

Canonical Correlation Analysis

Goal:

Identify and quantify the association between two sets of variables.

Examples

- | | |
|-----------------------------|--------------------------|
| 1. aptitude variables | achievement variables |
| 2. personality variables | ability measures |
| 3. price indices | production indices |
| 4. psychological attributes | physiological attributes |

Ideas

Consider the correlations between the linear combinations of the variables in the first set and the linear combinations of the variables in the second set.

Suppose there are p variables in the first set, denoted by $X^{(1)}$; q variables in the second set denoted by $X^{(2)}$. And we assume

$$\text{Cov} \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

The covariance between $X^{(1)}$ and $X^{(2)}$ is represented by Σ_{12} which involves $p \times q$ parameters. Canonical correlation analysis is to summarize the associations between the $X^{(1)}$ and $X^{(2)}$ sets in terms of a few carefully chosen covariances (correlations) rather than the pq covariances in Σ_{12} .

Definition

Let $U = a'X^{(1)}$ and $V = b'X^{(2)}$, the canonical variables and canonical correlations are defined as follows:

First pair: $U_1 = a'_1 X^{(1)}$, $V_1 = b'_1 X^{(2)}$

$$\text{Var}(U_1) = a'_1 \Sigma_{11} a_1 = 1$$

$$\text{Var}(V_1) = b'_1 \Sigma_{22} b_1 = 1$$

$$\text{Corr}(U_1, V_1) = \max \text{Corr}(U, V) = \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a} \sqrt{b' \Sigma_{22} b}}$$

Second pair: $U_2 = a'_2 X^{(1)}$, $V_2 = b'_2 X^{(2)}$

$$\text{Corr}(U_2, V_2) = \max \text{Corr}(U, V)$$

$$\text{Var}(U_2) = \text{Var}(V_2) = 1$$

$$\text{Cov}(U_2, U_1) = \text{Cov}(U_2, V_1) = 0$$

$$\text{Cov}(V_2, U_1) = \text{Cov}(V_2, V_1) = 0$$

k th pair: U_k and V_k having unit variances which maximize the correlation among all choices uncorrelated with the previous $k - 1$ canonical variable pairs.

$$\begin{aligned} \max_{a,b} \text{Corr}(a' X^{(1)}, b' X^{(2)}) &= \max_{a,b} \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a} \sqrt{b' \Sigma_{22} b}} \\ &= \max_{\substack{\|\hat{a}\| = 1 \\ \|\hat{b}\| = 1 \\ \hat{b} \propto \Sigma_{21} \hat{a}}} (\hat{a}' \tilde{\Sigma}_{12} \tilde{\Sigma}_{21} \hat{a})^{1/2} = \max_{\substack{\|\hat{b}\| = 1 \\ \|\hat{a}\| = 1 \\ \hat{a} \propto \Sigma_{12} \hat{b}}} (\hat{b}' \tilde{\Sigma}_{21} \tilde{\Sigma}_{12} \hat{b})^{1/2} \end{aligned} \quad (1)$$

Where $a = \Sigma_{11}^{-1/2} \hat{a}$, $b = \Sigma_{22}^{-1/2} \hat{b}$, $\tilde{\Sigma}_{12} = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$, and $\tilde{\Sigma}_{21} = \tilde{\Sigma}'_{12}$. Based on (1), the canonical correlations and variables can be calculated as follows.

$$\begin{array}{ll} \tilde{\Sigma}_{12} \tilde{\Sigma}_{21} = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2} & \tilde{\Sigma}_{21} \tilde{\Sigma}_{12} = \Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1/2} \\ (e_1, \rho_1^2), (e_2, \rho_2^2), \dots, (e_p, \rho_p^2) & (f_1, \rho_1^2), (f_2, \rho_2^2), \dots, (f_p, \rho_p^2), (f_{p+1}, 0), \dots, (f_q, 0) \\ \hat{a}_1 = e_1, a_1 = \Sigma_{11}^{-1/2} e_1, U_1 = a'_1 X^{(1)} & \rho_1 \quad V_1 = b'_1 X^{(2)}, b_1 = \Sigma_{22}^{-1/2} \hat{b}_1, \hat{b}_1 = f_1 \\ \hat{a}_2 = e_2, a_2 = \Sigma_{11}^{-1/2} e_2, U_2 = a'_2 X^{(1)} & \rho_2 \quad V_2 = b'_2 X^{(2)}, b_2 = \Sigma_{22}^{-1/2} \hat{b}_2, \hat{b}_2 = f_2 \\ \vdots & \vdots \\ \hat{a}_p = e_p, a_p = \Sigma_{11}^{-1/2} e_p, U_p = a'_p X^{(1)} & \rho_p \quad V_p = b'_p X^{(2)}, b_p = \Sigma_{22}^{-1/2} \hat{b}_p, \hat{b}_p = f_p \\ \text{none} & \text{na} \quad V_{p+1} = (\Sigma_{22}^{-1/2} f_{p+1}) X^{(2)}, \dots, V_q = (\Sigma_{22}^{-1/2} f_q)' X^{(2)} \end{array}$$

