High Dimensional Data Analysis

Canonical Correlation Analysis

Canonical Correlation Analysis

- Introduction
- Principle of Canonical Correlation Analysis
- Sample Canonical Correlation

Introduction

Canonical Correlation Analysis (CCA) is a multivariate statistical analysis method used to analyze the correlation between two sets of random variables. It effectively reveals the mutual linear dependence between two sets of random variables. This method was first proposed by Hotelling (1935).
In practice, it is often necessary to study the correlation between one set of variables and another set of variables.

Introduction

For example:

- In business and economic research, consider the correlation between one set of price indices and another set of price indices.
- In sports training, examine the correlation between various physical indicators of athletes and various training items.
- In factories, examine the correlation between the main quality indicators of raw materials and the quality indicators of products.
- In pedagogy, study the correlation between the scores of various subjects in the college entrance examination for senior high school students and the scores of main subjects in the second year of high school, etc.

Introduction

We consider using the idea of principal component analysis to transform the study of the correlation between two random vectors $x = (x_1, x_2, \cdots, x_p)^t$ and $y = (y_1, y_2, \cdots, y_q)^t$ into the study of the correlation between two composite variables u and v. That is, find coefficient vectors $a = (a_1, a_2, \cdots, a_p)^t$ and $b = (b_1, b_2, \cdots, b_q)^t$, such that

$$u = a^t x = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p,$$

$$v = b^t y = b_1 y_1 + b_2 y_2 + \cdots + b_q y_q$$

have the maximum correlation between them;

Introduction

$$u = a^t x = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p,$$

$$v = b^t y = b_1 y_1 + b_2 y_2 + \cdots + b_q y_q$$

if these two composite variables are insufficient to represent the correlation between the two original sets of variables, we can continue to find the second linear combination in each group, so that they have the maximum correlation with each other among the linear combinations that are uncorrelated with the first linear combination, and so on. This is the basic idea of canonical correlation analysis.

Principle of Canonical Correlation Analysis

We want to find coefficient vectors a and b, such that the correlation coefficient between $u = a'x$ and $v = b'y$ reaches the maximum. By the definition of correlation coefficient,

$$\rho(u, v) = \frac{\text{cov}(a'x, b'y)}{\sqrt{\text{var}(a'x)\text{var}(b'y)}} = \frac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}ab'\Sigma_{22}b}}$$

To obtain a unique solution, we impose the following conditions on a and b

$$\text{var}(a'x) = a'\Sigma_{11}a = 1, \quad \text{var}(b'y) = b'\Sigma_{22}b = 1.$$

Principle of Canonical Correlation Analysis

We want to find a and b under the above conditions, such that $\rho(u, v)$ reaches the maximum. From the method of conditional extremes in conditional differential calculus, we know that a and b satisfy

$$M_1 a = \lambda^2 a, \quad M_2 b = \lambda^2 b,$$

where

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \quad M_2 = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Principle of Canonical Correlation Analysis

Matrices M_1 and M_2 have common nonzero eigenvalues, denoted as $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_m^2 > 0$. We call $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ the canonical correlation coefficients. Let a_i and b_i be the eigenvectors of matrices M_1 and M_2 corresponding to λ_i^2, respectively, satisfying $a_i' \Sigma_{11} a_i = 1$ and $b_i' \Sigma_{22} b_i = 1$. Let

$$(u_i, v_i) = (a_i' x, b_i' y), i = 1, 2, \cdots, m,$$

then (u_i, v_i) is called the i-th pair of canonical correlation variables, and $\rho(u_i, v_i) = a_i' \Sigma_{12} b_i = \lambda_i$ is called the i-th canonical correlation coefficient.

Sample Canonical Correlation

Assume the data matrix composed of n observations of the $p+q$-dimensional random vector $z = (x_1, \cdots, x_p, y_1, \cdots, y_q)$ is

$$Z = (X, Y) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_{11} & y_{12} & \cdots & y_{1q} \\ x_{21} & x_{22} & \cdots & x_{2p} & y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} z'_{(1)} \\ z'_{(2)} \\ \vdots \\ z'_{(n)} \end{pmatrix}$$

Sample Canonical Correlation

If we assume $z \sim N_{p+q}(\mu, \Sigma)$, then the unbiased estimate of the covariance matrix Σ is

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (z_{(i)} - \bar{z})(z_{(i)} - \bar{z})',$$

where $\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_{(i)}$. The matrix S is called the sample covariance matrix.

