \hline Y_1\\ \vdots\\ Y_q\end{array}\right)$$

일 때 $\boldsymbol{W}$의 표본평균벡터는

$$\overline{W}=\begin{pmatrix}\overline{X}\\---\\\overline{Y}\end{pmatrix}$$

여기서 $\overline{X}=\frac{1}{n}\sum_{j=1}^{n}X_j,\ \overline{Y}=\frac{1}{n}\sum_{j=1}^{n}Y_j$ 이다.

표본공분산행렬

$$S_{(p+q)\times(p+q)}=\begin{pmatrix}S_{XX}S_{XY}\\S_{YX}S_{YY}\end{pmatrix}$$

여기서

$$S_{XX}=\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})(X_j-\overline{X})'$$

$$S_{YY}=\frac{1}{n-1}\sum_{j=1}^{n}(Y_j-\overline{Y})(Y_j-\overline{Y})'$$

$$S_{XY}=\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})(Y_j-\overline{Y})'$$

상수계수벡터 $p\times 1$ 벡터 a와 $q\times 1$ 벡터 b에 대해 변수들의 선형결합식(일변량 확률변수)

$$\hat{U}=a'\boldsymbol{X}$$
$$\hat{V}=b'\boldsymbol{Y}$$

$\hat{U}$와 $\hat{V}$의 분산 추정량 :

$$\widehat{Var}(\hat{U})=a'\widehat{Cov}(\boldsymbol{X})a=a'\boldsymbol{S}_{XX}a$$
$$\widehat{Var}(\hat{V})=b'\widehat{Cov}(\boldsymbol{Y})b=b'\boldsymbol{S}_{YY}b$$

$\hat{U}$와 $\hat{V}$의 상관계수 :

$$r=Corr(\hat{U},\hat{V})=\frac{a'\boldsymbol{S}_{XY}b}{\sqrt{a'\boldsymbol{S}_{XX}a}\ \sqrt{b'\boldsymbol{S}_{YY}b}}$$

모공분산행렬 대신 추정량인 표본공분산행렬을 이용해 표본정준변수와 표본정준상관계수를 구하게 되며 이는 아래의 정리 9.2에서 설명된다.

정리 9.2 $p \le q$이며, $p \times 1$ 확률벡터 $\boldsymbol{X}$와 $q \times 1$ 확률벡터 $\boldsymbol{Y}$가 표본공분산행렬 $\boldsymbol{S}_{XX}, \boldsymbol{S}_{YY}, \boldsymbol{S}_{XY}, \boldsymbol{S}_{YX}$ 를 가질 때 첫 번째 표본정준상관변수는

$$\widehat{U}_1 = \boldsymbol{e}_1{}' \boldsymbol{S}_{XX}^{-1/2} \boldsymbol{X} \text{와 } \widehat{V}_1 = \boldsymbol{f}_1{}' \boldsymbol{S}_{YY}^{-1/2} \boldsymbol{Y}$$

로 주어진다. 또한 k번째 표본정준상관변수는 $k = 2,3,...,p$일 때

$$\widehat{U}_k = \boldsymbol{e}_k{}' \boldsymbol{S}_{XX}^{-1/2} \boldsymbol{X} \text{ 와 } \widehat{V}_k = \boldsymbol{f}_k{}' \boldsymbol{S}_{YY}^{-1/2} \boldsymbol{Y}$$

로 주어지며 $i = 1,2,...,k-1$번째 정준상관변수와 서로 독립이면서

$$Corr(\widehat{U}_k, \widehat{V}_k) = \widehat{\rho}_k^*$$

여기서 $\widehat{\rho}_1^{*2} \ge \widehat{\rho}_2^2 \ge \cdots \ge \widehat{\rho}_p^*$는 $\boldsymbol{S}_{XX}^{-1/2} \boldsymbol{S}_{XY} \boldsymbol{S}_{YY}^{-1} \boldsymbol{S}_{YX} \boldsymbol{S}_{XX}^{-1/2}$의 고유값이며

해당 고유벡터는 $\widehat{e}_1, \widehat{e}_2, ..., \widehat{e}_p$이다. $\boldsymbol{S}_{YY}^{-1/2} \boldsymbol{S}_{YX} \boldsymbol{S}_{XX}^{-1} \boldsymbol{S}_{XY} \boldsymbol{S}_{YY}^{-1/2}$의 고유벡터 $\widehat{f}_k$는

$$\widehat{f}_k = (1/\widehat{\rho}_k^*) \boldsymbol{S}_{YY}^{-1/2} \boldsymbol{S}_{YX} \boldsymbol{S}_{XX}^{-1/2} \widehat{e}_k$$

▶ 표본공분산행렬에 정준상관분석을 적용하여 보자.

