MULTIVARIATE STATISTICS, Lesson 10.
Canonical Correlation Analysis

Let $\mathbf{X} = (X_1, \dots, X_p)^T$ and $\mathbf{Y} = (Y_1, \dots, Y_q)^T$ be p- and q-dimensional random vectors with $\mathbf{0}$ expectation, respectively. Let $\mathbf{C}_{11}$, $\mathbf{C}_{22}$, $\mathbf{C}_{12}$ denote their covariance- and cross-covariance matrices (the covariance matrices are not singular).

- **Problem**: We are looking for vectors $\mathbf{a}_1 \in \mathbb{R}^p$ and $\mathbf{b}_1 \in \mathbb{R}^q$ such that

$$\mathrm{Corr}\,(\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y})$$

is maximum. Then for $k = 2, 3, \dots, l = \min\{p, q\}$ we are looking for vectors $\mathbf{a}_k \in \mathbb{R}^p$ and $\mathbf{b}_k \in \mathbb{R}^q$ such that

$$\mathrm{Corr}\,(\mathbf{a}_k^T\mathbf{X}, \mathbf{b}_k^T\mathbf{Y})$$

is maximum on the constraints that

$$\mathrm{Corr}\,(\mathbf{a}_k^T\mathbf{X}, \mathbf{a}_i^T\mathbf{X}) = 0, \quad \mathrm{Corr}\,(\mathbf{b}_k^T\mathbf{Y}, \mathbf{b}_i^T\mathbf{Y}) = 0 \qquad (i = 1, \dots, k-1).$$

The value of the k-th maximum, ϱ_k is called k-th *canonical correlation coefficient*.

- The **solution** is obtained by the SVD of the matrix

$$\mathbf{D} = \mathbf{C}_{11}^{-1/2}\mathbf{C}_{12}\mathbf{C}_{22}^{-1/2}.$$

(Remark that $\mathrm{rang}\,(\mathbf{D}) = \mathrm{rang}\,(\mathbf{C}_{12}) = r$.) The singular values are $1 \geq \rho_1 \geq \cdots \geq \rho_l \geq 0$, and denoting by $\mathbf{v}_1, \dots, \mathbf{v}_l$ and $\mathbf{u}_1, \dots, \mathbf{u}_l$ the corresponding left- and right-hand side singular vectors,

$$\mathbf{a}_k = \mathbf{C}_{11}^{-1/2}\mathbf{v}_k, \quad \mathbf{b}_k = \mathbf{C}_{22}^{-1/2}\mathbf{u}_k \qquad (k = 1, \dots, l)$$

are the *canonical vector pairs*, while the r.v.'s $\mathbf{a}_k^T\mathbf{X}$, $\mathbf{b}_k^T\mathbf{Y}$, $(k = 1, \dots, l)$ are the *canonical variable pairs*.

- **Estimation** of the canonical correlations and variables: by the SVD of the matrix $\hat{\mathbf{D}}$ based on the empirical covariance matrices.

- **Hypothesis testing**. Based on multivariate normality, for testing the hypothesis

$$H_{0k} : \rho_{k+1} = \cdots = \rho_l = 0,$$

that is k canonical correlations are enough to explain the connection between $\mathbf{X}$ and $\mathbf{Y}$, we use the transformed likelihood ratio test statistic

$$-2\ln\lambda_n = -n\sum_{i=k+1}^{\min\{p,q\}} \ln(1 - \hat{\rho}_i^2)$$

that (for large sample size n) asymptotically follows $\chi^2((p-k)(q-k))$-distribution under H_{0k}. (Here $\hat{\rho}$'s denote the sample estimates of ρ's.)

The sequential testing goes on for $k = 0, 1, 2, \dots$ until accepting one. Remark that $k \leq r-1$, and the $k = 0$ case corresponds to testing the independence of $\mathbf{X}$ and $\mathbf{Y}$.

- **Results**: the caninical loadings (coordinates of the canonical vectors) help to interpret the canonical variable pairs. **Spacial representation**: by the canonical scores $\mathbf{a}_i^T\mathbf{x}_j$, $\mathbf{b}_i^T\mathbf{y}_j$ $(i = 1, \dots, k)$ based on measurements $\mathbf{x}_j, \mathbf{y}_j$, $j = 1, \dots, n$ (store them for further analysis).