

연세대학교 통계 데이터 사이언스 학회 ESC 23-2 FALL WEEK4

Canonical Correlation Analysis

[ESC 정규세션 학술부] 장덕재 허정웅



Contents

1. Introduction
2. Canonical Variates and Canonical Correlations
3. Interpreting the Population Canonical Variables
4. The Sample Canonical Variates and Sample Canonical Correlations
5. Additional Interpretation
6. Large Sample Inference
7. R code Implementation





1.Introduction

Introduction

Canonical Correlation Analysis

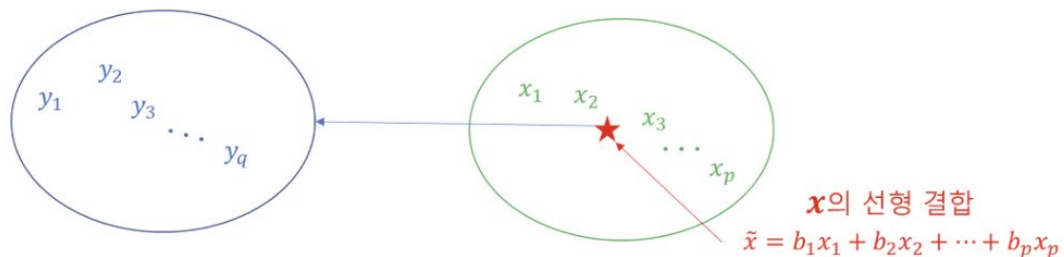
- $\mathbf{X} = [X_1, X_2, \dots, X_p]', \mathbf{Y} = [Y_1, Y_2, \dots, Y_q]'$
- \mathbf{X} 와 \mathbf{Y} 두 개의 set이 있을 때, \mathbf{X} 와 \mathbf{Y} 가 어떠한 관계를 가지고 있는지 밝혀내는 것이 바로 CCA이다.
- 기본적으로 \mathbf{X} 와 \mathbf{Y} 의 관계를 파악하기 위해 $\Sigma_{\mathbf{X}, \mathbf{Y}}$ 를 살펴볼 수 있는데, 이때 각 원소들이 나타내는 것은 개별적인 x_i 와 y_j 의 관계(공분산)다. 그런데 p 와 q 의 크기가 증가하면 개별 공분산으로부터 전체적인 관계를 파악하는 것이 매우 힘들어진다.
- 그렇기에 x_i 들의 선형 결합으로 x_i 들을 대표하는 새로운 변수를 만들고, y_i 들의 선형 결합으로 y_i 들을 대표하는 새로운 변수를 만들어 이를 비교하게 되는 것이다.
- CCA는 두 변수 집단 간의 회귀분석으로도 생각할 수 있다.



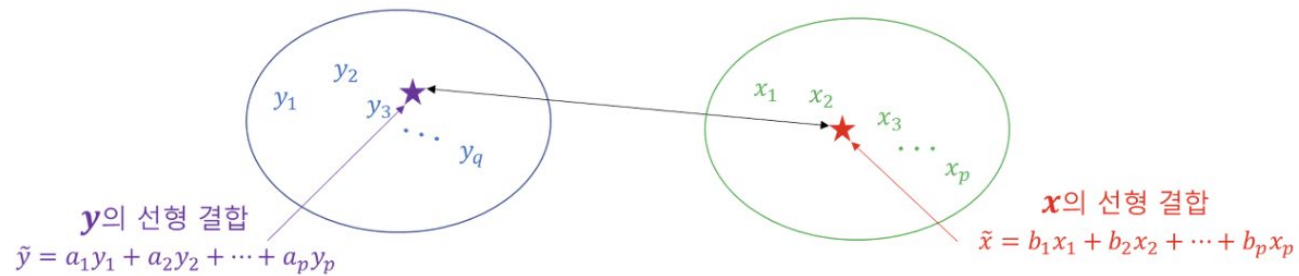
Introduction

Canonical Correlation Analysis

다변량 회귀 분석



정준 상관 분석



Introduction

CCA vs FA

- 관측 데이터로부터 실제로 존재하지 않는 데이터를 찾는다는 점에서 Factor Analysis와 매우 유사하다.
- 차이점1:
 - FA는 어떠한 요인(F_i)이 기존의 변수(X_i)에 영향을 많이 주는지를 분석
 - $X_i = l_{i1}F_1 + l_{i2}F_2 + \dots + l_{im}F_m + \epsilon_i$
 - CCA는 기존의 변수(X_i)들로 그것들의 대표 변수(U)를 찾고자 함
 - $U = a_1X_1 + a_2X_2 + \dots + a_pX_p$
- 차이점2:
 - FA는 서로 다른 요인들 간에 상관계수가 0이라고 가정함
 - CCA는 대표 변수들 간에 상관계수가 최대화가 되는 coefficient를 찾고자 함
- CCA는 구분되는 두 개의 set이 존재하고, set들 간에 집단적인 관계성을 밝혀내고자 할 때 사용
- FA는 특징을 구분 짓지 못한 변수들을 잘 정리하고 싶을 때 사용



Introduction

Example

	SBP	DBP	Height	Weight
1	120	76	165	60
2	109	80	180	80
3	130	82	170	70
4	121	78	185	85
5	135	85	180	90
6	140	87	187	87

	SBP	DBP	Height	Weight
SBP	1			
DBP	0.79	1		
Height	0.25	0.54	1	
Weight	0.37	0.66	0.92	1

Correlation Matrix

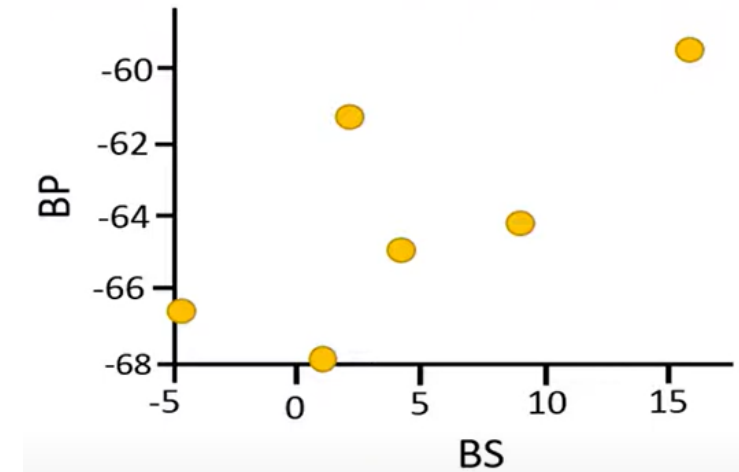
<https://www.youtube.com/watch?v=2tUuyWTtPqM&t=1s>



Introduction

Example

	SBP	DBP	Height	Weight	BP	BS
1	120	76	165	60	-59.6	16.0
2	109	80	180	80	-65.0	4.3
3	130	82	170	70	-64.3	9.1
4	121	78	185	85	-61.5	1.9
5	135	85	180	90	-66.6	-4.8
6	140	87	187	87	-67.9	0.9



Introduction

Example

	SBP	DBP	Height	Weight
1	120	76	165	60
2	109	80	180	80
3	130	82	170	70
4	121	78	185	85
5	135	85	180	90
6	140	87	187	87

$$S_{11} = \begin{bmatrix} 128.567 & 37.267 \\ 37.267 & 17.467 \end{bmatrix}$$

$$S_{22} = \begin{bmatrix} 74.167 & 91.333 \\ 91.333 & 132.667 \end{bmatrix}$$

$$S_{12} = \begin{bmatrix} 24.167 & 48.333 \\ 19.267 & 31.933 \end{bmatrix}$$

$$S_{21} = \begin{bmatrix} 24.167 & 19.267 \\ 48.333 & 31.933 \end{bmatrix}$$



Introduction

Example

	SBP	DBP	Height	Weight
1	120	76	165	60
2	109	80	180	80
3	130	82	170	70
4	121	78	185	85
5	135	85	180	90
6	140	87	187	87

$$s_1 = \begin{bmatrix} -0.093 & -0.081 \\ 0.989 & 0.650 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} 0.131 & -0.526 \\ -0.991 & 0.850 \end{bmatrix}$$

$$\lambda_1 = 0.514, \lambda_2 = 0.009$$

$$r_1 = 0.72, r_2 = 0.03$$



Introduction

Example

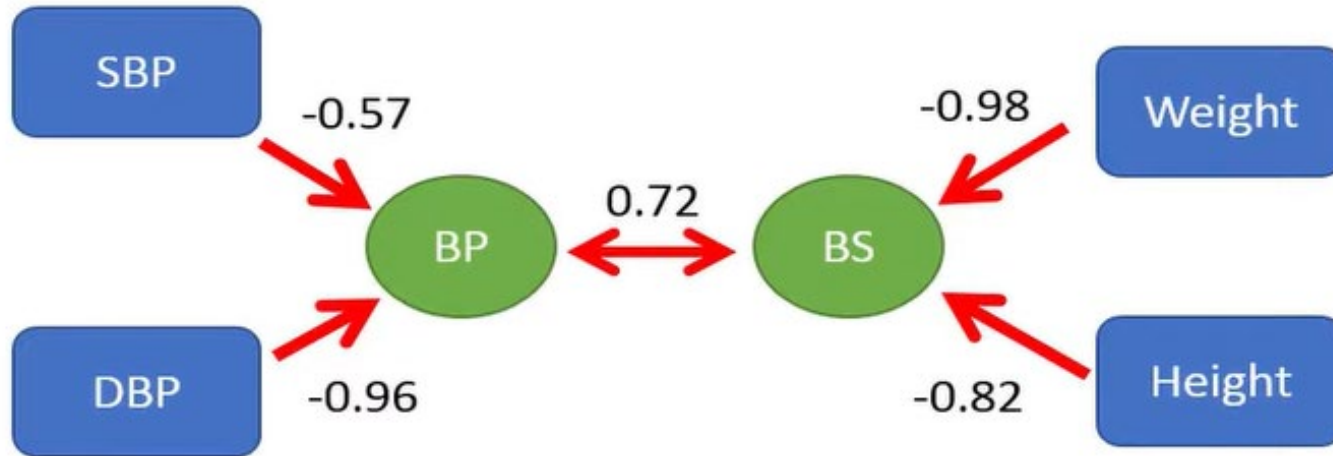
	SBP	DBP	Height	Weight	BP	BS
1	120	76	165	60	-59.6	16.0
2	109	80	180	80	-65.0	4.3
3	130	82	170	70	-64.3	9.1
4	121	78	185	85	-61.5	1.9
5	135	85	180	90	-66.6	-4.8
6	140	87	187	87	-67.9	0.9
Mean	125.8	81.3	177.8	78.7	-64.1	4.6
SD	11.3	4.2	8.6	11.5	3.1	7.2