Sample Canonical Correlation

Let $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, where Σ_{11} is a $p \times p$ matrix and Σ_{22} is a $q \times q$ matrix. Partition the sample covariance matrix S similarly as $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$. Below we discuss how to obtain the sample canonical correlation variables starting from the sample covariance matrix or sample correlation matrix.

Sample Canonical Correlation

Similar to the method for obtaining population canonical correlation variables and canonical correlation coefficients, the coefficient vectors $\hat{a}_i, \hat{b}_i$ of the sample canonical correlation variables and the sample canonical correlation coefficients $\hat{\lambda}_i (i = 1, 2, \cdots, m)$ can be directly obtained by solving the following equation systems:

$$\hat{M}_1 \hat{a}_i = \hat{\lambda}_i^2 \hat{a}_i, \quad \hat{M}_2 \hat{b}_i = \hat{\lambda}_i^2 \hat{b}_i,$$

where

$$\hat{M}_1 = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}, \quad \hat{M}_2 = S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}$$

Sample Canonical Correlation

In practice, data is often first centered and standardized. Below we provide the sample canonical correlation variables starting from the sample correlation matrix. Assume all components of the original random vectors x and y are standardized.
Let the sample correlation matrix be $R = (r_{ij})_{p\times p}$, where $r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$.
Similar to the partitioning of S, partition R as

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

Let

$$\tilde{M}_1 = R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}, \quad \tilde{M}_2 = R_{22}^{-1}R_{21}R_{11}^{-1}R_{12},$$

Sample Canonical Correlation

It can be proved that,
$\tilde{M}_1, \tilde{M}_2$ and $\tilde{M}_1, \tilde{M}_2$ have the same nonzero eigenvalues, namely $\hat{x}_1^2 \geq \hat{x}_2^2 \cdots \geq \hat{x}_m^2 > 0$. Let a_i and b_i be the eigenvectors of $\tilde{M}_1$ and $\tilde{M}_2$ corresponding to the eigenvalue $\hat{x}_i^2$, respectively, and satisfy $a_i' R_{11} a_i = 1$, $b_i' R_{22} b_i = 1$. Then $u_i = a_i' x, v_i = b_i' y$ are the i-th pair of canonical correlation variables of x and y.

Example 1

A rehabilitation club measured three physiological indicators: weight (x1), waist circumference (x2), pulse (x3) and three training indicators: pull-ups (y1), sit-ups (y2), jump counts (y3) for 20 middle-aged people. The measurement data are listed in Table 1. Calculate the canonical correlation coefficients and canonical correlation variables starting from the sample correlation matrix.

Table 1

No.	x1	x2	x3	y1	y2	y3
1	191	36	50	5	162	60
2	189	37	52	2	110	60
3	193	38	58	12	101	101
4	162	35	62	12	105	37
5	189	35	46	13	155	58
6	182	36	56	4	101	42
7	211	38	56	8	101	38
8	167	34	60	6	125	40
9	176	31	74	15	200	40
10	154	33	56	17	251	250
11	169	34	50	17	120	38
12	166	33	52	13	210	115
13	154	34	64	14	215	105
14	247	46	50	1	50	50
15	193	36	46	6	70	31
16	202	37	62	12	210	120
17	176	37	54	4	60	25
18	157	32	42	11	230	80
19	156	33	54	15	225	73
20	138	33	68	2	110	43

Example 1

Solution: In this example $p = q = 3, n = 20$. After calculation, the sample correlation matrices are:

$$R_{11} = \begin{pmatrix} 1.000 & 0.870 & -0.366 \\ 0.870 & 1.000 & -0.353 \\ -0.366 & -0.353 & 1.000 \end{pmatrix}, R_{22} = \begin{pmatrix} 1.000 & 0.696 & 0.496 \\ 0.696 & 1.000 & 0.669 \\ 0.496 & 0.669 & 1.000 \end{pmatrix},$$

$$R_{12} = R_{21}^{-1} = \begin{pmatrix} -0.390 & -0.493 & -0.226 \\ -0.552 & -0.646 & -0.192 \\ 0.151 & 0.225 & 0.035 \end{pmatrix}.$$

