Canonical Correlation Analysis

Recall: The sample covariance matrix:

$$\underset{p\times p}{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{11} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix}$$

where

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_k\right)$$

The sample correlation matrix:

$$\underset{p\times p}{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

where

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)\left(x_{kj}-\bar{x}_k\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2}\sqrt{\sum_{j=1}^{n}\left(x_{kj}-\bar{x}_k\right)^2}}$$

Tests for independence and zero correlation

Test for zero correlation (Independence between two normally distributed variables):

$$H_0: \boldsymbol{\rho}_{ij} = 0 \text{ vs. } H_a: \boldsymbol{\rho}_{ij} \neq 0$$

The test statistic $$t = r_{ij}\sqrt{\frac{(n-2)}{1-r_{ij}^2}}$$

If independence is true then the test statistic t will have a t - distributions with $\nu = n - 2$ degrees of freedom.

The test is to reject ***independence*** if:

$$|t| > t_{\alpha/2}^{(n-2)}$$

Test for specific correlation (H_0: $\rho = \rho_0$ vs. H_a: $\rho \neq \rho_0$)

Large sample test statistic:

$$z = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}}$$

If H_0 is true the test statistic z will have approximately a standard normal distribution

We then ***reject*** H_0 if:

$$|z| > z_{\alpha/2}$$

Partial Correlation

Conditional Independence

There are two groups of variables:

The first group $\boldsymbol{X}^{(1)}$ has p-variate Normal distribution

The second group $\boldsymbol{X}^{(2)}$ has q-variate Normal distribution

Let $\mathrm{E}(\boldsymbol{X}^{(1)}) = \boldsymbol{\mu}^{(1)}$ and $\mathrm{E}(\boldsymbol{X}^{(2)}) = \boldsymbol{\mu}^{(2)}$

$\mathrm{cov}(\boldsymbol{X}^{(1)}) = \boldsymbol{\Sigma}_{11}$ and $\mathrm{cov}(\boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma}_{22}$

$\mathrm{cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma}_{12}$

Then the two groups have a joint covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix}$$

The matrix $\Sigma_{2\cdot 1} = \Sigma_{22} - \Sigma_{12}' \Sigma_{11}^{-1} \Sigma_{12}$

is called the ***matrix of partial variances and covariances***.

The $(i, j)^{\text{th}}$ element of the matrix $\Sigma_{2\cdot 1}$

$$\boldsymbol{\sigma}_{ij\cdot 1,\,2,\,\ldots,\,p}$$

is called the ***partial covariance*** (variance if $i = j$) between x_i and x_j given $\boldsymbol{X}^{(1)} = x_1, \ldots, x_p$. Further,

$$\rho_{ij\cdot 1,2,\ldots,p} = \frac{\sigma_{ij\cdot 1,2,\ldots,p}}{\sqrt{\sigma_{ii\cdot 1,2,\ldots,p}\sigma_{jj\cdot 1,2,\ldots,p}}}$$

is called the ***partial correlation*** between x_i and x_j given $\boldsymbol{X}^{(1)} = x_1, \ldots, x_p$

Sample partial covariance

Let: $$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}' & S_{22} \end{bmatrix}$$

denote the sample Covariance matrix

Let $S_{2\bullet 1} = S_{22} - S_{12}' S_{11}^{-1} S_{12}$

The $(i, j)^{\text{th}}$ element of the matrix $S_{2\bullet 1}$

$$s_{ij\cdot 1,\,2,\,\ldots,\,p}$$

is called the ***sample partial covariance*** (variance if $i = j$) between x_i and x_j given $x_1, \ldots, x_p$

Also,

$$r_{ij\cdot 1,2,\ldots,p} = \frac{s_{ij\cdot 1,2,\ldots,p}}{\sqrt{s_{ii\cdot 1,2,\ldots,p} s_{jj\cdot 1,2,\ldots,p}}}$$

is called the ***sample partial correlation*** between x_i and x_j given $x_1, \ldots, x_p$

Test for zero partial correlation correlation

(Conditional independence between a two variables given a set of *p* variables)

The test statistic

$$t = r_{ij.x_1,\ldots,x_p}\sqrt{\frac{(n-p-2)}{1-r^2_{ij.x_1,\ldots,x_p}}}$$

$r_{ij.x_1,\ldots,x_p}$ = the partial correlation between y_i and y_j given $x_1, \ldots, x_p$.