	BP	BS
SBP	-0.57	-0.41
DBP	-0.96	-0.69
Height	-0.59	-0.82
Weight	-0.71	-0.98

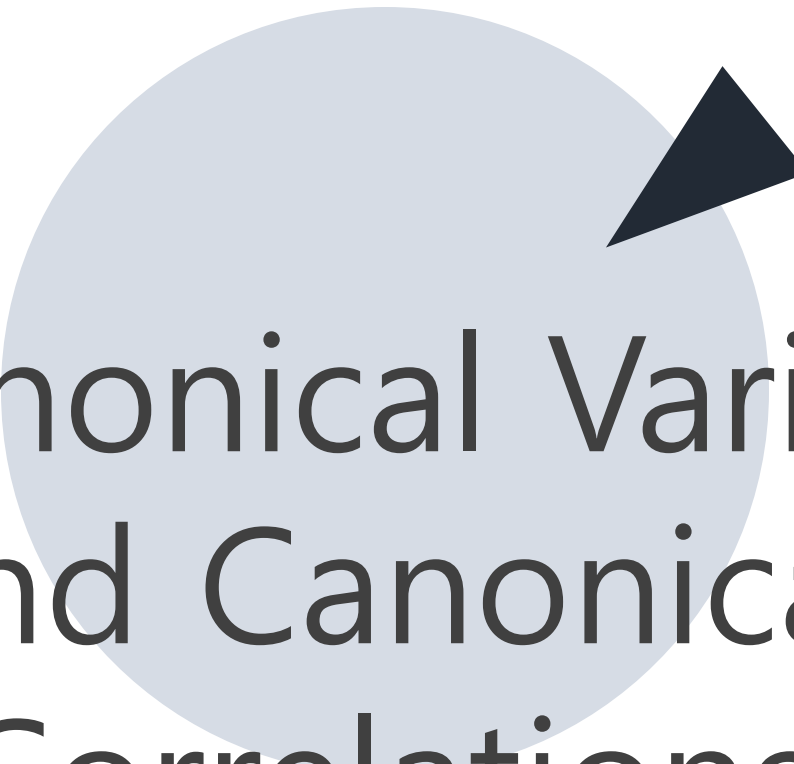


Introduction

Example



	BP	BS
SBP	-0.57	-0.41
DBP	-0.96	-0.69
Height	-0.59	-0.82
Weight	-0.71	-0.98



2. Canonical Variates and Canonical Correlations

Canonical Variates and Canonical Correlation

Notation



Canonical Variates and Canonical Correlation

Canonical Variates

- The first pair of canonical variables
 - (U_1, V_1)
 - Unit Variance
 - Maximum correlation
- The second pair of canonical variables
 - (U_2, V_2)
 - Unit Variance
 - Uncorrelated with (U_1, V_1)
 - Maximum correlation
- The kth pair of canonical variables
 - (U_k, V_k)
 - Unit Variance
 - Uncorrelated with $(U_1, V_1), (U_2, V_2), \dots, (U_{k-1}, V_{k-1})$
 - Maximum correlation

Canonical Correlation

- The kth canonical correlation
 - The correlation between the kth pair of canonical variables
 - $\text{Corr}(U_k, V_k) = \rho_k^*$



Canonical Variates and Canonical Correlation

How to find Canonical Variates and Canonical Correlation

1. Compute $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$
2. Find eigenvalues of $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$ and its corresponding eigenvectors
3. $\rho_1^{*2} \geq \rho_2^{*2} \geq \dots \geq \rho_p^{*2}$ are eigenvalues of $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$ and its corresponding eigenvector are e_1, e_2, \dots, e_p
4. $a_1 = \Sigma_{11}^{-\frac{1}{2}}e_1, a_2 = \Sigma_{11}^{-\frac{1}{2}}e_2, \dots, a_p = \Sigma_{11}^{-\frac{1}{2}}e_p$
5. $U_1 = a_1'X^{(1)}, U_2 = a_2'X^{(1)}, \dots, U_p = a_p'X^{(1)}$
6. Compute $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$
7. Find eigenvalues of $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ and its corresponding eigenvectors
8. $\rho_1^{*2} \geq \rho_2^{*2} \geq \dots \geq \rho_p^{*2}$ are eigenvalues of $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ and its corresponding eigenvector are f_1, f_2, \dots, f_p
9. $b_1 = \Sigma_{22}^{-\frac{1}{2}}f_1, b_2 = \Sigma_{22}^{-\frac{1}{2}}f_2, \dots, b_p = \Sigma_{22}^{-\frac{1}{2}}f_p$
10. $V_1 = b_1'X^{(2)}, V_2 = b_2'X^{(2)}, \dots, V_p = b_p'X^{(2)}$



Canonical Variates and Canonical Correlation

How to find Canonical Variates and Canonical Correlation

Why $\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}$?



Canonical Variates and Canonical Correlation

How to find Canonical Variates and Canonical Correlation

- If original variables are standardized, then we can compute Canonical variates and correlation by Correlation matrices
- $U_k = \mathbf{a}_k^{*'} \mathbf{Z}^{(1)} = \mathbf{e}_k^{*'} \boldsymbol{\rho}_{11}^{-\frac{1}{2}} \mathbf{Z}^{(1)}$
- $V_k = \mathbf{b}_k^{*'} \mathbf{Z}^{(2)} = \mathbf{f}_k^{*'} \boldsymbol{\rho}_{22}^{-\frac{1}{2}} \mathbf{Z}^{(2)}$
- Find eigenvalues and eigenvectors of $\boldsymbol{\rho}_{11}^{-\frac{1}{2}} \boldsymbol{\rho}_{12} \boldsymbol{\rho}_{22}^{-1} \boldsymbol{\rho}_{21} \boldsymbol{\rho}_{11}^{-\frac{1}{2}}$
- Find eigenvalues and eigenvectors of $\boldsymbol{\rho}_{22}^{-\frac{1}{2}} \boldsymbol{\rho}_{21} \boldsymbol{\rho}_{11}^{-1} \boldsymbol{\rho}_{12} \boldsymbol{\rho}_{22}^{-\frac{1}{2}}$
- Relationship between coefficient vector of X and Z
 - If $\mathbf{a}_k' \mathbf{X} = U_k$ then $\mathbf{a}_k' V_{11}^{\frac{1}{2}} \mathbf{Z} = U_k \rightarrow$ when coefficient vector of X is \mathbf{a}_k' , then coefficient vector of Z is $\mathbf{a}_k' V_{11}^{\frac{1}{2}}$
- Canonical Correlations are UNCHANGED by the standardization



Canonical Variates and Canonical Correlation

How to find Canonical Variates and Canonical Correlation



Canonical Variates and Canonical Correlation

Example 10.1

Calculating canonical variates and canonical correlation for standardized variables


$$\mathbf{Z}^{(1)} = [Z_1^{(1)}, Z_2^{(1)}]'$$

$$\mathbf{Z}^{(2)} = [Z_1^{(2)}, Z_2^{(2)}]'$$

$$\mathbf{Z} = [\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}]'$$

$$\text{Cov}(\mathbf{Z}) = \begin{bmatrix} \boldsymbol{\rho}_{11} & \boldsymbol{\rho}_{12} \\ \boldsymbol{\rho}_{21} & \boldsymbol{\rho}_{22} \end{bmatrix} = \begin{bmatrix} 1.0 & .4 & .5 & .6 \\ .4 & 1.0 & .3 & .4 \\ .5 & .3 & 1.0 & .2 \\ .6 & .4 & .2 & 1.0 \end{bmatrix}$$





3. Interpreting the Population Canonical Variables

Interpreting the Population Canonical Variables

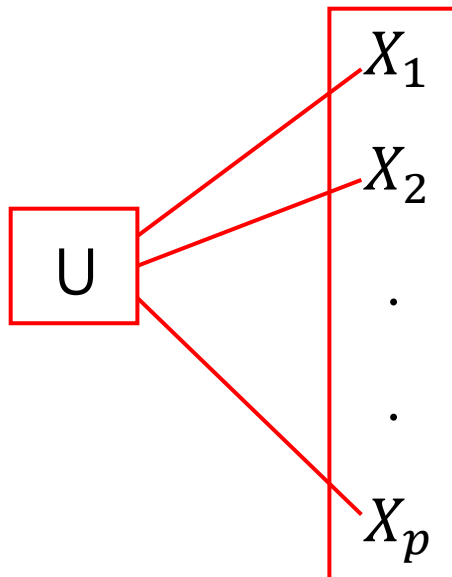
1. Identifying the Canonical Variables
 - 어떻게 Canonical Variables를 이해할 수 있을까?
2. Canonical Correlations as Generalizations of Other Correlation Coefficients
 - 어떻게 Canonical Correlation을 이해할 수 있을까?
3. A Geometrical Interpretation of the Population Canonical Correlation Analysis
 - Original Variables가 어떠한 기하적인 변화를 거쳐 Canonical Variables가 되는 것일까?
4. The First r Canonical Variables as a Summary of Variability



Interpreting the Population Canonical Variables

Identifying the Canonical Variables

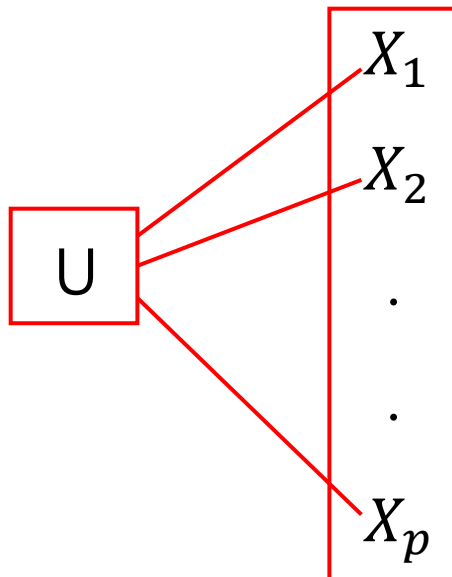
- Canonical Variable U 는 실제 변수가 아니라 x_i 들의 선형 결합으로 만들어진 인위적인 변수이다.
- 하지만 우리는 U 를 집합 $\mathbf{x} = [x_1, \dots, x_p]$ 의 맥락에서 파악할 수 있다.
 - How? U 와 x_i 들의 관계를 살펴봄으로써



Interpreting the Population Canonical Variables

Identifying the Canonical Variables

- Canonical Variable U 는 실제 변수가 아니라 X_i 들의 선형 결합으로 만들어진 인위적인 변수이다.
- 하지만 우리는 U 를 집합 $\mathbf{X} = [X_1, \dots, X_p]$ 의 맥락에서 파악할 수 있다.
 - How? U 와 X_i 들의 관계를 살펴봄으로써



주의해야 할 점:

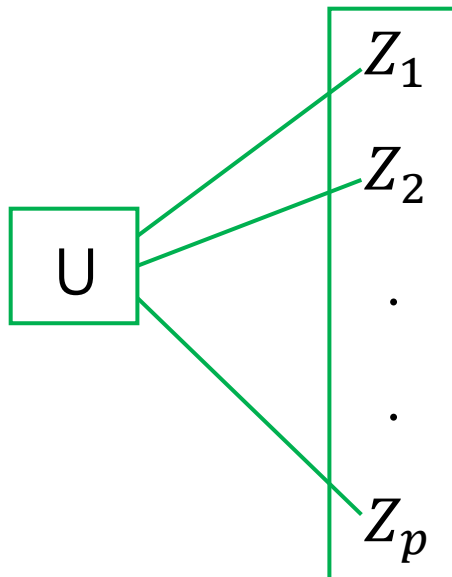
- 이러한 관계는 univariate한 정보만을 전달함
- 따라서 공분산으로 관계를 살펴볼 경우, scale에 따라 단위가 달라지는 문제가 발생할 수 있음
- 그렇기에 상관계수를 활용
- 또는 기존의 변수들을 정규화해서 비교!