The eigenvalues of $M_1 = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21}$ are
$\lambda_1^2 = 0.633, \lambda_2^2 = 0.0402, \lambda_3^2 = 0.0053$, thus the three sample canonical correlation coefficients are:

$$\lambda_1 = 0.79561, \quad \lambda_2 = 0.20056, \quad \lambda_3 = 0.07257.$$

Example 1

The normalized coefficient vectors of the sample canonical correlation variables are:

$$a_1 = \begin{pmatrix} -0.775 \\ 1.579 \\ -0.059 \end{pmatrix}, \quad a_2 = \begin{pmatrix} -1.844 \\ 1.181 \\ -0.231 \end{pmatrix}, \quad a_3 = \begin{pmatrix} 0.191 \\ -0.506 \\ -1.051 \end{pmatrix},$$

$$b_1 = \begin{pmatrix} -0.350 \\ -1.054 \\ 0.716 \end{pmatrix}, \quad b_2 = \begin{pmatrix} -0.376 \\ 0.124 \\ 1.062 \end{pmatrix}, \quad b_3 = \begin{pmatrix} 1.297 \\ -1.237 \\ 0.419 \end{pmatrix}.$$

Therefore, the first pair of sample canonical correlation variables are:

$$u_1 = a_1' x^* = -0.775 x_1^* + 1.579 x_2^* - 0.059 x_3^*,$$

$$v_1 = b_1' y^* = -0.350 y_1^* - 1.054 y_2^* + 0.716 y_3^*.$$

Example 1

The second pair of sample canonical correlation variables are:

$$u_2 = a_2' x^* = -1.844 x_1^* + 1.181 x_2^* - 0.231 x_3^*,$$

$$v_2 = b_2' y^* = -0.376 y_1^* + 0.124 y_2^* + 1.062 y_3^*.$$

Correlation coefficients between canonical correlation variables and original variables:

$$\rho(u_1, x_1) = 0.6206, \quad \rho(u_1, x_2) = 0.9254, \quad \rho(u_1, x_3) = -0.3328,$$

$$\rho(v_1, y_1) = -0.7276, \quad \rho(v_1, y_2) = -0.8177, \quad \rho(v_1, y_3) = -0.1622.$$

Example 1

Next, calculate the scores of sample data under canonical correlation variables. Let

$$A = (a_1, a_2, a_3), \quad B = (b_1, b_2, b_3)$$

From

$$U = (U_1, U_2, U_3) = XA, \quad V = (V_1, V_2, V_3) = YB$$

Calculate the score matrices.

Example 1

Following figure shows the score plots of the first two pairs of canonical correlation variables on the data matrix. The left plot shows the scatter plot of the scores of the first pair of canonical variables (U_1, V_1), and the right plot shows the scatter plot of the scores of the second pair of canonical variables (U_2, V_2). Clearly, the points in the left plot are almost along a straight line, while the points in the right plot are more scattered. This is because the correlation of the first pair of canonical correlation variables is large, with the first canonical correlation coefficient being 0.79561, so the scatter plot shows a significant linear relationship. The second canonical correlation coefficient is small, so the points in the scatter plot do not show an obvious linear correlation.