If independence is true then the test statistic t will have a t - distributions with $\nu = n - p - 2$ degrees of freedom.

The test is to reject ***independence*** if:

$$|t| > t^{(n-p-2)}_{\alpha/2}$$

Large Sample Test for a specific partial correlation:

$$H_0 : \rho_{ij.x_1,\ldots,x_p} = \rho^0_{ij.x_1,\ldots,x_p}$$

The test statistic

$$z = \frac{\frac{1}{2}\ln\left(\frac{1+r^0_{ij.x_1,\ldots,x_p}}{1-r^0_{ij.x_1,\ldots,x_p}}\right) - \frac{1}{2}\ln\left(\frac{1+\rho^0_{ij.x_1,\ldots,x_p}}{1-\rho^0_{ij.x_1,\ldots,x_p}}\right)}{\sqrt{\frac{1}{n-p-3}}}$$

If H_0 is true the test statistic z will have approximately a Standard Normal distribution

We then ***reject*** H_0 if: $|z| > z_{\alpha/2}$

The Multiple Correlation Coefficient

Testing independence between a single variable and a group of variables

Definition:

Suppose $\begin{bmatrix} \boldsymbol{X} \\ y \end{bmatrix}$ has (p +1)-variate Normal distribution

with covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{\boldsymbol{X}} & \Sigma_{\boldsymbol{Xy}} \\ \Sigma_{\boldsymbol{Xy}} & \sigma_y \end{bmatrix}$$

We are interested if the variable y is independent of the vector $\boldsymbol{X}$

The **multiple correlation coefficient** is the maximum correlation between y and a linear combination of the components of $\boldsymbol{X}$

It can be shown that the **multiple correlation coefficient** is:

$$\rho_{y|x_1,\ldots,x_p} = \sqrt{\frac{\Sigma_{Xy}\Sigma_X^{-1}\Sigma'_{Xy}}{\sigma_y}}$$

The sample multiple correlation coefficient:

Let the sample correlation matrix be

$$\boldsymbol{S} = \begin{bmatrix} S_{\boldsymbol{X}} & S_{\boldsymbol{Xy}} \\ S_{\boldsymbol{Xy}} & s_y \end{bmatrix}$$

Then the sample multiple correlation coefficient is

$$r_{y|x_1,\ldots,x_p} = \sqrt{\frac{S_{Xy}S_X^{-1}S'_{Xy}}{s_y}}$$

We are interested if the variable y is independent of the vector $\boldsymbol{X}$

That is,

$$H_0\colon \rho_{y|x_1,\ldots,x_p} = \sqrt{\frac{\Sigma_{Xy}\Sigma_X^{-1}\Sigma'_{Xy}}{\sigma_y}} = 0$$

The test statistic is:

$$F = \frac{n-p-1}{p}\frac{r^2_{y|x_1,\ldots,x_p}}{1-r^2_{y|x_1,\ldots,x_p}} = \frac{n-p-1}{p}\frac{S_{Xy}S_X^{-1}S'_{Xy}}{s_y - S_{Xy}S_X^{-1}S'_{Xy}}$$

If independence is true then the test statistic F will have an F-distributions with $\nu_1 = p$ degrees of freedom in the numerator and $\nu_1 = n - p + 1$ degrees of freedom in the denominator

The test is to reject ***independence*** if: $F > F_\alpha\left(p, n-p-1\right)$

Canonical Correlation Analysis

The problem

Quite often when one has collected data on several variables.

The variables are grouped into two (or more) sets of variables and the researcher is interested in whether one set of variables is independent of the other set.

In addition if it is found that the two sets of variates are dependent, it is then important to describe and understand the nature of this dependence.

The appropriate statistical procedure in this case is called *Canonical Correlation Analysis*.

Canonical Correlation: An Example

In the following study the researcher was interested in whether specific instructions on how to *relax* when taking tests and how to increase *Motivation* , would affect performance on standardized achievement tests

- Reading,
- Language and
- Mathematics

A group of 65 third- and fourth-grade students were rated after the instruction and immediately prior taking the Scholastic Achievement tests on:

- how **relaxed** they were (X_1) and
- how **motivated** they were (X_2).

In addition data was collected on the three achievement tests

- Reading (Y_1),
- Language (Y_2) and
- Mathematics (Y_3).