Interpreting the Population Canonical Variables

Identifying the Canonical Variables

- Canonical Variable U 는 실제 변수가 아니라 x_i 들의 선형 결합으로 만들어진 인위적인 변수이다.
- 하지만 우리는 U 를 집합 $\mathbf{X} = [X_1, \dots, X_p]$ 의 맥락에서 파악할 수 있다.
 - How? U 와 x_i 들의 관계를 살펴봄으로써



주의해야 할 점:

- 이러한 관계는 univariate한 정보만을 전달함
- 따라서 공분산으로 관계를 살펴볼 경우, scale에 따라 단위가 달라지는 문제가 발생할 수 있음
- 그렇기에 상관계수를 활용
- 또는 기존의 변수들을 정규화해서 비교!



Interpreting the Population Canonical Variables

Identifying the Canonical Variables

- To identify the canonical variables, we will compute correlation between \mathbf{U} and $\mathbf{Z}^{(1)}$
- Notation

$$\underset{(p \times p)}{\mathbf{A}} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p]'$$

$$\text{Cov}(\mathbf{U}, \mathbf{X}^{(1)}) = \text{Cov}(\mathbf{A}\mathbf{X}^{(1)}, \mathbf{X}^{(1)}) = \mathbf{A}\boldsymbol{\Sigma}_{11}$$

$$\underset{(q \times q)}{\mathbf{B}} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_q]'$$

$$\underset{(p \times p)}{\boldsymbol{\rho}_{\mathbf{U}, \mathbf{X}^{(1)}}} = \mathbf{A}\boldsymbol{\Sigma}_{11}\mathbf{V}_{11}^{-1/2}$$

$$\underset{(q \times q)}{\boldsymbol{\rho}_{\mathbf{V}, \mathbf{X}^{(2)}}} = \mathbf{B}\boldsymbol{\Sigma}_{22}\mathbf{V}_{22}^{-1/2}$$

$$\underset{(p \times 1)}{\mathbf{U}} = \mathbf{A}\mathbf{X}^{(1)} \quad \underset{(q \times 1)}{\mathbf{V}} = \mathbf{B}\mathbf{X}^{(2)}$$

$$\underset{(p \times q)}{\boldsymbol{\rho}_{\mathbf{U}, \mathbf{X}^{(2)}}} = \mathbf{A}\boldsymbol{\Sigma}_{12}\mathbf{V}_{22}^{-1/2}$$

$$\underset{(q \times p)}{\boldsymbol{\rho}_{\mathbf{V}, \mathbf{X}^{(1)}}} = \mathbf{B}\boldsymbol{\Sigma}_{21}\mathbf{V}_{11}^{-1/2}$$

$$\boldsymbol{\rho}_{\mathbf{U}, \mathbf{Z}^{(1)}} = \mathbf{A}_z \boldsymbol{\rho}_{11}$$

$$\boldsymbol{\rho}_{\mathbf{V}, \mathbf{Z}^{(2)}} = \mathbf{B}_z \boldsymbol{\rho}_{22}$$

$$\boldsymbol{\rho}_{\mathbf{U}, \mathbf{Z}^{(2)}} = \mathbf{A}_z \boldsymbol{\rho}_{12}$$

$$\boldsymbol{\rho}_{\mathbf{V}, \mathbf{Z}^{(1)}} = \mathbf{B}_z \boldsymbol{\rho}_{21}$$



Interpreting the Population Canonical Variables

Identifying the Canonical Variables



Interpreting the Population Canonical Variables

Identifying the Canonical Variables



Interpreting the Population Canonical Variables

Identifying the Canonical Variables

- Example 10.2: Computing correlations between canonical variates and their component variables

$$\mathbf{Z}^{(1)} = [Z_1^{(1)}, Z_2^{(1)}]'$$

$$\mathbf{Z}^{(2)} = [Z_1^{(2)}, Z_2^{(2)}]'$$

$$\mathbf{Z} = [\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}]'$$

$$\text{Cov}(\mathbf{Z}) = \begin{bmatrix} \boldsymbol{\rho}_{11} & \boldsymbol{\rho}_{12} \\ \boldsymbol{\rho}_{21} & \boldsymbol{\rho}_{22} \end{bmatrix} = \begin{bmatrix} 1.0 & .4 & .5 & .6 \\ .4 & 1.0 & .3 & .4 \\ .5 & .3 & 1.0 & .2 \\ .6 & .4 & .2 & 1.0 \end{bmatrix}$$



Interpreting the Population Canonical Variables

Identifying the Canonical Variables

- Example 10.2: Computing correlations between canonical variates and their component variables

$$\boldsymbol{\rho}_{U_1, \mathbf{Z}^{(1)}} = \mathbf{A}_z \boldsymbol{\rho}_{11} = [.86, .28] \begin{bmatrix} 1.0 & .4 \\ .4 & 1.0 \end{bmatrix} = [.97, .62]$$

$$\boldsymbol{\rho}_{V_1, \mathbf{Z}^{(2)}} = \mathbf{B}_z \boldsymbol{\rho}_{22} = [.54, .74] \begin{bmatrix} 1.0 & .2 \\ .2 & 1.0 \end{bmatrix} = [.69, .85]$$



Interpreting the Population Canonical Variables

Canonical Correlations as Generalizations of Other Correlation Coefficients

1. The first canonical correlation is larger than the absolute value of any entry in $\boldsymbol{\rho}_{12}$



Interpreting the Population Canonical Variables

Canonical Correlations as Generalizations of Other Correlation Coefficients

2. The canonical correlations are also the multiple correlation coefficient of U_k with $X^{(2)}$ or the multiple correlation coefficients of V_k with $X^{(1)}$

- Multiple Correlation Coefficient(R): 변수 집합의 선형 함수를 사용하여 특정 변수를 얼마나 잘 예측할 수 있는지에 대한 척도
- R^2 : 결정계수. 종속변수의 분산을 독립변수 집합이 얼마나 많이 설명하는지에 대한 척도

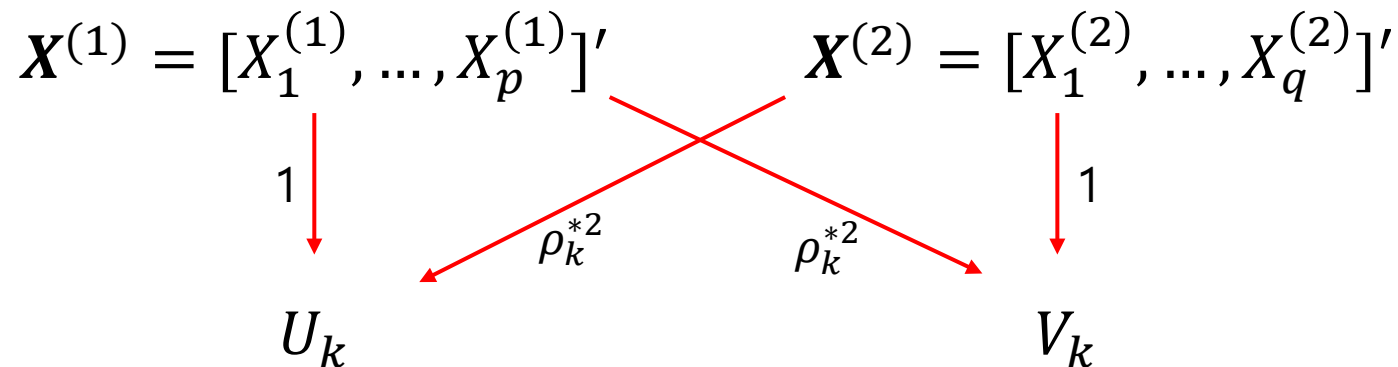


Interpreting the Population Canonical Variables

Canonical Correlations as Generalizations of Other Correlation Coefficients

2. The canonical correlations are also the multiple correlation coefficient of U_k with $X^{(2)}$ or the multiple correlation coefficients of V_k with $X^{(1)}$

- Multiple Correlation Coefficient(R): 변수 집합의 선형 함수를 사용하여 특정 변수를 얼마나 잘 예측할 수 있는지에 대한 척도
- R^2 : 결정계수. 종속변수의 분산을 독립변수 집합이 얼마나 많이 설명하는지에 대한 척도

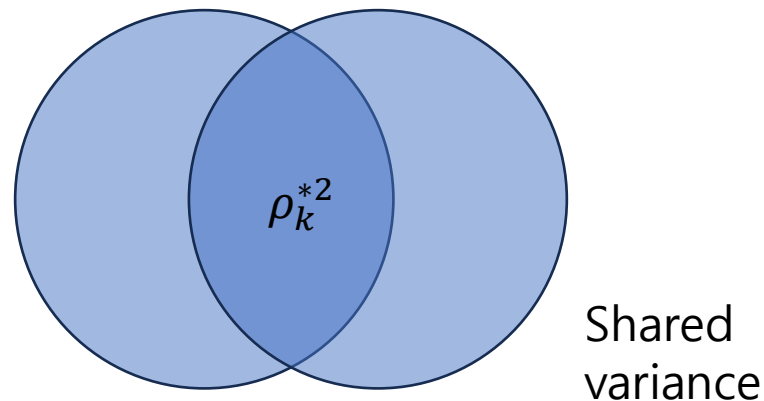


Interpreting the Population Canonical Variables

Canonical Correlations as Generalizations of Other Correlation Coefficients

2. The canonical correlations are also the multiple correlation coefficient of U_k with $X^{(2)}$ or the multiple correlation coefficients of V_k with $X^{(1)}$

- Multiple Correlation Coefficient(R): 변수 집합의 선형 함수를 사용하여 특정 변수를 얼마나 잘 예측할 수 있는지에 대한 척도
- R^2 : 결정계수. 종속변수의 분산을 독립변수 집합이 얼마나 많이 설명하는지에 대한 척도



Interpreting the Population Canonical Variables

A Geometrical Interpretation of the Population Canonical Correlation Analysis

- $U = AX^{(1)}$ 는 기하적으로 보았을 때, (1) $X^{(1)}$ 을 standardized principal component로 바꾼 다음 (2) $P_1(\Sigma_{11})$ 에 의해 결정된 직교행렬에 의해 회전된 다음 (3) E' (full covariance matrix Σ 에 의해 결정된)에 의해 회전된 것으로 볼 수 있다.
- $V = BX^{(1)}$ 도 비슷하게 해석할 수 있다.

$$U = AX^{(1)} = E'\Sigma_{11}^{-1/2}X^{(1)} = E'P_1\Lambda_1^{-1/2}P_1'X^{(1)}$$





4. The Sample Canonical Variates and Sample Canonical Correlations

The Sample Canonical Variates and Sample Canonical Correlations

Notation

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1p}^{(1)} & x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1q}^{(2)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2p}^{(1)} & x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2q}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1}^{(1)} & x_{n2}^{(1)} & \cdots & x_{np}^{(1)} & x_{n1}^{(2)} & x_{n2}^{(2)} & \cdots & x_{nq}^{(2)} \end{bmatrix}$$

$$\mathbf{S}_{(p+q) \times (p+q)} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$$