Example 1

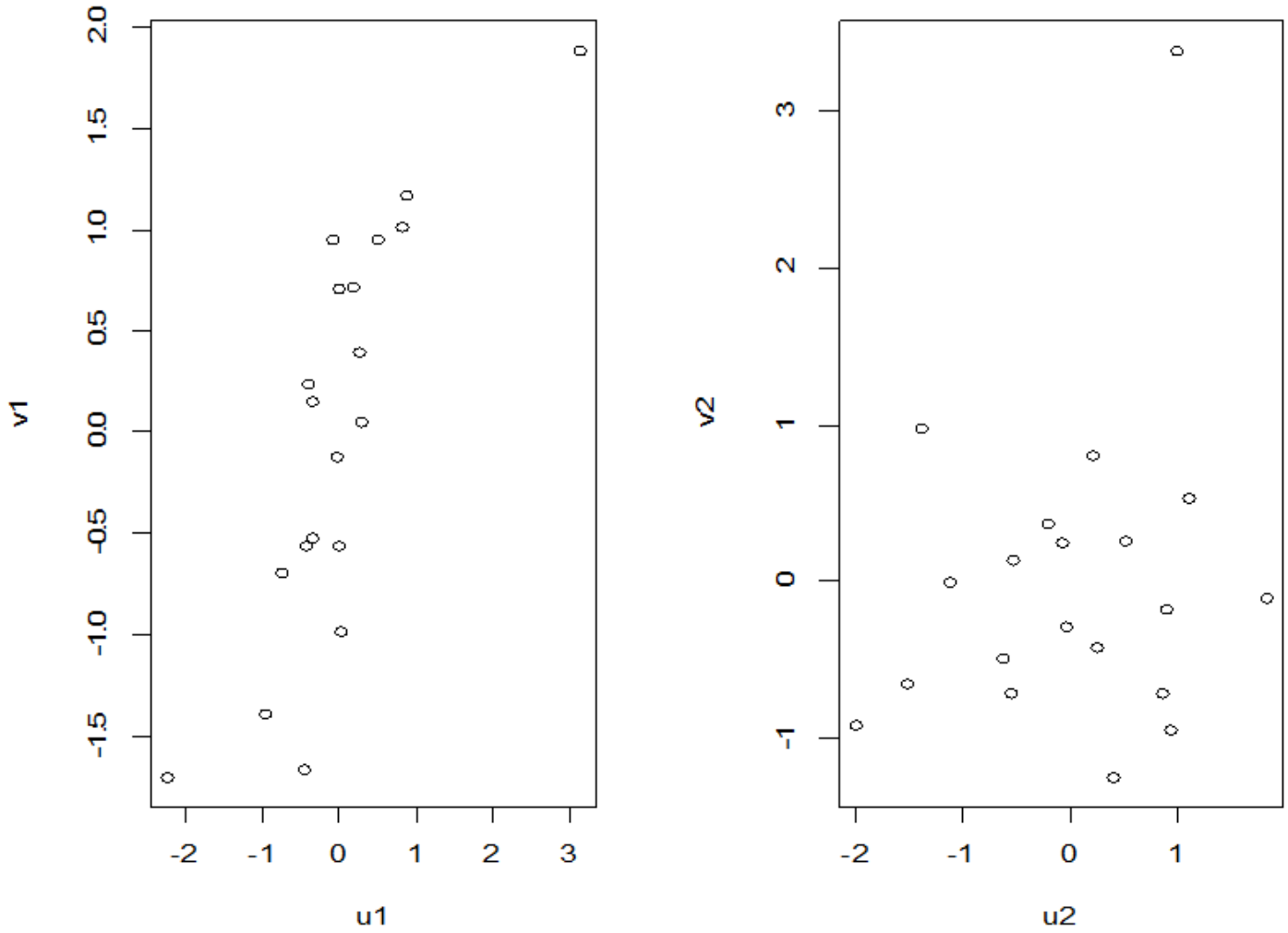


Figure: Score plot of sample data for canonical correlation variables

Sample Canonical Correlation

Significance Test for Canonical Correlation Coefficients

Now assume the first k canonical correlation coefficients are significant, i.e., the first k pairs of canonical correlation variables are significantly correlated. To test whether the (k+1)-th canonical correlation coefficient is significant, test the following hypothesis:

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_m = 0,$$

$$H_1 : \text{At least one of } \lambda_{k+1}, \lambda_{k+2}, \cdots, \lambda_m \neq 0.$$

Sample Canonical Correlation

The likelihood ratio test statistic is

$$\Lambda_{k+1} = \prod_{i=k+1}^{m} (1 - \hat{\lambda}_i^2).$$

According to Box's result, for sufficiently large n, when the null hypothesis H_0 holds, the statistic

$$Q_{k+1} = \left[n - k - \frac{1}{2}(p + q + 3) + \sum_{i=1}^{k} \lambda_i^{-2} \right] \ln \Lambda_{k+1}$$

approximately follows a χ^2 distribution with degrees of freedom $f_{k+1} = (p - k)(q - k)$.