Canonical correlations and variable for standardized variables

Let $Z^{(1)'} = (Z_1^{(1)}, Z_2^{(1)}, \dots, Z_p^{(1)})$ and $Z^{(2)'} = (Z_1^{(2)}, Z_2^{(2)}, \dots, Z_q^{(2)})$, and

$$\text{Cov}(Z^{(1)}, Z^{(2)}) = \rho = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix}$$

Denote the pairs of canonical variables are

$$(U_1^z, V_1^z), (U_2^z, V_2^z), \dots, (U_p^z, V_p^z),$$

which can be calculated as follows.

$$\begin{array}{ll} \begin{array}{l} \rho_{11}^{-1/2} \rho_{12} \rho_{22}^{-1} \rho_{21} \rho_{11}^{-1/2} \\ (e_1^z, \rho_1^2), (e_2^z, \rho_2^2), \dots, (e_p^z, \rho_p^2) \\ a_1^z = \rho_{11}^{-1/2} e_1^z, U_1^z = a_1^{z'} Z^{(1)} \\ \vdots \\ a_p^z = \rho_{11}^{-1/2} e_p^z, U_p^z = a_p^{z'} Z^{(1)} \\ \text{none} \end{array} & \begin{array}{l} \rho_{22}^{-1/2} \rho_{21} \rho_{11}^{-1} \rho_{12} \rho_{22}^{-1/2} \\ (f_1^z, \rho_1^2), (f_2^z, \rho_2^2), \dots, (f_p^z, \rho_p^2), (f_{p+1}^z, 0), \dots, (f_q^z, 0) \\ \rho_1 \quad V_1^z = b_1^{z'} Z^{(2)}, b_1^z = \rho_{22}^{-1/2} f_1^z \\ \vdots \\ \rho_p \quad V_p^z = b_p^{z'} Z^{(2)}, b_p^z = \rho_{22}^{-1/2} f_p^z \\ \text{na} \quad V_{p+1}^z = (\rho_{22}^{-1/2} f_{p+1}^z) Z^{(2)}, \dots, V_q^z = (\rho_{22}^{-1/2} f_q^z) Z^{(2)} \end{array} \end{array}$$

Notice that the canonical correlations do not change.

Relations between (U_k, V_k) and (U_k^z, V_k^z)

$$\begin{array}{ll} U_k = a_k'(X^{(1)} - \mu^{(1)}) & V_k = b_k'(X^{(2)} - \mu^{(2)}) \\ = a_{k1}(X_1^{(1)} - \mu_1^{(1)}) + \dots + a_{kp}(X_p^{(1)} - \mu_p^{(1)}) & = b_{k1}(X_1^{(2)} - \mu_1^{(2)}) + \dots + b_{kq}(X_q^{(2)} - \mu_q^{(2)}) \\ = (a_{k1}\sqrt{\sigma_{11}^{(1)}}, \dots, a_{kp}\sqrt{\sigma_{pp}^{(1)}}) \begin{pmatrix} Z_1^{(1)} \\ \vdots \\ Z_p^{(1)} \end{pmatrix} & = (b_{k1}\sqrt{\sigma_{11}^{(2)}}, \dots, b_{kq}\sqrt{\sigma_{qq}^{(2)}}) \begin{pmatrix} Z_1^{(2)} \\ \vdots \\ Z_q^{(2)} \end{pmatrix} \\ = U_k^z & = V_k^z \end{array}$$

Hence, $a_k^z = V_{11}^{1/2} a_k$ and $b_k^z = V_{22}^{1/2} b_k$ for $k = 1, 2, \dots, p$ and $V_{11} = \text{diag}(\sigma_{11}^{(1)}, \dots, \sigma_{pp}^{(1)})$ and $V_{22} = \text{diag}(\sigma_{11}^{(2)}, \dots, \sigma_{qq}^{(2)})$ **Another way to calculate ρ_k^2 , a_k , b_k**

1.

$$\det(\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \rho_k^2 I) = 0$$

2.

$$\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} (a_k) = \rho_k^2 (a_k)$$

3.

$$\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} (b_k) = \rho_k^2 (b_k)$$

Interpreting canonical variables and relations

The canonical variables can be interpreted much as in principal components.