$A = S_{YY}^{-1} S_{YX}$, $B = S_{XX}^{-1} S_{XY}$ 라 놓고 $BA = S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX}$ 와 $AB = S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY}$ 으로부터 고유값을 구하면 고유벡터는 다르지만 같은 고유값 $r_1^2, r_2^2, \ldots, r_s^2$을 얻게 된다.

여기서 $s = \min(p, q)$이다.

다음의 특성방정식(characteristic equation)

$$|S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX} - r^2 I| = 0,$$
$$|S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY} - r^2 I| = 0$$

의 해(solution)로 고유값을 구하고 다음의 고유벡터방정식

$$S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX} a = r^2 a, \qquad (9.33)$$
$$S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY} b = r^2 b$$

으로부터 고유벡터를 구한 후 정준계수벡터 a와 b를 계산하면 된다.

▶ 표본상관행렬 $$R_{(p+q)\times(p+q)} = \begin{pmatrix} R_{XX} R_{XY} \\ R_{YX} R_{YY} \end{pmatrix}$$

을 이용할 경우에도 마찬가지 방법으로 정준변수와 정준계수를 구할 수 있다.

일변량 Y에 대해 $R^2 = r_{XY}' R_{XX}^{-1} r_{XY}$는 다변량 $_Y$에 대해 $R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY}$로 확장된다.

$$|R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX} - r^2 I \ | = 0,$$
$$|R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY} - r^2 I \ | = 0$$

의 해로 고유값을 구하고 다음의 고유벡터방정식

$$R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX} c = r^2 c,$$
$$R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY} d = r^2 d$$

으로부터 고유벡터를 구한 후 정준계수벡터는 a와 b를 계산하면 된다.

식(9.33)의 a, b와는

$$c = D_X a \text{ 그리고 } d = D_Y b$$

성립한다. 여기서 $D_X = diag(s_{X_1}, s_{X_2}, ..., s_{X_p})$, $D_Y = diag(s_{Y_1}, s_{Y_2}, ..., s_{Y_q})$이다.

9.4 정준변수에 대한 해석 및 특성

▶ 정준변수는 인공적으로 만들어 낸 변수이므로 인자나 주성분과 같은 절대적 의미를 부여하기는 힘들며 관심있는 변수 집단에 대해 연관성을 알고자 할 때 주로 이용할 수 있다.

▶ 변수를 표준화 하더라도 정준상관계수는 변하지 않으므로 단위의 표준화와 해석을 위해서는 표준화 변수들에 대한 정준상관분석을 권장한다.

▶ 정준변수 특성

(1) 정준상관계수는 변수들의 척도변환에 불변이다(scale invariant).

(2) 첫 번째 정준상관계수 ρ_1^*는 두 변수 집단간의 최대 상관계수이며 두 변수 집단에서 단순상관계수 또는 다중상관계수를 구할 때 ρ_1^*를 넘지 않는다.

▶ 정준변수에 대한 해석을 할 때는 다음과 같은 측면을 이용하여 설명할 수 있도록 한다.

(1) 표준화된 계수: 정준계수는 정준상관계수에 각 변수가 기여하는 정도를 나타낸다.

(2) 정준변수 $V_j = b_j{}' \boldsymbol{Y}$와 q개 변수 $Y_1,...,Y_q$의 (표본)상관관계는 $\boldsymbol{R_{YY}} d_j$로 나타난다. 여기서 $\boldsymbol{R_{YY}}$는 $\boldsymbol{Y}$의 상관행렬이고 d_j는 식(9.36)에서 정의된 바와 같으며 $d_j{}' \boldsymbol{R_{YY}} d_j = 1$을 만족한다.