Sample Canonical Correlation

At a given significance level α, if $Q_{k+1} \geq \chi^2_\alpha(f_{k+1})$, then reject the null hypothesis H_0, meaning that the $(k+1)$-th pair of canonical variables is significantly correlated; otherwise, consider the $(k+1)$-th pair of canonical variables uncorrelated, and only keep the first k pairs of canonical correlation variables.

Example 2

To study the relationship between children's morphology and pulmonary ventilation function, morphological indicators: height (x1), weight (x2), chest circumference (x3) and pulmonary ventilation function indicators: vital capacity (y1), resting ventilation volume (y2), maximum ventilation per minute (y3) were measured for 40 healthy children aged 8-12 in a primary school. The data are shown in the table. Analyze the correlation between children's morphological indicators and pulmonary ventilation indicators, and determine the number of pairs of canonical correlation variables.

Example 2 (Part 1)

No.	x1	x2	x3	y1	y2	y3
1	140.6	43.7	77.9	2.67	7.00	108.0
2	135.7	39.5	63.9	2.08	6.98	91.7
3	140.2	45.0	75.0	2.62	6.17	101.8
4	152.1	52.3	88.1	2.89	10.42	112.5
5	132.2	36.7	62.4	2.14	7.47	97.5
6	147.1	45.2	78.9	2.96	9.23	92.4
7	147.5	47.4	76.2	3.14	8.78	95.4
8	130.6	38.4	61.8	2.03	5.31	77.2
9	154.9	48.2	87.2	2.91	10.69	80.8
10	142.4	42.6	74.1	2.33	11.15	76.7
11	136.5	38.4	66.6	1.99	7.77	49.9
12	162.0	45.7	93.6	3.29	3.35	58.0
13	148.9	42.4	80.6	2.74	10.11	82.4
14	136.3	38.1	68.3	2.44	7.82	76.5
15	159.5	49.1	87.7	2.98	11.77	88.1
16	165.9	55.7	93.5	3.17	13.14	110.3
17	134.5	41.6	61.9	2.25	8.75	73.1
18	152.5	53.4	83.2	2.96	6.60	71.5
19	138.2	35.5	66.1	2.13	6.62	105.4
20	144.2	42.0	76.2	2.52	5.59	82.0

Example 2 (Part 2)

No.	x1	x2	x3	y1	y2	y3
21	128.1	37.3	57.0	1.92	5.81	92.7
22	127.5	32.0	57.9	2.02	6.42	73.2
23	140.7	44.7	73.7	2.64	8.00	89.1
24	150.4	49.7	82.4	2.97	9.09	61.8
25	151.5	45.5	81.3	2.71	10.20	98.9
26	151.3	47.2	84.3	2.92	6.16	83.7
27	150.2	48.1	85.8	2.79	9.50	84.0
28	139.4	33.6	67.0	2.27	8.92	71.0
29	150.6	45.6	84.9	2.86	12.03	123.4
30	140.6	46.7	67.9	2.67	7.00	108.0
31	135.7	47.5	57.9	2.38	6.98	91.7
32	140.2	48.0	71.0	2.62	6.17	101.8
33	152.1	50.3	88.1	2.89	10.42	112.5
34	132.2	43.7	62.4	2.14	7.47	97.5
35	147.1	41.2	78.9	2.66	9.23	92.4
36	147.5	45.4	76.2	2.75	8.78	95.4
37	130.6	38.4	65.8	2.13	5.31	77.2
38	154.9	48.2	91.2	2.91	10.69	80.8
39	142.4	42.6	83.1	2.63	11.15	76.7
40	136.5	40.4	69.6	2.01	7.77	49.9

Example 2

Solution: After calculation, the sample correlation matrices are:

$$R_{11} = \begin{pmatrix} 1.000 & 0.807 & 0.947 \\ 0.807 & 1.000 & 0.762 \\ 0.947 & 0.762 & 1.000 \end{pmatrix}, \quad R_{22} = \begin{pmatrix} 1.000 & 0.395 & 0.248 \\ 0.395 & 1.000 & 0.270 \\ 0.248 & 0.270 & 1.000 \end{pmatrix},$$

$$R_{12} = R_{21}' = \begin{pmatrix} 0.911 & 0.523 & 0.136 \\ 0.833 & 0.243 & 0.199 \\ 0.881 & 0.504 & 0.097 \end{pmatrix}.$$