1. by coefficients

$$\vec{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{pmatrix} = \begin{pmatrix} a'_1 \\ a'_2 \\ \vdots \\ a'_p \end{pmatrix} X^{(1)} = AX^{(1)}$$

$$\vec{V} = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_p \end{pmatrix} = \begin{pmatrix} b'_1 \\ b'_2 \\ \vdots \\ b'_q \end{pmatrix} X^{(2)} = BX^{(2)}$$

2. By correlations with the original variables.

$$\rho_{U, X^{(1)}} = \text{Cov}(U, X^{(1)}) = \text{Cov}(U, V_{11}^{-1/2}) = A \Sigma_{11} V_{11}^{-1/2}$$

Similarly, $\rho_{U, X^{(2)}} = A \Sigma_{12} V_{22}^{-1/2}$, $\rho_{V, X^{(2)}} = B \Sigma_{22} V_{22}^{-1/2}$ and $\rho_{V, X^{(1)}} = B \Sigma_{21} V_{11}^{-1/2}$

Shared variances

$$\rho_{U_k(X^{(1)})} = \max_b \text{Corr}(U_k, b' X^{(2)}) = \text{Corr}(U_k, V_k) = \rho_k$$

$$\rho_{V_k(X^{(1)})} = \max_a \text{Corr}(a' X^{(1)}, V_k) = \text{Corr}(U_k, V_k) = \rho_k$$

ρ_k^2 : the proportion of the variance of U_k explained by $X^{(2)}$ or of V_k explained by $X^{(1)}$.

Canonical correlations and variables based on S and R

Same analyses can be developed as above for S and R .

Canonical variables used as summaries of $X^{(1)}$ and $X^{(2)}$

Counterexample 10.3

$$\text{Cov} \begin{pmatrix} X_1^{(1)} \\ X_2^{(1)} \\ X_1^{(2)} \\ X_2^{(2)} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 1 & .95 & 0 \\ 0 & .95 & 1 & 0 \\ 0 & 0 & 0 & 100 \end{pmatrix}$$

Verify that the first pair of canonical variables are $U_1 = X_2^{(1)}$ and $V_1 = X_1^{(2)}$ and $\rho_1 = .95$ but they don't summarize Σ_{11} and Σ_{22} well.

When the canonical variables are good summaries?

$$X^{(1)} = A^{-1}U = (a^{(1)}, a^{(2)}, \dots, a^{(p)})U$$

Where $a^{(j)}$ are the j th column. Please note that $a^{(j)}$ consists of the coefficients associated with U_j . Similarly, we have

$$X^{(2)} = B^{-1}V = (b^{(1)}, b^{(2)}, \dots, b^{(q)})V$$

1. $\text{Cov}(U) = AS_{11}A' \Rightarrow S_{11} = A^{-1}(A^{-1})'$ (why)?

$$S_{11} = a^{(1)}a^{(1)'} + a^{(2)}a^{(2)'} + \dots + a^{(p)}a^{(p)'}$$

2. $\text{Cov}(V) = BS_{22}B' \Rightarrow S_{22} = B^{-1}(B^{-1})'$

$$S_{22} = b^{(1)}b^{(1)'} + b^{(2)}b^{(2)'} + \dots + b^{(q)}b^{(q)'}$$

3. $\text{Cov}(U, V)_{p \times q} = AS_{12}B$

$$S_{12} = \rho_1 a^{(1)}b^{(1)'} + \rho_2 a^{(2)}b^{(2)'} + \dots + \rho_p a^{(p)}b^{(p)'}$$

Represent $X^{(1)}$ and $X^{(2)}$ in terms of the first r pairs of canonical variables

$$\tilde{X}^{(1)} = (a^{(1)}, a^{(2)}, \dots, a^{(r)}) \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_r \end{pmatrix}$$

$$\tilde{X}^{(2)} = (b^{(1)}, b^{(2)}, \dots, b^{(r)}) \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_r \end{pmatrix}$$

So,

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

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

$$\tilde{S}_{12} = \rho_1 a^{(1)}b^{(1)'} + \dots + \rho_r a^{(r)}b^{(r)'}$$

Matrices of errors of approximations:

$$S_{11} - \tilde{S}_{11} = a^{(r+1)}a^{(r+1)'} + \dots + a^{(p)}a^{(p)'}$$

$$S_{22} - \tilde{S}_{22} = b^{(r+1)}b^{(r+1)'} + \dots + b^{(q)}b^{(q)'}$$

$$S_{12} - \tilde{S}_{12} = \rho_{r+1}a^{(r+1)}b^{(r+1)'} + \dots + \rho_p a^{(p)}b^{(q)'}$$