$(p \times p)$ $(p \times q)$
 $(q \times p)$ $(q \times q)$

$$\hat{U} = \hat{\mathbf{a}}' \mathbf{x}^{(1)}; \quad \hat{V} = \hat{\mathbf{b}}' \mathbf{x}^{(2)}$$

$$r_{\hat{U}, \hat{V}} = \frac{\hat{\mathbf{a}}' \mathbf{S}_{12} \hat{\mathbf{b}}}{\sqrt{\hat{\mathbf{a}}' \mathbf{S}_{11} \hat{\mathbf{a}}} \sqrt{\hat{\mathbf{b}}' \mathbf{S}_{22} \hat{\mathbf{b}}}}$$



The Sample Canonical Variates and Sample Canonical Correlations

Notation

$$\mathbf{S}_{11}^{-1/2} \mathbf{S}_{12} \mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11}^{-1/2}$$

$$\mathbf{S}_{22}^{-1/2} \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} \mathbf{S}_{22}^{-1/2}$$

$$r_{\hat{U}_k, \hat{V}_k} = \widehat{\rho_k^*}$$

$$\hat{U}_k = \underbrace{\hat{\mathbf{e}}_k' \mathbf{S}_{11}^{-1/2} \mathbf{x}^{(1)}}_{\hat{\mathbf{a}}_k'}$$

$$\hat{V}_k = \underbrace{\hat{\mathbf{f}}_k' \mathbf{S}_{22}^{-1/2} \mathbf{x}^{(2)}}_{\hat{\mathbf{b}}_k'}$$

2부



Contents

1. Additional Interpretation

- Purpose of CCA
- Matrices of Errors of Approximations : first r canonical variable이 유효한가?
- Proportions of Explained Sample Variance

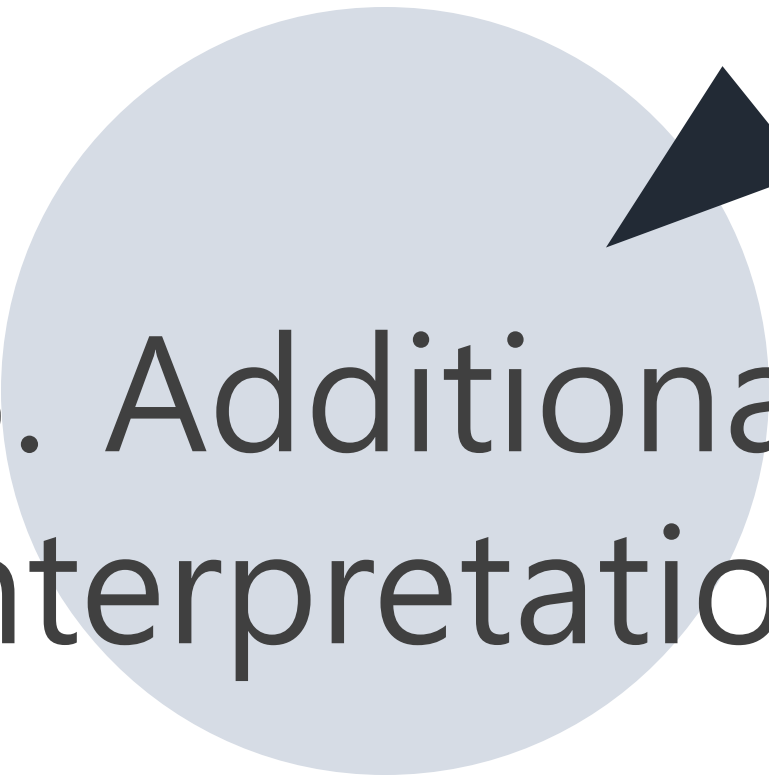
2. Large Sample Inference

- Likelihood Ratio Test

3. R code implementation

- Notations
- code





5. Additional Interpretation

Additional Interpretation – Purpose of CCA

- CCA의 목적

1. 변수 집단 간 관계 파악 : 두 data set 사이의 상관관계를 분석하여 한 set의 변수들이 다른 set과 어떻게 연관되어있는지 파악
eg. 심리적 요인들과 학업적 요인들의 상호작용 파악
2. 다중공선성 문제 해결 : canonical variable들은 orthogonal \Rightarrow canonical variable로 분석을 할 경우 다중공선성 문제 해결
3. 차원 축소 : CCA를 통해 $(p+q)*n$ 개의 데이터를 $(r+r)*n$ 개의 데이터(canonical score)로 축소할 수가 있음
4. 다변량 분석 : canonical score로 클러스터링, 시각화, 회귀분석과 같은 다양한 분석을 할 수 있음

**Canonical Score : 실제 data set에 canonical coefficient(weight)를 곱한 값 \Rightarrow 실제 data set - $p*n$ (X), $q*n$ (Y) , Canonical score - $r*n$ (U), $r*n$ (V)



Additional Interpretation – Residual Matrix

- Matrices of Errors of Approximations (Residual Matrix)
 - 결국에는 전체 p 개의 canonical variable을 사용하는 것이 아닌 두 집단 간의 상관관계, 집단 내의 variance를 잘 설명할 수 있는 r 개의 canonical variable을 사용할 것임 → r 을 어떻게 선택? 검정?
 - 우선은 선택한 r 개의 canonical variable이 얼마나 집단 간의 관계, 집단 내의 variance를 잘 설명하는지 알아보자
 - 어떻게 파악할 수 있을까? r 개의 canonical variable을 사용하는 것은 일종의 estimation
⇒ 회귀분석에서 residual 개념을 사용(true value - estimated value)
⇒ residual matrix를 이용해서 파악 ($\mathbf{S}_{11} - \tilde{\mathbf{S}}_{11}$, $\mathbf{S}_{22} - \tilde{\mathbf{S}}_{22}$, $\mathbf{S}_{12} - \tilde{\mathbf{S}}_{12}$)



Additional Interpretation – Residual Matrix

- 일단 $\mathbf{S}_{11}, \mathbf{S}_{22}, \mathbf{S}_{12}$ 를 canonical correlation과 canonical weight로 표현해보자

let $\hat{\mathbf{a}}^{(i)}, \hat{\mathbf{b}}^{(i)}$ denote the i th column of $\hat{\mathbf{A}}^{-1}, \hat{\mathbf{B}}^{-1}$

since $\hat{\mathbf{U}} = \hat{\mathbf{A}}\mathbf{x}^{(1)}, \hat{\mathbf{V}} = \hat{\mathbf{B}}\mathbf{x}^{(2)}$

$\Rightarrow \mathbf{x}^{(1)} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{U}}, \mathbf{x}^{(2)} = \hat{\mathbf{B}}^{-1}\hat{\mathbf{V}}$ where $\mathbf{x}^{(1)}$ is $(p \times 1)$ $\mathbf{x}^{(2)}$ is $(q \times 1)$



Additional Interpretation – Residual Matrix

- 일단 $\mathbf{S}_{11}, \mathbf{S}_{22}, \mathbf{S}_{12}$ 를 canonical correlation과 canonical weight로 표현해보자

$$\text{Cov}(\hat{\mathbf{U}}, \hat{\mathbf{V}}) = \hat{\mathbf{A}}\mathbf{S}_{12}\hat{\mathbf{B}}', \quad \text{Cov}(\hat{\mathbf{U}}) = \hat{\mathbf{A}}\mathbf{S}_{11}\hat{\mathbf{A}}' = \mathbf{I}, \quad \text{Cov}(\hat{\mathbf{V}}) = \hat{\mathbf{B}}\mathbf{S}_{22}\hat{\mathbf{B}}' = \mathbf{I}$$

$$\Rightarrow \mathbf{S}_{12} = \hat{\mathbf{A}}^{-1} \begin{bmatrix} \widehat{\rho_1^*} & 0 & \cdots & 0 \\ 0 & \widehat{\rho_2^*} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\rho_p^*} \end{bmatrix} (\hat{\mathbf{B}}^{-1})' = \widehat{\rho_1^*} \hat{\mathbf{a}}^{(1)} \hat{\mathbf{b}}^{(1)'} + \widehat{\rho_2^*} \hat{\mathbf{a}}^{(2)} \hat{\mathbf{b}}^{(2)'} + \cdots + \widehat{\rho_p^*} \hat{\mathbf{a}}^{(p)} \hat{\mathbf{b}}^{(p)'} \quad (10-32)$$

$$\mathbf{S}_{11} = (\hat{\mathbf{A}}^{-1})(\hat{\mathbf{A}}^{-1})' = \hat{\mathbf{a}}^{(1)}\hat{\mathbf{a}}^{(1)'} + \hat{\mathbf{a}}^{(2)}\hat{\mathbf{a}}^{(2)'} + \cdots + \hat{\mathbf{a}}^{(p)}\hat{\mathbf{a}}^{(p)'}$$

$$\mathbf{S}_{22} = (\hat{\mathbf{B}}^{-1})(\hat{\mathbf{B}}^{-1})' = \hat{\mathbf{b}}^{(1)}\hat{\mathbf{b}}^{(1)'} + \hat{\mathbf{b}}^{(2)}\hat{\mathbf{b}}^{(2)'} + \cdots + \hat{\mathbf{b}}^{(q)}\hat{\mathbf{b}}^{(q)'}$$



Additional Interpretation – Residual Matrix

- Estimated Sets

$$\tilde{\mathbf{x}}^{(1)} = [\hat{\mathbf{a}}^{(1)} \mid \hat{\mathbf{a}}^{(2)} \mid \dots \mid \hat{\mathbf{a}}^{(r)}] \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_r \end{bmatrix}$$

$$\tilde{\mathbf{x}}^{(2)} = [\hat{\mathbf{b}}^{(1)} \mid \hat{\mathbf{b}}^{(2)} \mid \dots \mid \hat{\mathbf{b}}^{(r)}] \begin{bmatrix} \hat{V}_1 \\ \hat{V}_2 \\ \vdots \\ \hat{V}_r \end{bmatrix}$$

- Estimated Covariance

$$\tilde{\mathbf{S}}_{11} = \hat{\mathbf{a}}^{(1)} \hat{\mathbf{a}}^{(1)'} + \dots + \hat{\mathbf{a}}^{(r)} \hat{\mathbf{a}}^{(r)'}$$

$$\tilde{\mathbf{S}}_{22} = \hat{\mathbf{b}}^{(1)} \hat{\mathbf{b}}^{(1)'} + \dots + \hat{\mathbf{b}}^{(r)} \hat{\mathbf{b}}^{(r)'}$$

$$\tilde{\mathbf{S}}_{12} = \hat{\rho}_1^* \hat{\mathbf{a}}^{(1)} \hat{\mathbf{b}}^{(1)'} + \dots + \hat{\rho}_r^* \hat{\mathbf{a}}^{(r)} \hat{\mathbf{a}}^{(r)'}$$