정리 9.3 **변수 Y_j와 정준변수 $V_1,...,V_s$ 와의 상관계수의 가중합은 $R^2_{Y_j|\boldsymbol{X}}$ 즉 Y_j 와 $\boldsymbol{X}$의 다중상관계수의 제곱으로 나타난다. 즉**

$$\sum_{i=1}^{s} r^2_{Y_j, U_i} r_i^2 = R^2_{Y_j|\boldsymbol{X}}$$

여기서 r_j는 j 번째 정준상관계수이다. 변수 X_j와 정준변수 $U_1,...,U_s$에 대해서도 마찬가지로 성립한다.

$$\sum_{i=1}^{s} r^2_{X_j, V_i} r_i^2 = R^2_{X_j|\boldsymbol{Y}}$$

(3) 각 변수와 정준변수와의 상관성 :

정리 9.3에서와 같이 변수 Y_j와 정준변수 $U_1,...,U_s$와의 상관계수의 가중합은 $R^2_{Y_j|\boldsymbol{X}}$ 즉 Y_j와 $\boldsymbol{X}$의 다중상관계수의 제곱으로 나타난다. 즉 변수 Y_j와 정준변수 $U_1,...,U_s$의 상관계수는 종속변수 Y_j를 설명변수벡터 $\boldsymbol{X}$의 회귀식으로 구했을 때의 다중상관계수형태로 나타난다. 그러나 Y_j가 $\boldsymbol{X}$로 구성된 정준변수 $U_1,...,U_s$에 어떻게 기여하는지는 나타내지 못한다.

9.5 상관성에 대한 검정

정준상관분석은 두 변수 집단 간의 연관성에 대해 설명하고자 하는 것이므로 두 변수 집단 간에 상관성이 존재할 때만 의미있는 분석이 된다.

$$H_0 : \Sigma_{XY} = 0 \qquad \text{에 대해} \qquad H_1 : \Sigma_{XY} \neq 0$$

즉 H_0가 사실이면 X와 Y간에 연관성이 없으며 구해지는 정준상관계수 $r_1, r_2, \ldots, r_s$는 통계적인 의미가 없다. 다음의 검정통계량

$$\Lambda_1 = \frac{|S|}{|S_{YY}||S_{XX}|} = \frac{|R|}{|R_{YY}||R_{XX}|} = \prod_{i=1}^{s}(1-r_i^2)$$

은 H_0 하에서 $\Lambda_{p,q,n-1-q}$ 분포를 따르므로 $\Lambda_1 \leq \Lambda_{p,q,n-1-q}(\alpha)$이면 H_0를 기각한다.

근사적으로 $\chi^2 = -\left[n - \frac{1}{2}(p+q+3)\right]\ln\Lambda_1$

는 자유도가 pq인 카이제곱분포를 따르므로 $\chi^2 \geqq \chi^2_{pq}(\alpha)$이면 H_0를 기각한다.

≪예제 9.2≫ ≪예제 9.1≫의 데이터가 $n=30$에 대해 얻은 결과라 하자. $\boldsymbol{R}$을 이용하여 두 변수 집단 간에 상관성 존재 여부에 대한 검정을 하고자 한다.

$H_0: \boldsymbol{\Sigma_{XY}}=0$에 대해 $H_1: \boldsymbol{\Sigma_{XY}}\neq 0$, 검정통계량은

$$\Lambda_1 = \frac{|\boldsymbol{R}|}{|\boldsymbol{R_{YY}}||\boldsymbol{R_{XX}}|} = \prod_{i=1}^{s}(1-r_i^2) = (1-0.0009)(1-0.5458) = 0.4538$$

이고 $\Lambda_1 = 0.4538 \leq \Lambda_{2,2,27}(0.05) = 0.699$이므로 H_0를 기각한다.

9.6 R을 이용한 정준상관분석

≪예 9.2≫ [표 9.1]이용하여 (Y_1, Y_2)와 (X_1, X_2, X_3)의 정준상관계수와 정준변수를 구하자.

설명변수: $X_1 =$ 온도, $X_2 =$농도, $X_3 =$ 시간

반응변수: $Y_1 =$ 변화하지 않고 남은 양, $Y_2 =$ 반응 후 생성된 양.

▶표 9.1 화학반응실험 결과 화학공정 자료

번호	Y_1	Y_2	X_1	X_2	X_3
1	41.5	45.9	162	23.0	3.0
2	33.8	53.3	162	23.0	8.0
3	27.7	57.5	162	30.0	5.0
4	21.7	58.8	162	30.0	8.0
5	19.9	60.6	172	25.0	5.0
6	15.0	58.0	172	25.0	8.0
7	12.2	58.6	172	30.0	5.0
8	4.3	52.4	172	30.0	8.0
9	19.3	56.9	167	27.5	6.5
10	6.4	55.4	177	27.5	6.5
11	37.6	46.9	157	27.5	6.5
12	18.0	57.3	167	32.5	6.5

● R에서 정준상관분석: cancor() 함수 이용.

