# MULTIVARIATE STATISTICAL ANALYSIS

# Lecture 6
# Canonical Correlation Analysis

Associate Professor Lý Quốc Ngọc

TRUONG DAI HOC KHOA HOC TU NHIEN
KHTN
TP. HO CHI MINH

cdio 4.0

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

fit@hcmus

# Contents

# 6.1. Purpose of CCA

Canonical correlation analysis seeks to identify and quantify the associations between two sets of variables.
CCA focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set.

# 6.2. Problem statement



Consider the first group, of p variables, is represented by the random vector $X^{(1)}(p\times 1)$ .

Consider the second group, of q variables, is represented by the random vector $X^{(2)}(q\times 1)$ .

Consider the linear combination of $X^{(1)}$ and $X^{(2)}$

$$U = a^T X^{(1)}$$

$$V = b^T X^{(2)}$$

We shall seek coefficient vectors $a,b$ such that correlation of $U,V$ is as large as possible.

$$Corr(U,V) = \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_{11} a}\sqrt{b^T \Sigma_{22} b}}$$

# 6.2. Problem statement

With the canonical variates satisfy the constraints of variance, covariance as follows:

$$Var(U_k) = Var(V_k) = 1$$

$$Cov(U_k, U_l) = Corr(U_k, U_l) = 0 \quad \mathrm{k} \neq 1$$

$$Cov(V_k, V_l) = Corr(V_k, V_l) = 0 \quad \mathrm{k} \neq 1$$

$$Cov(U_k, V_l) = Corr(U_k, V_l) = 0 \quad \mathrm{k} \neq 1$$

$$\mathrm{k}, 1 = 1, 2, \ldots, \mathrm{p}$$

# 6.3. Method

Seek $a,b$ to maximize $Corr(U,V)$ with constraints:

$$\max_{a,b} Corr(U,V) = \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_{11} a}\sqrt{b^T \Sigma_{22} b}}$$

$$a^T \Sigma_{11} a = 1$$

$$b^T \Sigma_{22} b = 1$$

Consider Lagrange function:

$$L(a,b,\lambda) = a^T \Sigma_{12} b + \frac{\lambda_a}{2}(1 - a^T \Sigma_{11} a) + \frac{\lambda_b}{2}(1 - b^T \Sigma_{22} b)$$

# 6.3. Method

fit@hcmus

The necessary condition for $Corr(U,V)$ reaching the maximum with constraints is

$$\frac{\partial L(a,b,\lambda)}{\partial a} = \Sigma_{12} b - \lambda_a \Sigma_{11} a = 0 \quad (1)$$

$$\frac{\partial L(a,b,\lambda)}{\partial b} = \Sigma_{12}{}^T a - \lambda_b \Sigma_{22} b = \Sigma_{21} a - \lambda_b \Sigma_{22} b = 0 \quad (2)$$

Từ (1) và (2) ta có:

$$\mathrm{a}^{\mathrm{T}} \Sigma_{12} b - \lambda_a \mathrm{a}^{\mathrm{T}} \Sigma_{11} a + \lambda_b \mathrm{b}^{\mathrm{T}} \Sigma_{22} b - \mathrm{b}^{\mathrm{T}} \Sigma_{12} a = 0$$

$$\lambda_b \mathrm{b}^{\mathrm{T}} \Sigma_{22} b - \lambda_a \mathrm{a}^{\mathrm{T}} \Sigma_{11} a = 0$$

$$\lambda_b - \lambda_a = 0$$

Từ (1) ta có: $a = \dfrac{\Sigma_{11}^{-1} \Sigma_{12} b}{\lambda}$

# 6.3. Method

Replace $a$ into (2) we have:

$$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\,b=\lambda^2 b$$

Similarly we have:

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\,a=\lambda^2 a$$

Therefore $a$ is the eigenvector of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

$b$ is the eigenvector of $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

fit@hcmus

# 6.3. Method

The first pair of canonical variables:

$$U_1 = a_1^T X = e_1^T \Sigma_{11}^{-1/2} X, \quad V_1 = b_1^T Y = f_1^T \Sigma_{22}^{-1/2} Y$$

Generalization for the kth pair of canonical variables::

$$U_k = a_k^T X = e_k^T \Sigma_{11}^{-1/2} X$$

$$V_k = b_k^T Y = f_k^T \Sigma_{22}^{-1/2} Y$$

$$\max_{a,b} Corr(U_k, V_k) = \rho_k^*$$

Assume $\rho_1^{*2} \geq \rho_2^{*2} \geq \ldots \geq \rho_p^{*2}$ and $(\rho_k^{*2}, e_k), (\rho_k^{*2}, f_k)$ are eigen values, eigenvectors of:

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

# 6.4. Geometrical Explanation

Discuss in class

# 6.5. Model Checking

Discuss in class

# 6.6. Case study

Discuss in class