Additional Interpretation – Residual Matrix

- Residual Matrix

$$\mathbf{S}_{11} = (\hat{\mathbf{a}}^{(1)}\hat{\mathbf{a}}^{(1)'} + \hat{\mathbf{a}}^{(2)}\hat{\mathbf{a}}^{(2)'} + \dots + \hat{\mathbf{a}}^{(r)}\hat{\mathbf{a}}^{(r)'}) = \hat{\mathbf{a}}^{(r+1)}\hat{\mathbf{a}}^{(r+1)'} + \dots + \hat{\mathbf{a}}^{(p)}\hat{\mathbf{a}}^{(p)'}$$

$$\mathbf{S}_{22} = (\hat{\mathbf{b}}^{(1)}\hat{\mathbf{b}}^{(1)'} + \hat{\mathbf{b}}^{(2)}\hat{\mathbf{b}}^{(2)'} + \dots + \hat{\mathbf{b}}^{(r)}\hat{\mathbf{b}}^{(r)'}) = \hat{\mathbf{b}}^{(r+1)}\hat{\mathbf{b}}^{(r+1)'} + \dots + \hat{\mathbf{b}}^{(q)}\hat{\mathbf{b}}^{(q)'}$$

$$\begin{aligned}\mathbf{S}_{12} &= (\hat{\rho}_1^* \hat{\mathbf{a}}^{(1)}\hat{\mathbf{b}}^{(1)'} + \hat{\rho}_2^* \hat{\mathbf{a}}^{(2)}\hat{\mathbf{b}}^{(2)'} + \dots + \hat{\rho}_r^* \hat{\mathbf{a}}^{(r)}\hat{\mathbf{b}}^{(r)'}) \\ &= \hat{\rho}_{r+1}^* \hat{\mathbf{a}}^{(r+1)}\hat{\mathbf{b}}^{(r+1)'} + \dots + \hat{\rho}_p^* \hat{\mathbf{a}}^{(p)}\hat{\mathbf{b}}^{(p)'}\end{aligned}$$



Additional Interpretation – Residual Matrix

$$\begin{aligned}\mathbf{S}_{11} - (\hat{\mathbf{a}}^{(1)}\hat{\mathbf{a}}^{(1)'} + \hat{\mathbf{a}}^{(2)}\hat{\mathbf{a}}^{(2)'} + \dots + \hat{\mathbf{a}}^{(r)}\hat{\mathbf{a}}^{(r)'}) &= \hat{\mathbf{a}}^{(r+1)}\hat{\mathbf{a}}^{(r+1)'} + \dots + \hat{\mathbf{a}}^{(p)}\hat{\mathbf{a}}^{(p)'} \\ \mathbf{S}_{22} - (\hat{\mathbf{b}}^{(1)}\hat{\mathbf{b}}^{(1)'} + \hat{\mathbf{b}}^{(2)}\hat{\mathbf{b}}^{(2)'} + \dots + \hat{\mathbf{b}}^{(r)}\hat{\mathbf{b}}^{(r)'}) &= \hat{\mathbf{b}}^{(r+1)}\hat{\mathbf{b}}^{(r+1)'} + \dots + \hat{\mathbf{b}}^{(q)}\hat{\mathbf{b}}^{(q)'} \\ \mathbf{S}_{12} - (\widehat{\rho}_1^* \hat{\mathbf{a}}^{(1)}\hat{\mathbf{b}}^{(1)'} + \widehat{\rho}_2^* \hat{\mathbf{a}}^{(2)}\hat{\mathbf{b}}^{(2)'} + \dots + \widehat{\rho}_r^* \hat{\mathbf{a}}^{(r)}\hat{\mathbf{b}}^{(r)'}) \\ &= \widehat{\rho}_{r+1}^* \hat{\mathbf{a}}^{(r+1)}\hat{\mathbf{b}}^{(r+1)'} + \dots + \widehat{\rho}_p^* \hat{\mathbf{a}}^{(p)}\hat{\mathbf{b}}^{(p)'}\end{aligned}$$

$\Rightarrow \mathbf{S}_{12} - \tilde{\mathbf{S}}_{12}$ 는 canonical correlations 과 canonical coefficients의 곱으로 이루어져 $\mathbf{S}_{11} - \tilde{\mathbf{S}}_{11}$ 와 $\mathbf{S}_{22} - \tilde{\mathbf{S}}_{22}$ 보다 상대적으로 작은 값을 가짐 $\rightarrow \mathbf{S}_{11} - \tilde{\mathbf{S}}_{11}$ 와 $\mathbf{S}_{22} - \tilde{\mathbf{S}}_{22}$ 보다 실제 covariance 값을 더 잘 표현할 수밖에 없음

\Rightarrow 반면에 $\mathbf{S}_{11} - \tilde{\mathbf{S}}_{11}$ 와 $\mathbf{S}_{22} - \tilde{\mathbf{S}}_{22}$ 는 실제 covariance를 잘 설명하지 못할 수 있음

(set $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$ 의 sampling variability를 효과적으로 설명하지 못할 수 있음)

\Rightarrow 얼마나 잘 설명하는지 척도가 있는 것은 아님 (얼마나 큰지, 얼마나 작은지 판단 불가)



Additional Interpretation – Residual Matrix

Example 10.6 (Calculating matrices of errors of approximation) In Example 10.4, we obtained the canonical correlations between the two head and the two leg variables for white leghorn fowl. Starting with the sample correlation matrix

$$\mathbf{R} = \left[\begin{array}{cc|cc} \mathbf{R}_{11} & \mathbf{R}_{12} & & \\ \mathbf{R}_{21} & \mathbf{R}_{22} & & \end{array} \right] = \left[\begin{array}{cc|cc} 1.0 & .505 & .569 & .602 \\ .505 & 1.0 & .422 & .467 \\ \hline .569 & .422 & 1.0 & .926 \\ .602 & .467 & .926 & 1.0 \end{array} \right] \Rightarrow \text{Standardized} \rightarrow \mathbf{R} = \mathbf{S}$$

we obtained the two sets of canonical correlations and variables

$$\begin{aligned} \hat{\rho}_1^* &= .631 & \begin{aligned} \hat{U}_1 &= .781z_1^{(1)} + .345z_2^{(1)} \\ \hat{V}_1 &= .060z_1^{(2)} + .944z_2^{(2)} \end{aligned} & \Rightarrow \text{높은 canonical correlation} & \rightarrow r = 1 \text{로 두기로 하자} \end{aligned}$$

and

$$\begin{aligned} \hat{\rho}_2^* &= .057 & \begin{aligned} \hat{U}_2 &= -.856z_1^{(1)} + 1.106z_2^{(1)} \\ \hat{V}_2 &= -2.648z_1^{(2)} + 2.475z_2^{(2)} \end{aligned} & \Rightarrow \text{상당히 낮은 canonical correlation} \end{aligned}$$



Additional Interpretation – Residual Matrix

where $z_i^{(1)}$, $i = 1, 2$ and $z_i^{(2)}$, $i = 1, 2$ are the standardized data values for sets 1 and 2, respectively.

We first calculate (see Panel 10.1)

$$\hat{\mathbf{A}}_z^{-1} = \begin{bmatrix} .781 & .345 \\ -.856 & 1.106 \end{bmatrix}^{-1} = \begin{bmatrix} \hat{\mathbf{a}}^{(1)} & \hat{\mathbf{a}}^{(2)} \\ .9548 & -.2974 \\ .7388 & .6739 \end{bmatrix}$$
$$\hat{\mathbf{B}}_z^{-1} = \begin{bmatrix} \hat{\mathbf{b}}^{(1)} & \hat{\mathbf{b}}^{(2)} \\ .9343 & -.3564 \\ .9997 & .0227 \end{bmatrix}$$



Additional Interpretation – Residual Matrix

Consequently, the matrices of errors of approximation created by using only the first canonical pair are

$$\mathbf{R}_{12} - \text{sample Cov}(\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)}) = \begin{matrix} \hat{\rho}_2^* & \hat{\mathbf{a}}^{(2)} \\ \hat{\mathbf{b}}^{(2)'} \end{matrix} \begin{bmatrix} -.2974 \\ .6739 \end{bmatrix} \begin{bmatrix} -.3564 & .0227 \end{bmatrix}$$

$$= \begin{bmatrix} .006 & -.000 \\ -.014 & .001 \end{bmatrix}$$

$$\mathbf{R}_{11} - \text{sample Cov}(\tilde{\mathbf{z}}^{(1)}) = \begin{matrix} \hat{\mathbf{a}}^{(2)} \\ \hat{\mathbf{a}}^{(2)'} \end{matrix} \begin{bmatrix} -.2974 \\ .6739 \end{bmatrix} \begin{bmatrix} -.2974 & .6739 \end{bmatrix}$$

$$= \begin{bmatrix} .088 & -.200 \\ -.200 & .454 \end{bmatrix}$$

$$\mathbf{R}_{22} - \text{sample Cov}(\tilde{\mathbf{z}}^{(2)}) = \begin{matrix} \hat{\mathbf{b}}^{(2)} \\ \hat{\mathbf{b}}^{(2)'} \end{matrix} \begin{bmatrix} -.3564 \\ .0227 \end{bmatrix} \begin{bmatrix} -.3564 & .0227 \end{bmatrix}$$

$$= \begin{bmatrix} .127 & -.008 \\ -.008 & .001 \end{bmatrix}$$

$\Rightarrow \mathbf{S}_{12} - \tilde{\mathbf{S}}_{12}$ 는 canonical correlations 과 canonical coefficients의 곱으로 이루어져 $\mathbf{S}_{11} - \tilde{\mathbf{S}}_{11}$ 와 $\mathbf{S}_{22} - \tilde{\mathbf{S}}_{22}$ 보다 상대적으로 작은 값을 가짐 $\rightarrow \mathbf{S}_{11} - \tilde{\mathbf{S}}_{11}$ 와 $\mathbf{S}_{22} - \tilde{\mathbf{S}}_{22}$ 보다 실제 covariance 값을 더 잘 표현할 수밖에 없음

\Rightarrow 반면에 $\mathbf{S}_{11} - \tilde{\mathbf{S}}_{11}$ 와 $\mathbf{S}_{22} - \tilde{\mathbf{S}}_{22}$ 는 실제 covariance를 잘 설명하지 못할 수 있음
(set $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$ 의 sampling variability를 효과적으로 설명하지 못할 수 있음)

\Rightarrow first pair of canonical variable 은 \mathbf{R}_{12} 를 잘 설명하지만, $\mathbf{R}_{11}, \mathbf{R}_{22}$ 를 잘 설명하지 못함



Additional Interpretation – Proportions of Variance

- 설명력 : 앞에서 구한 Residual Matrix보다 구체적으로 수치화 (집단 내 variability)
 - 각 set에서 도출한 canonical variable이 얼마나 집단을 잘 대표하고 있는가?
⇒ canonical variable의 설명력은 집단 내 총 분산과 canonical variable로 구한 estimated variance의 비율을 구하면 됨

각 set의 total variance를 first r canonical variable가 얼마나 설명할 수 있는가?

⇒ total variance와 estimated variance의 비율을 구하면 된다



Additional Interpretation – Proportions of Variance

- $\text{Cov}(\mathbf{X}^{(1)}, \hat{\mathbf{U}}) = \text{Cov}(\hat{\mathbf{A}}^{-1}\hat{\mathbf{U}}, \hat{\mathbf{U}}) = \hat{\mathbf{A}}^{-1}\text{Cov}(\hat{\mathbf{U}}) = \hat{\mathbf{A}}^{-1}$

\Rightarrow first r columns of $\hat{\mathbf{A}}^{-1}$ 는 $\hat{U}_1, \dots, \hat{U}_r$ 과 $X_1^{(1)}, \dots, X_p^{(1)}$ 와의 sample covariance와 같다.