* CCA 패키지를 사용할 경우 : cc() 함수 이용.

* yacca 패키지를 사용하는 경우 : cca() 함수.

[결과 9.1](2)에서는 첫 번째 정준상관계수는 0.990이고 두 번째 정준상관계수는 0.0929로 첫 번째 정준상관계수값이 아주 높은 편임을 알 수 있다.

첫 번째 정준변수 쌍:

$$u_1 = -0.1356X_1 - 0.1212X_2 - 0.1585X_3$$

$$v_1 = 0.0832\,Y_1 - 0.0076\,Y_2$$

두 번째 정준변수 쌍:

$$u_2 = 0.0682X_1 - 0.3102X_2 + 0.1735X_3$$

$$v_2 = -0.0633\,Y_1 - 0.2627\,Y_2$$

[프로그램 9.1] 화학공정 자료에 대한 정준상관분석

```
chem=read.csv("C:/data/chem.csv", header=T)
chem ; attach(chem)
n=dim(chem)[[1]] ; n
y=chem[,2:3]
x=chem[,4:6]

library(CCA)
matcor(x,y)   # correlation matrix

cc1 <- cc(x,y)
cc1
cc2<-comput(x,y cc1)
cc2[3:6]     # display canonical loadings

sd<-sd(x)
s1<-diag(sd)       # diagonal matrix of x sd's
s1 %*% cc1$xcoef   # standardized x canonical coefficients
```

```
sd<-sd(y)
s2<-diag(sd)        # diagonal matrix of y sd's
s2 %*% cc1$ycoef    # standardized y canonical coefficients

u1=cc1$scores$xscores[,1]
v1=cc1$scores$yscores[,1]
  plot(u1,v1,pch=18, main="first canonical plot")  # 그림 9.1
u2=cc1$scores$xscores[,2]
v2=cc1$scores$yscores[,2]
  plot(u2,v2,pch=15, main="second canonical plot") # 그림 9.2

plt.cc(cc1, type="v",var.label=TRUE)   # for variales 그림 9.3
plt.cc(cc1, type="i",var.label=TRUE)   # for individuals 그림 9.4

mtc=matcor(x,y)
img.matcor(mtc,type=1)  # images of the correlation matrices
img.matcor(mtc,type=2)  # 그림 9.5
```

[결과 9.1] 정준상관분석 결과 (1)

```
> n=dim(chem)[[1]]
> n
[1] 12
> y=chem[,2:3]
> x=chem[,4:6]

> library(CCA)
```

```
> matcor(x,y)   # correlation matrix
 $Xcor
           x1         x2         x3
x1 1.00000000 0.09857281 0.09505864
x2 0.09857281 1.00000000 0.17178695
x3 0.09505864 0.17178695 1.00000000
 $Ycor
           y1         y2
y1  1.0000000 -0.5822212
y2 -0.5822212  1.0000000
 $XYcor
            x1          x2          x3          y1          y2
x1  1.00000000  0.09857281  0.09505864 -0.8698175  0.5009080
x2  0.09857281  1.00000000  0.17178695 -0.4872731  0.3601848
x3  0.09505864  0.17178695  1.00000000 -0.3892286  0.2243497
y1 -0.86981747 -0.48727311 -0.38922862  1.0000000 -0.5822212
y2  0.50090799  0.36018482  0.22434975 -0.5822212  1.0000000
```

[결과 9.1] 정준상관분석 결과 (2)

```
> cc1[1]        # canonical correlations
$cor
[1] 0.99009125 0.09285027
> cc1[3:4]      # raw canonical coefficients
  $xcoef
         [,1]        [,2]
x1 -0.1355657  0.06823046
x2 -0.1212626 -0.31017337
x3 -0.1585513  0.17350695
  $ycoef
           [,1]        [,2]
y1  0.083239052 -0.06333019
y2 -0.007603168 -0.26270315
```