Propotion of explained sample variance (based on R)

$$U = A_z Z^{(1)} \Rightarrow Z^{(1)} = A_z^{-1}U$$

$$V = B_z Z^{(2)} \Rightarrow Z^{(2)} = B_z^{-1}V$$

$$A_z^{-1} = (a_z^{(1)}, a_z^{(2)}, \dots, a_z^{(p)}) = \begin{pmatrix} r_{U_1, z_1^1} & r_{U_2, z_1^1} & \cdots & r_{U_p, z_1^1} \\ r_{U_1, z_2^1} & r_{U_2, z_2^1} & \cdots & r_{U_p, z_2^1} \\ \vdots & \vdots & \cdots & \vdots \\ r_{U_1, z_p^1} & r_{U_2, z_p^1} & \cdots & r_{U_p, z_p^1} \end{pmatrix}$$

$$B_z^{-1} = (b_z^{(1)}, b_z^{(2)}, \dots, a_z^{(q)}) = \begin{pmatrix} r_{V_1, z_1^2} & r_{V_2, z_1^2} & \cdots & r_{V_q, z_1^2} \\ r_{V_1, z_2^2} & r_{V_2, z_2^2} & \cdots & r_{V_q, z_2^2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{V_1, z_q^2} & r_{V_2, z_q^2} & \cdots & r_{V_q, z_q^2} \end{pmatrix}$$

The total sample variances of $Z^{(1)}$ and $Z^{(2)}$ are

$$\text{trace}(R_{11}) = p$$

and

$$\text{trace}(R_{22}) = q$$

If the first r pairs of canonical variables are used to approximate the original variables.

$$Z^{(1)} \sim \tilde{Z}^{(1)} = (a_z^{(1)}, a_z^{(2)}, \dots, a_z^{(r)}) \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_r \end{pmatrix}$$

$$Z^{(2)} \sim \tilde{Z}^{(2)} = (b_z^{(1)}, b_z^{(2)}, \dots, b_z^{(r)}) \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_q \end{pmatrix}$$

The approximate variance and covariance matrices are

$$\tilde{R}_{11} = a_z^{(1)} a_z^{(1)'} + \cdots + a_z^{(r)} a_z^{(r)'}$$

$$\tilde{R}_{22} = b_z^{(1)} b_z^{(1)'} + \cdots + b_z^{(r)} b_z^{(r)'}$$

Then the approximate total sample variance based on $\tilde{Z}^{(1)}$ and $\tilde{Z}^{(2)}$ are

$$\text{trace} \tilde{R}_{11} = \text{trace} \left(\sum_{i=1}^r a_z^{(i)} a_z^{(i)'} \right) = \sum_{i=1}^p \sum_{j=1}^r r_{U_j, z_i^1}^2$$

$$\text{trace} \tilde{R}_{22} = \text{trace} \left(\sum_{i=1}^r b_z^{(i)} b_z^{(i)'} \right) = \sum_{i=1}^q \sum_{j=1}^r r_{U_j, z_i^1}^2$$

So,

$$R_{Z^{(1)}|U_1, \dots, U_r}^2 = \left(\begin{array}{c} \text{Proportion of total variance} \\ \text{explained by } U_1, \dots, U_r \end{array} \right) = \frac{\text{trace}(\tilde{R}_{11})}{p}$$

$$R_{Z^{(2)}|V_1, \dots, V_r}^2 = \left(\begin{array}{c} \text{Proportion of total variance} \\ \text{explained by } V_1, \dots, V_r \end{array} \right) = \frac{\text{trace}(\tilde{R}_{22})}{q}$$

Large sample inference

Under normal assumptions, formal tests are available to check if canonical correlation analysis is necessary.

$$H_0 : \Sigma_{12} = 0 \text{ versus } H_1 : \Sigma_{12} \neq 0$$

Likelihood ratio test with Bartlett correction: Reject H_0 at level α if

$$-(n-1 - \frac{1}{2}(p+q+1)) \log \prod_{i=1}^p (1 - \rho_i^2) > \chi_{pq}^2(\alpha)$$

If H_0 is rejected, then the following sequential tests can be further carried out.

$$H_0^k : \rho_1 \neq 0, \dots, \rho_k \neq 0, \rho_{k+1} = \dots = \rho_p = 0 \text{ v.s. } H_1^k : \rho_i \neq 0 \text{ for some } i > k$$

Reject H_0^k at level α if

$$-(n-1 - \frac{1}{2}(p+q+1)) \log \prod_{i=k+1}^p (1 - \rho_i^2) > \chi_{(p-k)(q-k)}^2(\alpha)$$