\Rightarrow sample Cov ($\mathbf{z}^{(1)}, \hat{\mathbf{U}}$) = sample Cov ($\hat{\mathbf{A}}_z^{-1}\hat{\mathbf{U}}, \hat{\mathbf{U}}$) = $\hat{\mathbf{A}}_z^{-1}$

sample Cov ($\mathbf{z}^{(2)}, \hat{\mathbf{V}}$) = sample Cov ($\hat{\mathbf{B}}_z^{-1}\hat{\mathbf{V}}, \hat{\mathbf{V}}$) = $\hat{\mathbf{B}}_z^{-1}$

$$\hat{\mathbf{A}}_z^{-1} = [\hat{\mathbf{a}}_z^{(1)}, \hat{\mathbf{a}}_z^{(2)}, \dots, \hat{\mathbf{a}}_z^{(p)}] = \begin{bmatrix} r_{\hat{U}_1, z_1^{(1)}} & r_{\hat{U}_2, z_1^{(1)}} & \cdots & r_{\hat{U}_p, z_1^{(1)}} \\ r_{\hat{U}_1, z_2^{(1)}} & r_{\hat{U}_2, z_2^{(1)}} & \cdots & r_{\hat{U}_p, z_2^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\hat{U}_1, z_p^{(1)}} & r_{\hat{U}_2, z_p^{(1)}} & \cdots & r_{\hat{U}_p, z_p^{(1)}} \end{bmatrix}$$

$$\hat{\mathbf{B}}_z^{-1} = [\hat{\mathbf{b}}_z^{(1)}, \hat{\mathbf{b}}_z^{(2)}, \dots, \hat{\mathbf{b}}_z^{(q)}] = \begin{bmatrix} r_{\hat{V}_1, z_1^{(2)}} & r_{\hat{V}_2, z_1^{(2)}} & \cdots & r_{\hat{V}_q, z_1^{(2)}} \\ r_{\hat{V}_1, z_2^{(2)}} & r_{\hat{V}_2, z_2^{(2)}} & \cdots & r_{\hat{V}_q, z_2^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\hat{V}_1, z_q^{(2)}} & r_{\hat{V}_2, z_q^{(2)}} & \cdots & r_{\hat{V}_q, z_q^{(2)}} \end{bmatrix}$$



Additional Interpretation – Proportions of Variance

- Total (standardized) sample variance in first set

$$= \text{tr}(\mathbf{R}_{11}) = \text{tr}(\hat{\mathbf{a}}_z^{(1)}\hat{\mathbf{a}}_z^{(1)'} + \hat{\mathbf{a}}_z^{(2)}\hat{\mathbf{a}}_z^{(2)'} + \cdots + \hat{\mathbf{a}}_z^{(p)}\hat{\mathbf{a}}_z^{(p)'}) = p \quad \Rightarrow \quad \text{Cov}(\mathbf{Z}_1^{(1)}) = \mathbf{1} \rightarrow \sum_{i=1}^p \text{Cov}(\mathbf{Z}_i^{(1)}) = p$$

- Total (standardized) sample variance in second set

$$= \text{tr}(\mathbf{R}_{22}) = \text{tr}(\hat{\mathbf{b}}_z^{(1)}\hat{\mathbf{b}}_z^{(1)'} + \hat{\mathbf{b}}_z^{(2)}\hat{\mathbf{b}}_z^{(2)'} + \cdots + \hat{\mathbf{b}}_z^{(q)}\hat{\mathbf{b}}_z^{(q)'}) = q \quad \Rightarrow \quad \text{Cov}(\mathbf{Z}_1^{(2)}) = \mathbf{1} \rightarrow \sum_{i=1}^q \text{Cov}(\mathbf{Z}_i^{(2)}) = q$$

- Estimated (standardized) sample variance in first set

$$\text{tr}(\hat{\mathbf{a}}_z^{(1)}\hat{\mathbf{a}}_z^{(1)'} + \hat{\mathbf{a}}_z^{(2)}\hat{\mathbf{a}}_z^{(2)'} + \cdots + \hat{\mathbf{a}}_z^{(r)}\hat{\mathbf{a}}_z^{(r)'}) = \sum_{i=1}^r \sum_{k=1}^p r_{\hat{U}_{i,z_k}^{(1)}}^2$$

⇒ Total 과 Estimated의 비율을 구하면 된다

- Estimated (standardized) sample variance in second set

$$\text{tr}(\hat{\mathbf{b}}_z^{(1)}\hat{\mathbf{b}}_z^{(1)'} + \hat{\mathbf{b}}_z^{(2)}\hat{\mathbf{b}}_z^{(2)'} + \cdots + \hat{\mathbf{b}}_z^{(r)}\hat{\mathbf{b}}_z^{(r)'}) = \sum_{i=1}^r \sum_{k=1}^p r_{\hat{V}_{i,z_k}^{(2)}}^2$$



Additional Interpretation – Proportions of Variance

$$\begin{aligned} R_{\mathbf{Z}^{(1)}|\hat{U}_1, \hat{U}_2, \dots, \hat{U}_r}^2 &= \left(\begin{array}{c} \text{proportion of total standardized} \\ \text{sample variance in first set} \\ \text{explained by } \hat{U}_1, \hat{U}_2, \dots, \hat{U}_r \end{array} \right) \\ &= \frac{\text{tr}(\hat{\mathbf{a}}_{\mathbf{Z}}^{(1)} \hat{\mathbf{a}}_{\mathbf{Z}}^{(1)'} + \dots + \hat{\mathbf{a}}_{\mathbf{Z}}^{(r)} \hat{\mathbf{a}}_{\mathbf{Z}}^{(r)'})}{\text{tr}(\mathbf{R}_{11})} \\ &= \frac{\sum_{i=1}^r \sum_{k=1}^p r_{\hat{U}_i, z_k^{(1)}}^2}{p} \end{aligned}$$

$$\begin{aligned} R_{\mathbf{Z}^{(2)}|\hat{V}_1, \hat{V}_2, \dots, \hat{V}_r}^2 &= \left(\begin{array}{c} \text{proportion of total standardized} \\ \text{sample variance in second set} \\ \text{explained by } \hat{V}_1, \hat{V}_2, \dots, \hat{V}_r \end{array} \right) \\ &= \frac{\text{tr}(\hat{\mathbf{b}}_{\mathbf{Z}}^{(1)} \hat{\mathbf{b}}_{\mathbf{Z}}^{(1)'} + \dots + \hat{\mathbf{b}}_{\mathbf{Z}}^{(r)} \hat{\mathbf{b}}_{\mathbf{Z}}^{(r)'})}{\text{tr}(\mathbf{R}_{22})} \\ &= \frac{\sum_{i=1}^r \sum_{k=1}^q r_{\hat{V}_i, z_k^{(2)}}^2}{q} \end{aligned}$$

⇒ total variance 와 estimated variace의 비율 → 각 set의 분산구조에 대한 설명력

⇒ canonical variable이 각 set을 잘 요약하는지 알아볼 수 있는 척도

⇒ 다른 set 간의 설명력은 구할 수가 없음





6. Large Sample Inference

Large Sample Inference

Likelihood Ratio Test를 통해서

1. Canonical Correlation Analysis 자체가 유의한가?
2. r 을 어떻게 선택할 것인가?

1. Canonical Correlation Analysis 자체가 유의한가?

$$\text{if } \Sigma_{12} = 0 \rightarrow \text{Cov}(U, V) = a' \Sigma_{12} b = 0$$

- Canonical Correlation Analysis 자체가 불가능
- 가설검정을 위해 Likelihood Ratio Test를 시행



Large Sample Inference

- LRT(Likelihood Ratio Test)

$$\lambda_{LR} = -2(l(\theta_0) - l(\hat{\theta})) \rightarrow \chi^2$$

H_0 에서의 θ_0 와 estimated parameter $\hat{\theta}$ 의 log-likelihood의 차이는 chi-square를 asymptotic하게 따른다.

$H_0 : \Sigma_{12} = 0 \quad vs \quad H_1 : \Sigma_{12} \neq 0 \quad \Rightarrow S_{12}$ 가 Σ_{12} 의 unbiased estimator 이므로 Sample Size가 크면 S_{12} 를 Σ_{12} 로 대체

$$\text{test statistic : } -2 \log \lambda = n \log \left(\frac{|S_{11}| |S_{22}|}{|S|} \right) = -n \log \prod_{i=1}^p (1 - \hat{\rho}_i^{*2})$$



Large Sample Inference

$$H_0 : \Sigma_{12} = 0 \quad vs \quad H_1 : \Sigma_{12} \neq 0$$

$$\text{test statistic : } -2 \log \lambda = n \log \left(\frac{|S_{11}| |S_{22}|}{|S|} \right) = -n \log \prod_{i=1}^p (1 - \hat{\rho}_i^{*2})$$

$$\begin{vmatrix} S_{11} & 0 \\ 0' & S_{22} \end{vmatrix} = |S_{11}| |S_{22}| \rightarrow H_0 \text{ 에서의 sample generalized variance}$$

$$|S| \rightarrow H_1 \text{ 에서의 sample generalized variance}$$



Large Sample Inference

for better approximation, replace n with $n - 1 - \frac{1}{2}(p + q + 1)$

\rightarrow *Reject* $H_0 : \Sigma_{12} = 0$ ($\rho_1^* = \dots = \rho_p^*$) ⁼⁰ at significance level α
if $-(n - 1 - \frac{1}{2}(p + q + 1)) \log \prod_{i=1}^p (1 - \hat{\rho}_i^{*2}) > \chi_{pq}^2(\alpha)$



Large Sample Inference

2. r 을 어떻게 선택할 것인가?

그렇다면 first k canonical correlation이 non-zero이고 나머지가 zero인지 검정하는 방법은?

$$H_0^k : \rho_1^* \neq 0, \rho_2^* \neq 0, \dots, \rho_k^* \neq 0, \rho_{k+1}^* = 0, \dots, \rho_p^* = 0$$

$$H_1^k : \rho_i^* \neq 0, \text{ for some } i \geq k+1$$

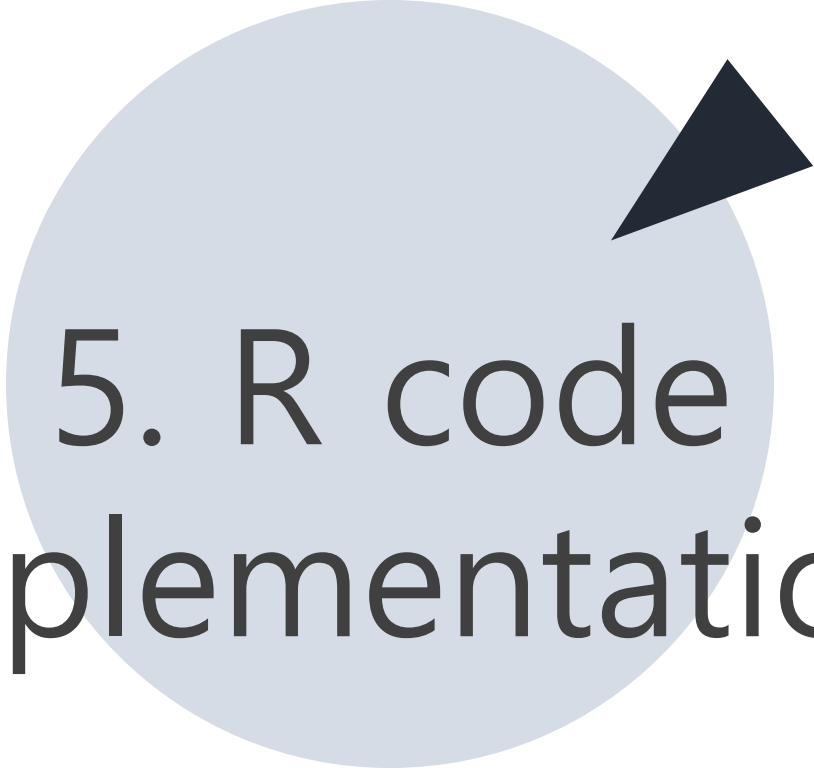
→ *Reject H_0^k at significance level α*

$$\text{if } -\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \log \prod_{i=k+1}^p (1 - \hat{\rho}_i^{*2}) > \chi_{(p-k)(q-k)}^2(\alpha)$$

→ rough guides for selecting the number of the important canonical variables

한계 : 여러번 test를 진행해야할 경우 유의수준은 α 가 아니게 됨.