[결과 9.1] 정준상관분석 결과 (3)

```
> cc2=comput(x,y, cc1) #  compute canonical loadings
> cc2[3:6]  # correlations between variables and canonical variates.
```

```
 $corr.X.xscores
          [,1]       [,2]
x1 -0.8779565  0.3440815
x2 -0.4946925 -0.8609206
x3 -0.3928780  0.1512975
  $corr.Y.xscores
          [,1]         [,2]
y1  0.9896768 -0.002686152
y2 -0.5994987 -0.073894562
```

```
 $corr.X.yscores
          [,1]        [,2]
x1 -0.8692570  0.03194806
x2 -0.4897907 -0.07993671
x3 -0.3889850  0.01404801
  $corr.Y.yscores
          [,1]        [,2]
y1  0.9995814 -0.02892994
y2 -0.6054984 -0.79584653
```

[결과 9.1](4)에서는 표준화 정준 변수를 보여준다.

첫 번째 표준화 정준변수 쌍:

$$u_{1s} = -0.8175X_1 - 0.3709X_2 - 0.2515X_3$$

$$v_{1s} = 0.9789\,Y_1 - 0.0356\,Y_2$$

두 번째 표준화 정준변수 쌍:

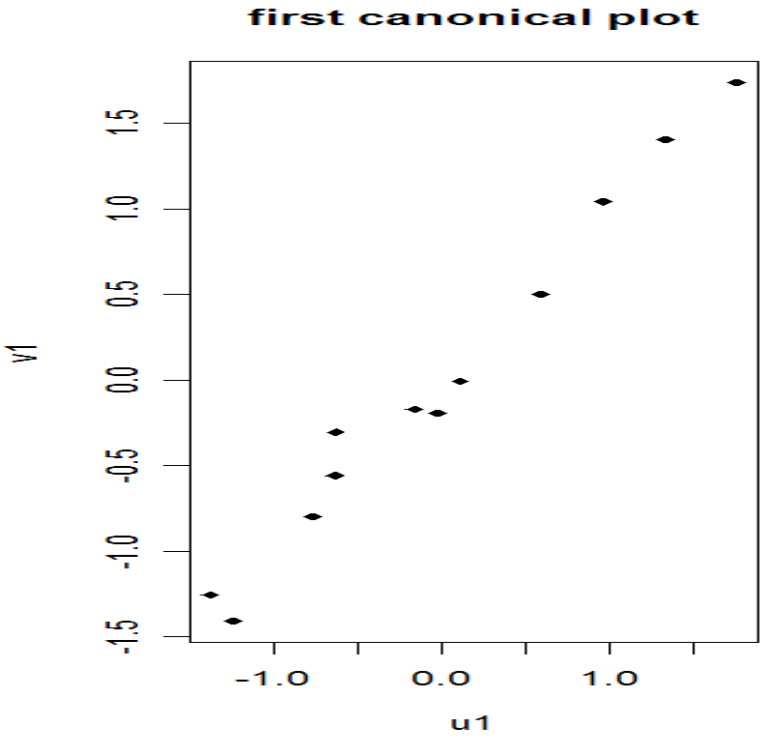
$$u_{2s} = 0.4114X_1 - 0.9487X_2 + 0.2752X_3$$