The three sample canonical correlation coefficients are:

$$\lambda_1 = 0.9423286, \quad \lambda_2 = 0.4754312, \quad \lambda_3 = 0.1041348.$$

Example 2

The normalized coefficient vectors of the canonical correlation variables are:

$$a_1 = \begin{pmatrix} -0.6559 \\ -0.1335 \\ -0.2429 \end{pmatrix}, a_2 = \begin{pmatrix} 1.2033 \\ -1.6873 \\ 0.2448 \end{pmatrix}, a_3 = \begin{pmatrix} 3.1544 \\ -0.1606 \\ -3.1047 \end{pmatrix},$$

$$b_1 = \begin{pmatrix} -0.9361 \\ -0.1958 \\ 0.1370 \end{pmatrix}, b_2 = \begin{pmatrix} -0.4922 \\ 1.0450 \\ -0.4733 \end{pmatrix}, b_3 = \begin{pmatrix} -0.3118 \\ 0.3174 \\ 0.9300 \end{pmatrix}$$

Therefore, the first pair of sample canonical correlation variables are:

$$u_1 = a_1' x^* = -0.6559 x_1^* - 0.1335 x_2^* - 0.2429 x_3^*,$$

$$v_1 = b_1' y^* = -0.9361 y_1^* - 0.1958 y_2^* + 0.1370 y_3^*.$$

Example 2

The second pair of sample canonical correlation variables are:

$$u_2 = a_2' x^* = 1.2033 x_1^* - 1.6873 x_2^* + 0.2448 x_3^*,$$

$$v_2 = b_2' y^* = -0.4922 y_1^* + 1.0450 y_2^* - 0.4733 y_3^*.$$

Correlation coefficients between the first pair of canonical correlation variables and original variables:

$$\rho(u_1, x_1) = -0.9940, \quad \rho(u_1, x_2) = -0.8490, \quad \rho(u_1, x_3) = -0.9660,$$

$$\rho(v_1, y_1) = -0.9794, \quad \rho(v_1, y_2) = -0.5284, \quad \rho(v_1, y_3) = -0.1481.$$

Let

$$A = (a_1, a_2, a_3), \quad B = (b_1, b_2, b_3)$$

$$U = (U_1, U_2, U_3) = XA, \quad V = (V_1, V_2, V_3) = YB$$

Example 2

Following figure shows the score plots of the first two pairs of canonical correlation variables on the data matrix. The left plot shows the scatter plot of the scores of the first pair of canonical variables, and the right plot shows the scatter plot of the scores of the second pair of canonical variables. Clearly, the points in the left plot are basically along a straight line, while the points in the right plot are more scattered but still show some linear trend. This is because the first canonical correlation coefficient is 0.9423286, close to 1, so the correlation of the first pair of canonical correlation variables is large, and the points in its scatter plot are almost on a straight line; while the second canonical correlation coefficient is 0.4754312, which is not too small either.

