

Canonical Correlation Analysis

Now the coordinates of our random vector are partitioned into two parts, and we want to find maximally correlated linear combinations of them. Let $(\mathbf{X}^T, \mathbf{Y}^T)^T$ be a $(p + q)$ -dimensional random vector. For simplicity, we assume that the coordinates have zero expectation. The covariance matrix \mathbf{C} is partitioned (with block sizes p and q) in the following way:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\mathbf{X}\mathbf{X}} & \mathbf{C}_{\mathbf{X}\mathbf{Y}} \\ \mathbf{C}_{\mathbf{Y}\mathbf{X}} & \mathbf{C}_{\mathbf{Y}\mathbf{Y}} \end{pmatrix}$$

where $\mathbf{C}_{\mathbf{X}\mathbf{X}}$, $\mathbf{C}_{\mathbf{Y}\mathbf{Y}}$ are covariance matrices of \mathbf{X} and \mathbf{Y} , whereas $\mathbf{C}_{\mathbf{Y}\mathbf{X}} = \mathbf{C}_{\mathbf{X}\mathbf{Y}}^T$ is the cross-covariance matrix. Assume that $\mathbf{C}_{\mathbf{X}\mathbf{X}}$, $\mathbf{C}_{\mathbf{Y}\mathbf{Y}}$ and \mathbf{C} are regular.

Consider the following successive maximization problem. In the first step we look for maximally correlated linear combination of the \mathbf{X} - and \mathbf{Y} -coordinates. Obviously, the problem is equivalent to finding unit variance linear combinations with maximum covariance as follows:

$$\begin{aligned} \max_{\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q} \text{Corr}(\mathbf{a}^T \mathbf{X}, \mathbf{b}^T \mathbf{Y}) &= \max_{\text{Var}(\mathbf{a}^T \mathbf{X}) = \text{Var}(\mathbf{b}^T \mathbf{Y}) = 1} \text{Cov}(\mathbf{a}^T \mathbf{X}, \mathbf{b}^T \mathbf{Y}) \\ &= \max_{\substack{\mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{X}} \mathbf{a} = 1 \\ \mathbf{b}^T \mathbf{C}_{\mathbf{Y}\mathbf{Y}} \mathbf{b} = 1}} \mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{Y}} \mathbf{b}. \end{aligned}$$

With the notation

$$\tilde{\mathbf{a}} = \mathbf{C}_{\mathbf{X}\mathbf{X}}^{1/2} \mathbf{a}, \quad \tilde{\mathbf{b}} = \mathbf{C}_{\mathbf{Y}\mathbf{Y}}^{1/2} \mathbf{b}, \quad \mathbf{B} = \mathbf{C}_{\mathbf{X}\mathbf{X}}^{-1/2} \mathbf{C}_{\mathbf{X}\mathbf{Y}} \mathbf{C}_{\mathbf{Y}\mathbf{Y}}^{-1/2} \quad (1)$$

the above maximization task can be treated with linear algebraic tools. Let $\mathbf{B} = \mathbf{V}\mathbf{S}\mathbf{U}^T$ be the singular value decomposition (briefly, SVD) of the $p \times q$ matrix \mathbf{B} of rank $r \leq \min\{p, q\}$ (see Theorem 3 of Lesson 1). Then, by Proposition 3 of Lesson 1,

$$\max_{\substack{\mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{X}} \mathbf{a} = 1 \\ \mathbf{b}^T \mathbf{C}_{\mathbf{Y}\mathbf{Y}} \mathbf{b} = 1}} \mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{Y}} \mathbf{b} = \max_{\substack{\tilde{\mathbf{a}} \in \mathbb{R}^p, \tilde{\mathbf{b}} \in \mathbb{R}^q \\ \|\tilde{\mathbf{a}}\| = 1, \|\tilde{\mathbf{b}}\| = 1}} \tilde{\mathbf{a}}^T \mathbf{B} \tilde{\mathbf{b}} = s_1$$

and it is attained with the choice $\tilde{\mathbf{a}} = \mathbf{v}_1$ and $\tilde{\mathbf{b}} = \mathbf{u}_1$ (uniquely if $s_1 > s_2$), where $\mathbf{v}_1, \mathbf{u}_1$ is the first singular vector pair with corresponding singular value s_1 . The first canonical vector pair is obtained by back transformation:

$$\mathbf{a}_1 = \mathbf{C}_{\mathbf{X}\mathbf{X}}^{-1/2} \mathbf{v}_1, \quad \mathbf{b}_1 = \mathbf{C}_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{u}_1.$$

The pair of linear combinations $\mathbf{a}_1^T \mathbf{X}$, $\mathbf{b}_1^T \mathbf{Y}$ is called first canonical variable pair, and their correlation s_1 is the *first canonical correlation*.