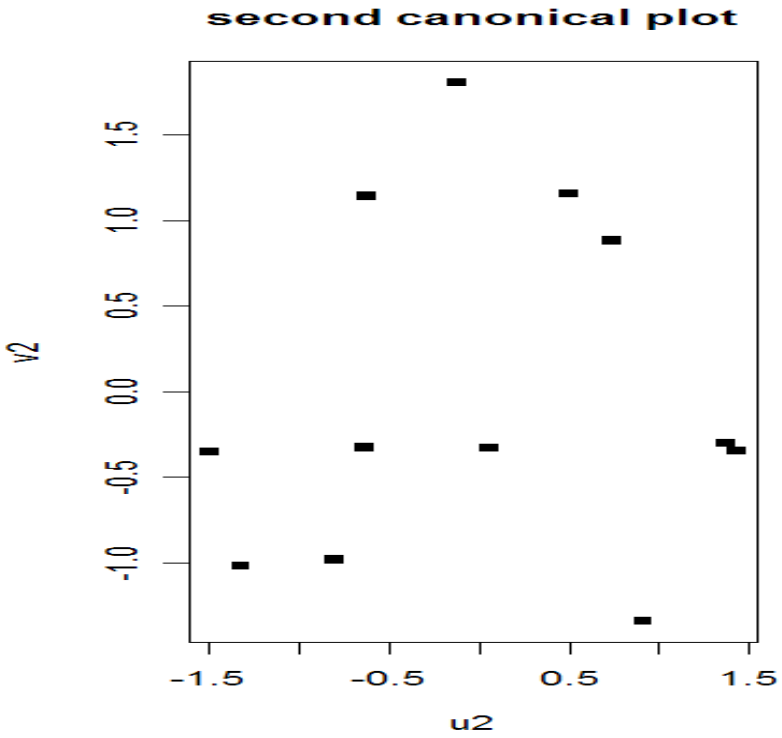
$$v_{2s} = -0.7448\,Y_1 - 1.2295\,Y_2$$

[결과 9.1] 정준상관분석 결과 (4): 표준화 정준 변수

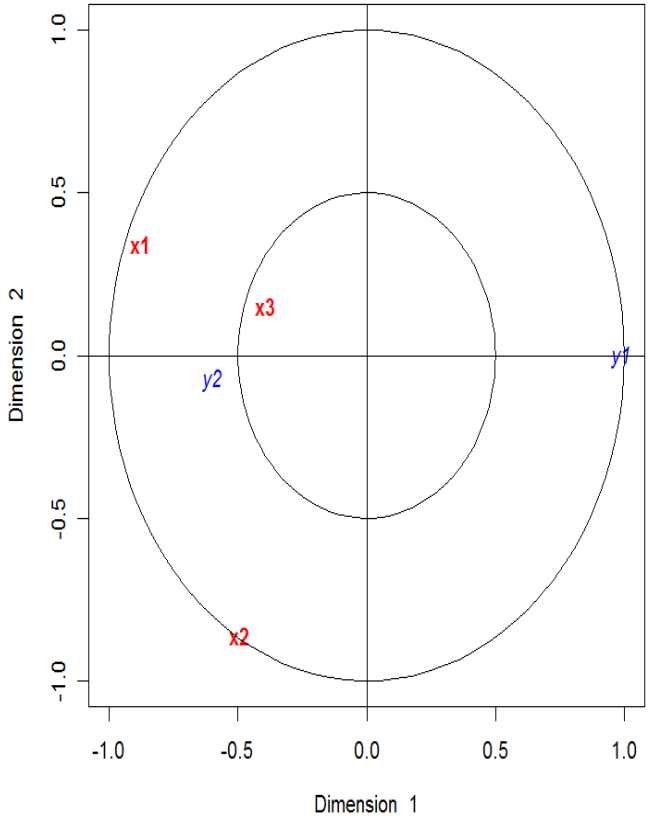
```
> # standardized x canonical coefficients
> sd=sd(x)
> s1=diag(sd)       # diagonal matrix of x sd's
> s1 %*% cc1$xcoef
           [,1]       [,2]
[1,] -0.8174920  0.4114452
[2,] -0.3709141 -0.9487483
[3,] -0.2514501  0.2751686

> # standardized y canonical coefficients
> sd=sd(y)
> s2=diag(sd)       # diagonal matrix of y sd's
> s2 %*% cc1$ycoef
            [,1]       [,2]
[1,]  0.97886436 -0.7447426
[2,] -0.03558284 -1.2294514
```