The data were tabulated on the next page

Case	Relaxation X_1	Motivation X_2	Reading Y_1	Language Y_2	Math Y_3	Case	Relaxation X_1	Motivation X_2	Reading Y_1	Language Y_2	Math Y_3
1	7	14	311	436	154	34	40	20	362	416	107
2	43	25	501	455	765	35	40	18	596	592	622
3	32	21	507	473	702	36	35	17	431	346	493
4	17	12	453	392	401	37	33	17	361	414	404
5	23	12	419	337	284	38	40	27	663	451	651
6	10	16	545	538	414	39	31	15	569	462	398
7	22	21	509	512	491	40	29	19	699	622	478
8	13	19	320	308	517	41	37	16	187	223	221
9	31	21	357	296	496	42	21	23	1132	839	1044
10	24	26	485	372	685	43	24	15	457	410	400
11	26	21	811	748	902	44	19	14	413	448	520
12	35	20	367	436	393	45	33	22	569	605	615
13	24	17	242	349	137	46	19	19	650	685	440
14	20	8	237	140	331	47	26	22	424	427	482
15	38	27	417	648	618	48	20	15	475	604	742
16	32	19	429	446	458	49	22	21	519	612	446
17	14	11	555	579	438	50	37	22	338	463	327
18	24	12	599	497	414	51	41	28	674	613	534
19	38	25	403	383	606	52	29	35	381	624	565
20	30	8	550	324	674	53	25	12	199	171	316
21	22	25	377	496	242	54	27	21	577	523	699
22	36	28	671	585	710	55	22	20	425	466	402
23	3	22	498	488	481	56	4	11	392	192	354
24	44	28	477	583	260	57	27	22	401	520	558
25	24	25	609	413	670	58	28	23	321	410	460
26	33	18	521	522	716	59	33	20	682	433	743
27	24	21	495	645	491	60	33	24	719	727	1052
28	28	20	400	555	624	61	31	33	672	705	650
29	34	7	258	175	276	62	20	11	366	309	537
30	39	20	466	541	348	63	26	25	581	558	386
31	7	19	709	757	589	64	23	10	681	530	581
32	13	17	586	472	492	65	30	22	1019	917	880
33	32	18	418	361	428						

Definition: (Canonical variates and Canonical correlations)

Let $\boldsymbol{X}^{(1)}$ have p-variate Normal distribution

and $\boldsymbol{X}^{(2)}$ have q-variate Normal distribution

Let $\mathrm{E}(\boldsymbol{X}^{(1)}) = \boldsymbol{\mu}^{(1)}$ and $\mathrm{E}(\boldsymbol{X}^{(2)}) = \boldsymbol{\mu}^{(2)}$

$\mathrm{cov}(\boldsymbol{X}^{(1)}) = \boldsymbol{\Sigma}_{11}$ and $\mathrm{cov}(\boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma}_{22}$

$\mathrm{cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma}_{12}$

Then the two groups have a joint covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix}$$

derivation: (1st pair of Canonical variates and Canonical correlation)

Let $U = \boldsymbol{a}'\boldsymbol{X}^{(1)}$

and $V = \boldsymbol{b}'\boldsymbol{X}^{(2)}$

We want to find U and V such that they achieve the maximum correlation $\phi_1 = \text{corr}(U, V)$

Then the solution U_1 and V_1 are called the first pair of ***canonical variates*** and ϕ_1 is called the first ***canonical correlation coefficient***.

derivation: (1st pair of Canonical variates and Canonical correlation)

We have that:

$\text{Var}(U) = \boldsymbol{a}'\text{cov}(\boldsymbol{X}^{(1)})\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}$

$\text{Var}(V) = \boldsymbol{b}'\text{cov}(\boldsymbol{X}^{(2)})\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}$

$\text{Cov}(U, V) = \boldsymbol{a}'\text{cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})\boldsymbol{b} = \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$

Hence:

$$\text{Corr}(U, V) = \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}}}$$

Thus we want to choose b and $\boldsymbol{b}$

so that

$$\rho_{UV} = \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}}} \text{ is at a maximum}$$

Solution:

The first pair of ***canonical variates***

$U_1 = \boldsymbol{a}_1'\boldsymbol{X}^{(1)}$ and $V_1 = \boldsymbol{b}_1'\boldsymbol{X}^{(2)}$

are found by finding $\vec{a}_1$ and $\vec{b}_1$, eigenvectors of the matrices

$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ and $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

associated with the largest eigenvalue (same for both matrices)

The largest eigenvalue of the two matrices is the square of the first canonical correlation coefficient ϕ_1

$$\phi_1 = \sqrt{\text{the largest eigenvalue of } \Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}}$$

The remaining canonical variates and canonical correlation coefficients

The second pair of ***canonical variates***

$U_2 = \boldsymbol{a}_2'\boldsymbol{X}^{(1)}$

and $V_2 = \boldsymbol{b}_2'\boldsymbol{X}^{(2)}$

are found by finding $\vec{a}_2$ and $\vec{b}_2$, so that

1. (U_2, V_2) are independent of (U_1, V_1).