5. R code implementation

R code implementation

- Canonical Weights : The canonical weights are the coefficients that are used to create the canonical variables from the original variables. They are similar to the regression coefficients in multiple regression, and they indicate how much each variable contributes to the formation of the canonical variable. The canonical weights can help you determine which variables are most important or influential in creating the canonical variables, and how they affect the canonical correlation. However, the canonical weights are not directly comparable across different canonical variables, because they depend on the scaling and ordering of the variables. Therefore, it is more common to use the standardized canonical weights, which are normalized by the standard deviations of the variables.
 - A, B가 Canonical Weights
 - 회귀계수처럼 각 변수가 canonical variable에 얼마만큼 기여를 하는지 알 수 있음
 - canonical weights를 통해 어떤 변수가 중요한지, 얼마만큼 canonical correlation에 영향을 주는지 알 수가 있음
 - Standardized Canonical Weights를 쓰는 이유? → 다른 Canonical Variable과 Canonical Weights를 비교하기가 어려움 → normalize해서 다른 variable의 weights와 비교



R code implementation

- Canonical Scores : The canonical scores are the values of the canonical variables for each observation or case in the data. They are computed by multiplying the original variables by the canonical weights, and they represent the position of each case on the canonical dimensions. The canonical scores can help you visualize and cluster the cases based on their similarities or differences on the canonical variables, and to perform further analyses such as classification, regression, or discriminant analysis. You can also use the canonical scores to examine the outliers or extreme cases that have high or low values on the canonical variables.
 - 실제 Canonical Variable의 값(AX_1, BX_2)
 - Canonical dimensions에 어떤 위치를 갖는지 나타냄
 - plotting을 할 때 Canonical Scores를 이용하여 시각화하고 클러스터링을 할 수 있음.
 - 또한 Canonical Scores를 이용해서 실제 회귀분석 등과 같은 데이터 분석도 가능



R code implementation

- Canonical Loadings : The canonical loadings are the correlations between the original variables and the canonical variables, which are the linear combinations of variables that maximize the canonical correlation coefficients. The canonical loadings can help you interpret the meaning and direction of the canonical variables, and how they relate to each other. For example, if a variable has a high positive loading on a canonical variable, it means that it contributes positively to the canonical correlation, and that it is positively associated with the other variables that have high positive loadings on the same canonical variable. The canonical loadings can also be used to compute the canonical redundancy, which is the amount of variance in one set of variables that is explained by the other set of variables.
 - $\rho_{U,Z^{(1)}} = A_z \rho_{11}$, $\rho_{V,Z^{(2)}} = B_z \rho_{22}$
 - 이를 통해 canonical variable의 의미와 방향을 알 수 있음
 - canonical loading이 높다는 것은 집단 내의 변수가 canonical variable에 큰 기여를 하고 있다는 것임 → 해당 변수 (*ex.* $Z_1^{(1)}$)가 canonical correlation에 positive한 영향을 미치고 있다고 볼 수 있음. 그리고 다른 높은 canonical loading을 가지는 다른 변수들과 positive하게 연관되어 있다고 볼 수 있음



R code implementation

- Canonical Cross Loadings : The canonical cross-loadings are the correlations between the original variables and the canonical variables of the other set. They can help you assess the degree of overlap or redundancy between the two sets of variables, and how they influence each other. For example, if a variable has a high positive cross-loading on a canonical variable of the other set, it means that it is highly correlated with that canonical variable, and that it shares some common variance with the variables that have high loadings on that canonical variable. The canonical cross-loadings can also be used to compute the canonical communality, which is the amount of variance in one set of variables that is shared with the other set of variables.
 - $\rho_{U,Z^{(2)}} = A_z \rho_{12}$, $\rho_{V,Z^{(1)}} = B_z \rho_{21}$
 - cross loading을 통해 overlap(집단 간에 공유되는 정보량)을 알아내는데 도움이 됨
 - 두 집단이 서로 어떤 영향을 주는지 알 수 있음
 - cross loading이 높다는 것은 한 집단 내의 변수가 다른 집단의 선형결합인 canonical variable에 큰 영향을 미치고 있다는 것임



R code implementation

```
mm = read.csv("https://stats.idre.ucla.edu/stat/data/mmreg.csv")
colnames(mm) <- c("Control", "Concept", "Motivation", "Read", "Write", "Math",
  "Science", "Gender")
summary(mm)
```

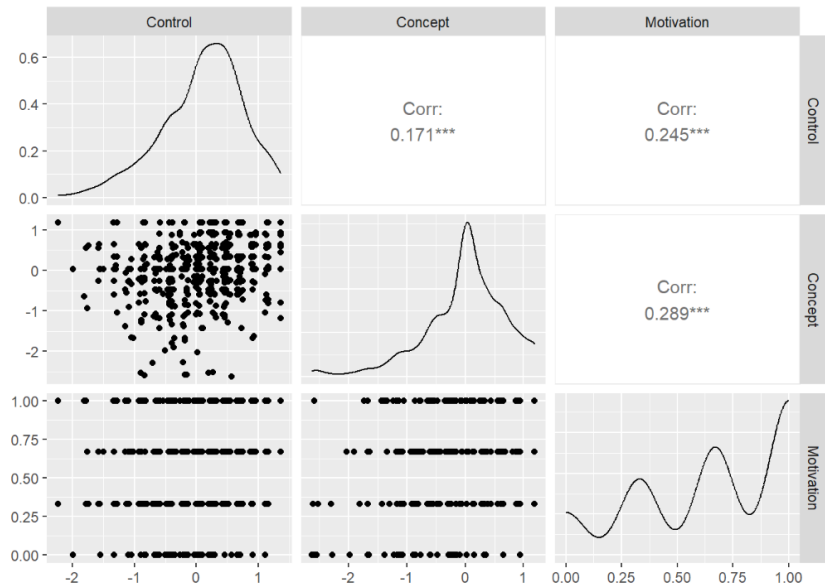
```
##      Control      Concept      Motivation      Read
## Min.   :-2.23000  Min.   :-2.620000  Min.    :0.0000  Min.    :28.3
## 1st Qu.: -0.37250  1st Qu.: -0.300000  1st Qu.: 0.3300  1st Qu.: 44.2
## Median : 0.21000  Median : 0.030000  Median : 0.6700  Median : 52.1
## Mean   : 0.09653  Mean   : 0.004917  Mean   : 0.6608  Mean   : 51.9
## 3rd Qu.: 0.51000  3rd Qu.: 0.440000  3rd Qu.: 1.0000  3rd Qu.: 60.1
## Max.    : 1.36000  Max.    : 1.190000  Max.    : 1.0000  Max.    : 76.0
##      Write      Math      Science      Gender
## Min.    :25.50   Min.    :31.80   Min.    :26.00   Min.    :0.000
## 1st Qu.: 44.30   1st Qu.: 44.50   1st Qu.: 44.40   1st Qu.: 0.000
## Median : 54.10   Median : 51.30   Median : 52.60   Median : 1.000
## Mean    : 52.38   Mean    : 51.85   Mean    : 51.76   Mean    : 0.545
## 3rd Qu.: 59.90   3rd Qu.: 58.38   3rd Qu.: 58.65   3rd Qu.: 1.000
## Max.    : 67.10   Max.    : 75.50   Max.    : 74.20   Max.    : 1.000
```



R code implementation

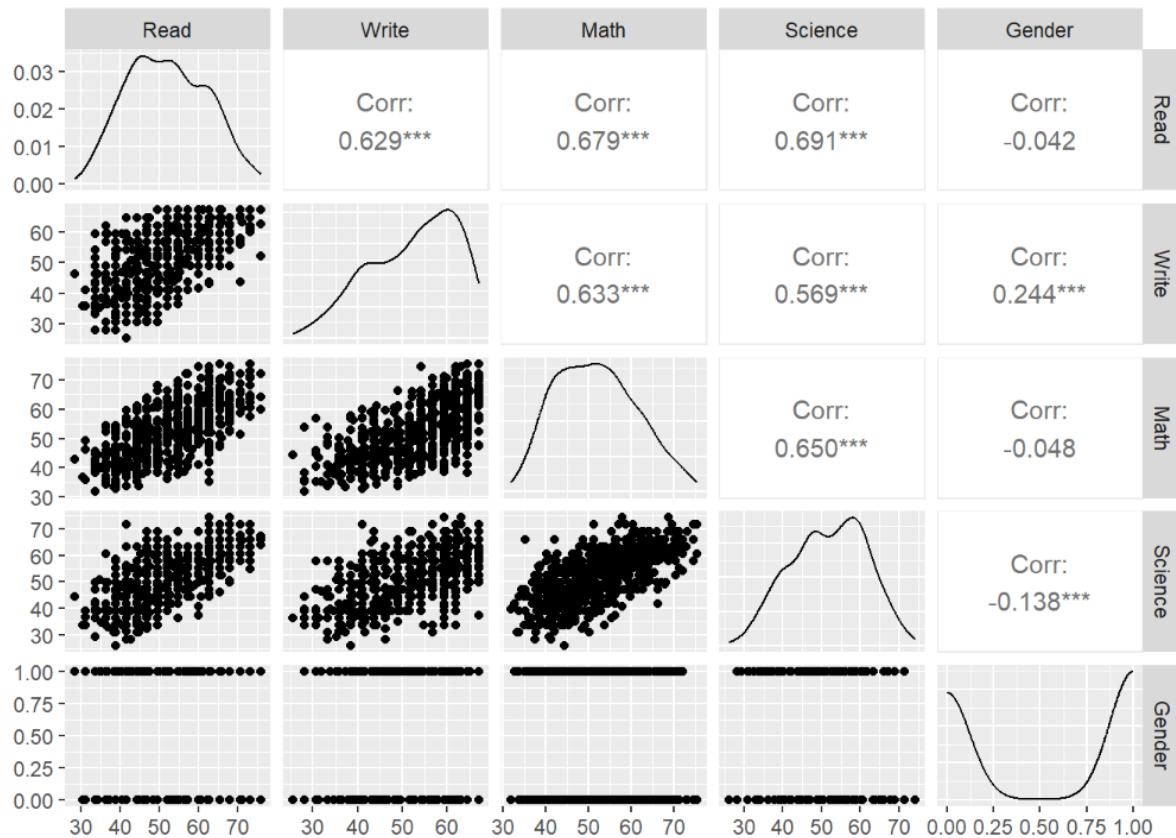
심리적 요인을 `psych set`으로 묶고, 학업적 요인을 `acad set`으로 묶어줌

```
psych = mm[, 1:3]  
acad = mm[, 4:8]  
  
ggpairs(psych)
```