This was the $k = 1$ case. For $k = 2, 3, \dots, r$ we are looking for maximally correlated linear combinations $\mathbf{a}^T \mathbf{X}$ and $\mathbf{b}^T \mathbf{Y}$ which are uncorrelated with the first $k-1$ canonical variables $\mathbf{a}_i^T \mathbf{X}$ and $\mathbf{b}_i^T \mathbf{Y}$ ($i = 1, \dots, k-1$), respectively. Since the correlations are again viewed as covariances of the unit-variance variables, and also making use of the relations

$$\text{Cov}(\mathbf{a}^T \mathbf{X}, \mathbf{a}_i^T \mathbf{X}) = \mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{X}} \mathbf{a}_i, \quad \text{Cov}(\mathbf{b}^T \mathbf{Y}, \mathbf{b}_i^T \mathbf{Y}) = \mathbf{b}^T \mathbf{C}_{\mathbf{Y}\mathbf{Y}} \mathbf{b}_i,$$

our maximization problem, making use of (1), again making use of by Proposition 3 of Lesson 1, has the following form and solution:

$$\max_{\substack{\mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{X}} \mathbf{a} = 1 \\ \mathbf{b}^T \mathbf{C}_{\mathbf{Y}\mathbf{Y}} \mathbf{b} = 1 \\ \mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{X}} \mathbf{a}_i = 0 \ (i=1, \dots, k-1) \\ \mathbf{b}^T \mathbf{C}_{\mathbf{Y}\mathbf{Y}} \mathbf{b}_i = 0 \ (i=1, \dots, k-1)}} \mathbf{a}^T \mathbf{C}_{\mathbf{X}\mathbf{Y}} \mathbf{b} = \max_{\substack{\tilde{\mathbf{a}} \in \mathbb{R}^p, \tilde{\mathbf{b}} \in \mathbb{R}^q \\ \|\tilde{\mathbf{a}}\| = 1, \|\tilde{\mathbf{b}}\| = 1 \\ \tilde{\mathbf{a}}^T \mathbf{u}_i = 0 \ (i=1, \dots, k-1) \\ \tilde{\mathbf{b}}^T \mathbf{v}_i = 0 \ (i=1, \dots, k-1)}} \tilde{\mathbf{a}}^T \mathbf{B} \tilde{\mathbf{b}} = s_k$$

and the maximum is attained with the choice $\tilde{\mathbf{a}} = \mathbf{v}_k$ and $\tilde{\mathbf{b}} = \mathbf{u}_k$ (uniquely if $s_k > s_{k+1}$).

The k th *canonical vector pair* is obtained by back transformation:

$$\mathbf{a}_k = \mathbf{C}_{\mathbf{X}\mathbf{X}}^{-1/2} \mathbf{v}_k, \quad \mathbf{b}_k = \mathbf{C}_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{u}_k.$$

The pair of linear combinations $\mathbf{a}_k^T \mathbf{X}$, $\mathbf{b}_k^T \mathbf{Y}$ is called k th *canonical variable pair*, and their correlation s_k is the k th *canonical correlation*.

Since the canonical correlations are, in fact, correlations, for the singular values of the matrix \mathbf{B} , the relation

$$1 \geq s_1 \geq s_2 \geq \cdots \geq s_r > 0$$

holds. In case of a multivariate Gaussian sample, a sequence of hypotheses can be tested for the number of canonical correlations significantly explaining the relation between \mathbf{X} and \mathbf{Y} . The likelihood ratio statistic is based on the singular values of the matrix $\hat{\mathbf{B}}$ calculated from the empirical covariances in the following way:

$$\hat{\mathbf{B}} = \hat{\mathbf{C}}_{\mathbf{X}\mathbf{X}}^{-1/2} \hat{\mathbf{C}}_{\mathbf{X}\mathbf{Y}} \hat{\mathbf{C}}_{\mathbf{Y}\mathbf{Y}}^{-1/2}.$$

By the invariance of the ML-estimation, the singular values and vector pairs of the matrix $\hat{\mathbf{B}}$ are ML-estimates of the singular values and vector pairs of the matrix \mathbf{B} (provided there are no multiple singular values, but this has zero probability).

Consider the sequence of hypotheses

$$H_{0k} : s_{k+1} = \cdots = s_{\min\{p,q\}} = 0 \quad (k = 0, \dots, r-1),$$

where $r = \text{rank}(\hat{\mathbf{B}})$.

Denoting the test statistic of the likelihood ratio test based on an n -element sample by λ_n , we can prove that

$$-2 \ln \lambda_n = -n \sum_{i=k+1}^r \ln(1 - r_i^2),$$

where r_i 's are the singular values of $\hat{\mathbf{B}}$, i.e., the ML-estimates of the canonical correlations.

The general theory guarantees that the test statistic $-2 \ln \lambda_n$ asymptotically follows $\chi^2(f)$ distribution as $n \rightarrow \infty$. The degree of freedom f is the number of the model's parameters (pq) minus the number of the parameters in the reduced rank model under H_{0k} , which is

$$\begin{aligned} & k + [(q-1) + (q-2) + \cdots + (q-k)] + [(p-1) + (p-2) + \cdots + (p-k)] = \\ & = k + \frac{(2q-k-1)k}{2} + \frac{(2p-k-1)k}{2} = qk + pk - k^2. \end{aligned}$$

Therefore,

$$f = pq - (qk + pk - k^2) = (p-k)(q-k).$$

For 'large' n , we will reject H_{0k} with significance α if $-2 \ln \lambda_n \geq \chi_\alpha^2(f)$.

We perform the sequential test for $k = 0, 1, \dots, r-1$, in this order, until H_{0k} is accepted for some k . This k will be the significant number of the canonical correlations that explains the interrelations between the two sets of variables.

Observe that in the case of $k = 0$, we investigate the independence of the two sets of variables.

Based on a sample entry with realization \mathbf{x}, \mathbf{y} , its *canonical scores* $\mathbf{a}_i^T \mathbf{x}, \mathbf{b}_i^T \mathbf{y}$ ($i = 1, \dots, k$) can be calculated and further used for k -dimensional representation or clustering of the data points.