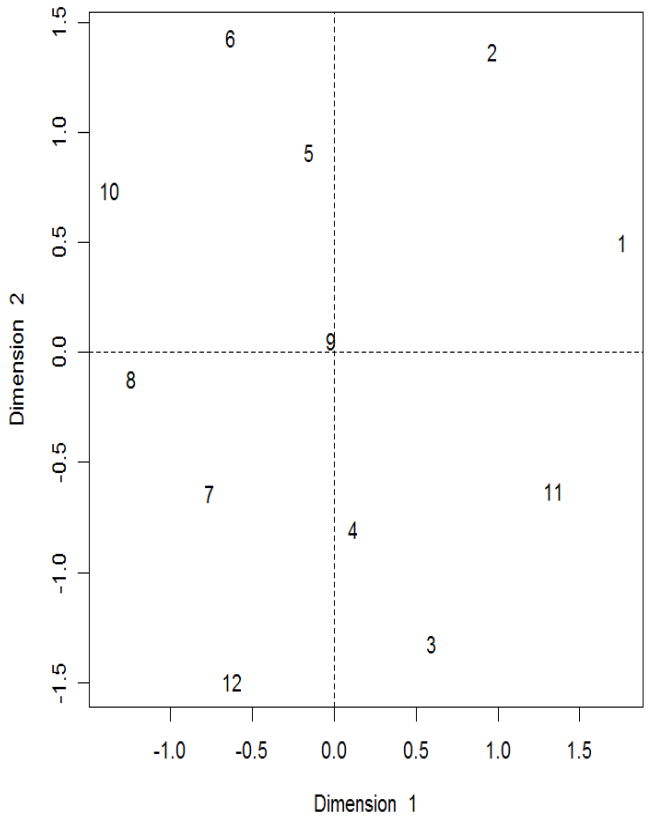




[그림 9.1] 첫 번째 정준 변수 그래프 [그림 9.2] 두 번째 정준 변수 그래프

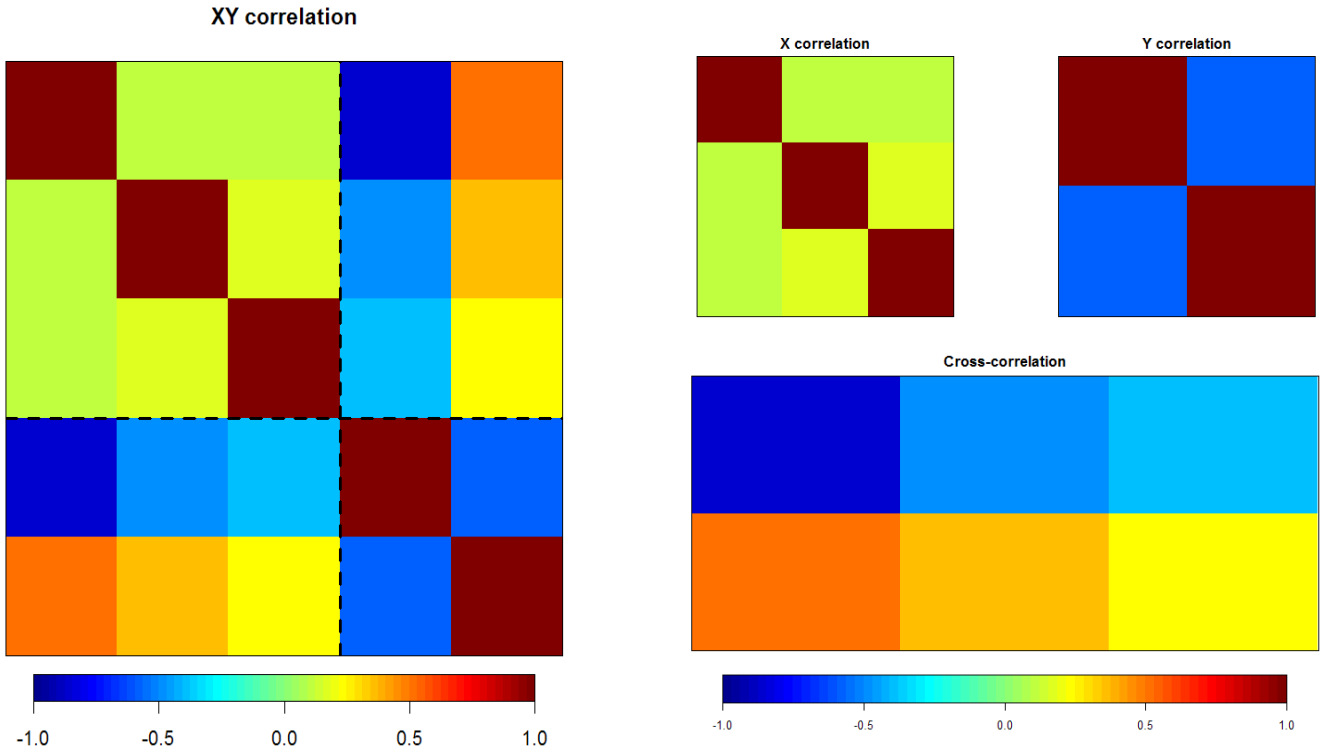


[그림 9.3] 첫 번째 정준 변수 그래프



[그림 9.4] 두 번째 정준 변수 그래프

[그림 9.3]은 2차원 정준변수 차원에서 각 변수의 위치를 나타내며 [그림 9.4]는 2차원 정준변수 차원에서 각 개체의 위치를 보여준다.



[그림 9.5] 두 행렬내 행렬간 상관성 이미지 그림