R code implementation

```
ggpairs(acad)
```



R code implementation

R_{11} , R_{22} , R_{12} 를 구함

```
matcor(psych, acad)
```

```
## $Xcor
##           Control  Concept Motivation
## Control  1.0000000 0.1711878  0.2451323
## Concept   0.1711878 1.0000000  0.2885707
## Motivation 0.2451323 0.2885707  1.0000000
##
## $Ycor
##           Read    Write    Math    Science    Gender
## Read    1.0000000 0.6285909  0.6792757  0.6906929 -0.04174278
## Write    0.6285909 1.0000000  0.6326664  0.5691498  0.24433183
## Math     0.6792758 0.6326664  1.0000000  0.6495261 -0.04821830
## Science  0.6906929 0.5691498  0.6495261  1.0000000 -0.13818587
## Gender  -0.04174278 0.2443318 -0.0482183 -0.1381859  1.00000000
##
## $XYcor
##           Control  Concept Motivation    Read    Write    Math
## Control  1.0000000 0.1711878 0.24513227 0.37356505 0.35887684 0.3372690
## Concept   0.1711878 1.0000000 0.28857075 0.06065584 0.01944856 0.0535977
## Motivation 0.2451323 0.28857075 1.00000000 0.21060992 0.25424818 0.1950135
## Read      0.3735650 0.06065584 0.21060992 1.00000000 0.62859089 0.6792757
## Write     0.3588768 0.01944856 0.25424818 0.62859089 1.00000000 0.6326664
## Math      0.3372690 0.05359770 0.19501347 0.67927568 0.63266640 1.0000000
## Science   0.3246269 0.06982633 0.11566948 0.69069291 0.56914983 0.6495261
## Gender    0.1134108 -0.12595132 0.09810277 -0.04174278 0.24433183 -0.0482183
##
##           Science    Gender
## Control  0.32462694 0.11341075
## Concept   0.06982633 -0.12595132
## Motivation 0.11566948 0.09810277
## Read      0.69069291 -0.04174278
## Write     0.56914983 0.24433183
## Math      0.64952612 -0.04821830
## Science   1.00000000 -0.13818587
## Gender    -0.13818587 1.00000000
```



R code implementation

CCA를 실행하는 함수

```
cc1 = cc(psych, acad)
```

Canonical Correlation

```
# display the canonical correlations  
cc1$cor
```

```
## [1] 0.4640861 0.1675092 0.1039911
```

Canonical coefficients(Weights)

```
# raw canonical coefficients  
cc1[3:4]
```

```
## $xcoef  
##           [,1]      [,2]      [,3]  
## Control  -1.2538339 -0.6214776 -0.6616896  
## Concept   0.3513499 -1.1876866  0.8267210  
## Motivation -1.2624204  2.0272641  2.0002283  
##  
## $ycoef  
##           [,1]      [,2]      [,3]  
## Read      -0.044620600 -0.004910024  0.021380576  
## Write     -0.035877112  0.042071478  0.091307329  
## Math      -0.023417185  0.004229478  0.009398182  
## Science   -0.005025152 -0.085162184 -0.109835014  
## Gender    -0.632119234  1.084642326 -1.794647036
```



R code implementation

Standardized Canonical coefficients(Weights)

```
s1 = diag(sqrt(diag(cov(psych))))  
s1 %*% cc1$xcoef
```

```
##           [,1]      [,2]      [,3]  
## [1,] -0.8404196 -0.4165639 -0.4435172  
## [2,]  0.2478818 -0.8379278  0.5832620  
## [3,] -0.4326685  0.6948029  0.6855370
```

```
s2 = diag(sqrt(diag(cov(acad))))  
s2 %*% cc1$ycoef
```

```
##           [,1]      [,2]      [,3]  
## [1,] -0.45080116 -0.04960589  0.21600760  
## [2,] -0.34895712  0.40920634  0.88809662  
## [3,] -0.22046662  0.03981942  0.08848141  
## [4,] -0.04877502 -0.82659938 -1.06607828  
## [5,] -0.31503962  0.54057096 -0.89442764
```

$$\Rightarrow \mathbf{V}_{11}^{1/2}$$

$$\Rightarrow \mathbf{A}_z = \mathbf{A} \mathbf{V}_{11}^{1/2}$$

$$\Rightarrow \mathbf{V}_{22}^{1/2}$$

$$\Rightarrow \mathbf{B}_z = \mathbf{B} \mathbf{V}_{22}^{1/2}$$



R code implementation

Canonical Loadings, Cross Loadings

```
# compute canonical loadings
cc2 = comput(psych, acad, cc1)

# display canonical loadings
cc2[3:6]
```

```
## $corr.X.xscores
##           [,1]      [,2]      [,3]
## Control   -0.90404631 -0.3896883 -0.1756227
## Concept    -0.02084327 -0.7087386  0.7051632
## Motivation -0.56715106  0.3508882  0.7451289
##
## $corr.Y.xscores
##           [,1]      [,2]      [,3]
## Read      -0.3900402 -0.06010654  0.01407661
## Write     -0.4067914  0.01086075  0.02647207
## Math      -0.3545378 -0.04990916  0.01536585
## Science   -0.3055607 -0.11336980 -0.02395489
## Gender    -0.1689796  0.12645737 -0.05650916
##
## $corr.X.yscores
##           [,1]      [,2]      [,3]
## Control   -0.419555307 -0.06527635 -0.01826320
## Concept    -0.009673069 -0.11872021  0.07333073
## Motivation -0.263206910  0.05877699  0.07748681
##
## $corr.Y.yscores
##           [,1]      [,2]      [,3]
## Read      -0.8404480 -0.35882541  0.1353635
## Write     -0.8765429  0.06483674  0.2545608
## Math      -0.7639483 -0.29794884  0.1477611
## Science   -0.6584139 -0.67679761 -0.2303551
## Gender    -0.3641127  0.75492811 -0.5434036
```



R code implementation

Likelihood Ratio Test -> r=2가 적절함을 알 수 있음

```
nrow = dim(cc1$scores$xscores)[1]
p = length(psych)
q = length(acad)
rho = cc1$cor
#LRT function
CCA_LRT = function(n = nrow, p = length(psych), q = length(acad), cor = rho){
  result = -(n-1-1/2*(p+q+1))*sum(log(1-cor^2))
  return(result)
}
```

```
#rho1, rho2, rho3 are zero?
r_3 = CCA_LRT()
c(r_3, qchisq(0.95, df=p*q))
```

```
## [1] 167.58008 24.99579 ⇒ reject  $H_0 (\rho_1 = \rho_2 = \rho_3 = 0)$ 
```

```
#r = 1
r_1 = CCA_LRT(cor = cc1$cor[c(2,3)])
c(r_1, qchisq(0.95, df=(p-1)*(q-1)))
```

```
## [1] 23.38380 15.50731 ⇒ reject  $H_0 (\rho_1 \neq 0, \rho_2 = 0, \rho_3 = 0)$ 
```

```
#r=2
r_2 = CCA_LRT(cor = cc1$cor[3])
c(r_2, qchisq(0.95, df=(p-2)*(q-2)))
```

```
## [1] 6.464032 7.814728
```

⇒ do not reject $H_0 (\rho_1 \neq 0, \rho_2 \neq 0, \rho_3 = 0)$



R code implementation

다른 test (F-test) => 패키지를 통해 쉽게 할 수 있음

```
# tests of canonical dimensions
rho = ccl$cor
# Define number of observations, number of variables in first set, and number of variables in the second set.
n = dim(psych)[1]
p = length(psych)
q = length(acad)

## Calculate p-values using the F-approximations of different test statistics:
p.asym(rho, n, p, q, tstat = "Wilks")
```

Wilks' Lambda, using F-approximation (Rao's F):

##		stat	approx	df1	df2	p.value
##	1 to 3:	0.7543611	11.715733	15	1634.653	0.000000000
##	2 to 3:	0.9614300	2.944459	8	1186.000	0.002905057
##	3 to 3:	0.9891858	2.164612	3	594.000	0.091092180

⇒ canonical variable 전부 사용했을 때 유의하다고 할 수 있음

⇒ 2,3번째 canonical variable만 사용했을 때 유의하다고 할 수 있음

⇒ 3번째 canonical variable만 사용했을 때는 유의하다고 할 수 없음



R code implementation

r=2일 때의 Residual matrix

```
#R11-estimated R11 (r=2)
solve(cc1$xccoef)[,3]*%t(solve(cc1$xccoef)[,3])
```

```
##           [,1]      [,2]      [,3]
## [1,]  0.03778332 -0.02337600 -0.04964012
## [2,] -0.02337600  0.01446239  0.03071163
## [3,] -0.04964012  0.03071163  0.06521770
```

⇒ r=2 일 때, Residual Matrix는 굉장히 낮은 entry들을 가짐

r=1일 때의 Residual matrix

```
#r=1
solve(cc1$xccoef)[,2]*%t(solve(cc1$xccoef)[,2]) + solve(cc1$xccoef)[,3]*%t(solve(cc1$xccoef)[,3])
```

```
##           [,1]      [,2]      [,3]
## [1,]  0.03799956 -0.01602305 -0.05695597
## [2,] -0.01602305  0.26448636 -0.21805102
## [3,] -0.05695597 -0.21805102  0.31272540
```

⇒ r=1 일 때, Residual Matrix는 높은 entry들을 가짐 ⇒ r=1 이라고 하기 힘들

⇒ but Residual Matrix는 estimated value에 대한 평가 기준이 없음



R code implementation

r=2일 때의 설명력

```
#proportions of explained sample variance(r=2)
sum(cc1$scores$corr.X.xscores[,c(1,2)]^2)/sum(diag(matcor(psych, acad)$Xcor))
```

```
## [1] 0.6388948
```

```
sum(cc1$scores$corr.Y.yscores[,c(1,2)]^2)/sum(diag(matcor(psych, acad)$Ycor))
```

```
## [1] 0.7748177
```

r=1일 때의 설명력

```
#r=1
sum(cc1$scores$corr.X.xscores[,1]^2)/sum(diag(matcor(psych, acad)$Xcor))
```

```
## [1] 0.3797982
```

```
sum(cc1$scores$corr.Y.yscores[,1]^2)/sum(diag(matcor(psych, acad)$Ycor))
```

```
## [1] 0.5248768
```

$$= \frac{\sum_{i=1}^r \sum_{k=1}^p r_{\hat{U}_i, z_k^{(1)}}^2}{p}$$

$$= \frac{\sum_{i=1}^r \sum_{k=1}^q r_{\hat{V}_i, z_k^{(2)}}^2}{q}$$

⇒ r = 2 일 때 설명력이 좋은 것을 확인할 수 있음

$$= \frac{\sum_{i=1}^r \sum_{k=1}^p r_{\hat{U}_i, z_k^{(1)}}^2}{p}$$

$$= \frac{\sum_{i=1}^r \sum_{k=1}^q r_{\hat{V}_i, z_k^{(2)}}^2}{q}$$



END

References

Richard A. Johnson & Dean W. Wichern (2021). Applied Multivariate Statistical Analysis(6th ed).

<https://www.youtube.com/watch?v=2tUuyWTtPqM&t=1s>

<https://zephyrus1111.tistory.com/459>

<https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/>

<https://www.linkedin.com/advice/1/how-do-you-interpret-canonical-correlation>