2. The correlation between U_2 and V_2 is maximized

The correlation, ϕ_2, between U_2 and V_2 is called the ***second canonical correlation coefficient.***

The i^{th} pair of ***canonical variates***

$U_i = \boldsymbol{a}_i'\boldsymbol{X}^{(1)}$

and $V_i = \boldsymbol{b}_i'\boldsymbol{X}^{(2)}$

are found by finding $\vec{a}_i$ and $\vec{b}_i$, so that

1. (U_i,V_i) are independent of (U_1,V_1), …, (U_{i-1},V_{i-1}).
2. The correlation between U_i and V_i is maximized

The correlation, ϕ_i, between U_i and V_i is called the ***i^{th} canonical correlation coefficient.***

It can be shown that $\boldsymbol{a}_2$ and $\boldsymbol{b}_2$

are eigenvectors of the matrices

$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ and $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

associated with the 2^{nd} largest eigenvalue (same for both matrices)

The 2^{nd} largest eigenvalue of the two matrices is the square of the 2^{nd} canonical correlation coefficient ϕ_2

$$\phi_2 = \sqrt{\text{the } 2^{nd} \text{ largest eigenvalue of } \Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}}$$

$$= \sqrt{\text{the } 2^{nd} \text{ largest eigenvalue of } \Sigma_{22}^{-1}\Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12}}$$

continuing

Coefficients for the i^{th} pair of canonical variates, $\vec{a}_i$ and $\vec{b}_i$ are eigenvectors of the matrices

$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ and $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

associated with the i^{th} largest eigenvalue (same for both matrices)

The i^{th} largest eigenvalue of the two matrices is the square of the i^{th} canonical correlation coefficient ϕ_i

$$\phi_i = \sqrt{\text{the } i^{th} \text{ largest eigenvalue of } \Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}}$$

$$= \sqrt{\text{the } i^{th} \text{largest eigenvalue of } \Sigma_{22}^{-1}\Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12}}$$

Example

Variables

- **relaxation Score** (X_1)
- **motivation score** (X_2).
  - **Reading** (Y_1),
  - **Language** (Y_2) and
  - **Mathematics** (Y_3).

Summary Statistics

UNIVARIATE SUMMARY STATISTICS

VARIABLE	MEAN	STANDARD DEVIATION
1 Relax	26.87692	9.50412
2 Mot	19.41538	5.83066
3 Read	499.03077	172.25508
4 Lang	485.83077	156.08957
5 Math	512.52308	195.18614

CORRELATIONS

		Relax	Mot	Read	Lang	Math
		1	2	3	4	5
Relax	1	1.000				
Mot	2	0.391	1.000			
Read	3	0.002	0.280	1.000		
Lang	4	0.050	0.510	0.781	1.000	
Math	5	0.127	0.340	0.713	0.556	1.000

Canonical Correlation Statistics

EIGENVALUE	CANONICAL CORRELATION	NUMBER OF EIGENVALUES	BARTLETT'S TEST FOR REMAINING EIGENVALUES		
			CHI-SQUARE	D.F.	TAIL PROB.
			27.86	6	0.0001
0.35029	0.59186	1	1.56	2	0.4586
0.02523	0.15885				

BARTLETT'S TEST ABOVE INDICATES THE NUMBER OF CANONICAL VARIABLES NECESSARY TO EXPRESS THE DEPENDENCY BETWEEN THE TWO SETS OF VARIABLES. THE NECESSARY NUMBER OF CANONICAL VARIABLES IS THE SMALLEST NUMBER OF EIGENVALUES SUCH THAT THE TEST OF THE REMAINING EIGENVALUES IS NON-SIGNIFICANT. FOR EXAMPLE, IF A TEST AT THE .01 LEVEL WERE DESIRED, THEN 1 VARIABLES WOULD BE CONSIDERED NECESSARY. HOWEVER, THE NUMBER OF CANONICAL VARIABLES OF PRACTICAL VALUE IS LIKELY TO BE SMALLER.