Example 2

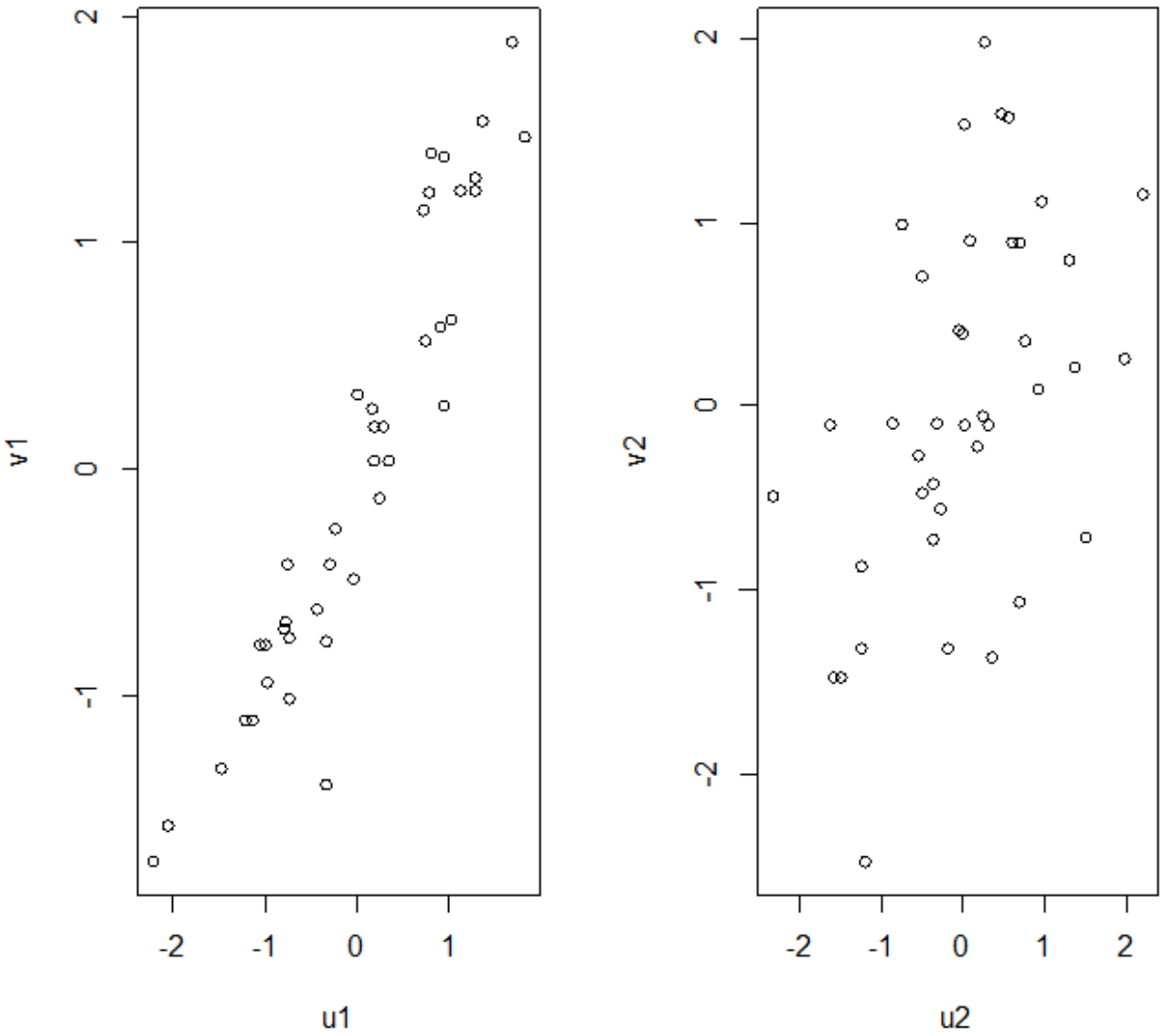


Figure: Score plot of sample data for canonical correlation variables

Example 2

The significance test results for the canonical correlation coefficients show that the first two canonical correlation coefficients are significant, meaning two pairs of canonical correlation variables should be retained, which is consistent with the previous analysis results. In fact, since $Q_1 = 87.19645 > 16.919 = \chi^2_{0.05}(9)$, we reject the hypothesis at significance level $\alpha = 0.05$:

$$H_0 : \lambda_1 = \lambda_2 = \lambda_3 = 0,$$

meaning at least the first pair of correlation variables is significantly correlated.

Example 2

Furthermore, since

$$Q_2 = 9.517 > 9.4877 = \chi^2_{0.05}(4),$$

we reject the hypothesis at significance level $\alpha = 0.05$:

$$H_0 : \lambda_2 = \lambda_3 = 0,$$

meaning the second pair of correlation variables is significantly correlated. Additionally,

$$Q_3 = 0.437 < 2.7055 = \chi^2_{0.1}(1),$$

so we accept the hypothesis at significance level $\alpha = 0.1$:

$$H_0 : \lambda_3 = 0,$$

meaning the third pair of correlation variables is not significantly correlated.