Statistical Inference and Multivariate Analysis (MA324)

Lecture Slides
Lecture 35

Canonical Correlation Analysis and Factor Analysis

Indian Institute of Technology Guwahati

Jan-May 2023

Canonical Correlation Analysis

- Canonical correlation analysis aims to **measure the associations between two sets of variables**.
- Relating your current **"academic performance"** variables with COVID period **"achievement"** variables.
- The objective of the canonical correlation analysis is to find the correlation **between a linear combination** of the variables in one set and a linear combination of the variables in another set.
- It is **high-dimensional relationship** between the two sets of variables into a few pairs of canonical variables.

Application ...

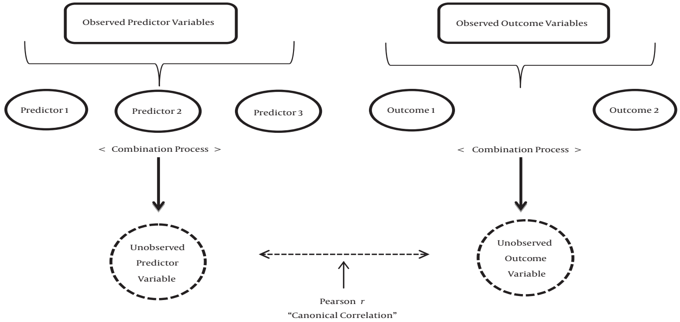


Figure 1. Illustration of the First Function in a Canonical Correlation Analysis With Three Predictors and Two Criterion Variables

- Reference: Sadoughi, F., Lotfnezhad Afshar, H., Olfatbakhsh, A., & Mehrdad, N. (2016). Application of Canonical Correlation Analysis for Detecting Risk Factors Leading to Recurrence of Breast Cancer. Iranian Red Crescent medical journal, 18(3), e23131.

Canonical Correlation Analysis

- Let $\underset{\sim}{X}^{(1)}_{p\times 1}$ and $\underset{\sim}{X}^{(2)}_{q\times 1}$ be two random vector.
- We assume that $p \leq q$.
- $E(\underset{\sim}{X}^{(1)}) = \underset{\sim}{\mu}^{(1)}$, $E(\underset{\sim}{X}^{(2)}) = \underset{\sim}{\mu}^{(2)}$.
- $Var(\underset{\sim}{X}^{(1)}) = \Sigma_{11}, Var(\underset{\sim}{X}^{(2)}) = \Sigma_{22}$
- $Cov(\underset{\sim}{X}^{(1)}, \underset{\sim}{X}^{(2)}) = (\Sigma_{12})_{p\times q}, Cov(\underset{\sim}{X}^{(2)}, \underset{\sim}{X}^{(1)}) = (\Sigma_{21})_{q\times p}$
  In this case $\Sigma_{12} = \Sigma_{21}'$

- $(\underset{\sim}{X})_{(p+q)\times 1} = \begin{pmatrix} \underset{\sim}{X}^{(1)} \\ \underset{\sim}{X}^{(2)} \end{pmatrix}$, $(\underset{\sim}{\mu})_{(p+q)\times 1} = \begin{pmatrix} \underset{\sim}{\mu}^{(1)} \\ \underset{\sim}{\mu}^{(2)} \end{pmatrix} = E(\underset{\sim}{X})$,

  $\Sigma_{(p+q)\times(p+q)} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = Var(\underset{\sim}{X})$

- First we want to find $\underset{\sim}{a}$ and $\underset{\sim}{b}$ such that $Corr(U_1 = \underset{\sim}{a}_1' \underset{\sim}{X}^{(1)}, V_1 = \underset{\sim}{b_1}' \underset{\sim}{X}^{(2)})$ is maximum.

- Note that,

$$Corr(U_1, V_1) = \frac{\underset{\sim}{a}_1' \Sigma_{12} \underset{\sim}{b}_1}{\sqrt{\underset{\sim}{a}_1' \Sigma_{11} \underset{\sim}{a}_1}\sqrt{\underset{\sim}{b_1}' \Sigma_{22} \underset{\sim}{b}_1}}$$

- We need to solve the optimization problem,

$$\max_{\underset{\sim}{a_1}, \underset{\sim}{b_1}} \frac{\underset{\sim}{a_1}' \Sigma_{12} \underset{\sim}{b_1}}{\sqrt{\underset{\sim}{a_1}' \Sigma_{11} \underset{\sim}{a_1}} \sqrt{\underset{\sim}{b_1}' \Sigma_{22} \underset{\sim}{b_1}}}, \ subject\ to\ Var(U_1) = Var(V_1) = 1$$

- The **first pair** of canonical variables is the **pair of linear combinations** U_1 and V_1 having **unit variance**, which **maximizes** the **correlation** between U_1 and V_1.

- The k^{th} pair of canonical variables is the **pair of linear combinations** U_k and V_k having **unit variance**, which **maximizes** the **correlation** between U_k and V_k among all choices **uncorrelated with the previous** $(k-1)$ canonical variables.

Theorem

Assume that Σ has full rank. The **first pair of canonical variables** are given by, $U_1 = \underset{\sim}{e_1}' \Sigma_{11}^{-\frac{1}{2}} \underset{\sim}{X}^{(1)}$, $V_1 = \underset{\sim}{f_1}' \Sigma_{22}^{-\frac{1}{2}} \underset{\sim}{X}^{(2)}$ and $Corr(U_1, V_1) = \rho_1^*$.

The k^{th} $(k = 1, 2, ..., p)$ pair of canonical variables are given by,
$U_k = \underset{\sim}{e_k}' \Sigma_{11}^{-\frac{1}{2}} \underset{\sim}{X}^{(1)}$, $V_k = \underset{\sim}{f_k}' \Sigma_{22}^{-\frac{1}{2}} \underset{\sim}{X}^{(2)}$ and $Corr(U_k, V_k) = \rho_k^*$.

Here $\rho_1^* \geq \rho_2^* \geq ... \geq \rho_p^*$ are the **eigen values** of $\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}$ and $\underset{\sim}{e_1}, ..., \underset{\sim}{e_p}$ are corresponding **eigen vectors**. The quantities $\rho_1^*, ..., \rho_p^*$ are **also** p

largest **eigen values** of the matrix, $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$ with corresponding **eigen vectors** $\underset{\sim}{f_1}, ..., \underset{\sim}{f_p}$.

Factor Analysis : Motivation

- Try and understand, the **covariance relationships among many variables** in terms of **a few** underlying, but **unobservable, random quantities** called **factors**.
- Suppose all variables **within** a particular **group** are **highly correlated** among themselves, but have relatively **small correlations** with variables in a **different group**.
- Then it is conceivable that **each group of variables** represents a **single** underlying **construct, or factor**, that is responsible for the observed correlations.
- For example, correlations from the **group of test scores** in classics, Assamese (Asamiya), English, mathematics, and music may suggest an underlying "**intelligence**" factor. A **second group** of variables representing **physical-fitness scores**, may corresponds to another factor.

Orthogonal Factor Model with m common Factors

- X be the observable random vector with p components with mean μ and variance-covariance matrix Σ.

$$
\begin{aligned}
\underset{(p\times 1)}{\mathbf{X}} &= \underset{(p\times 1)}{\boldsymbol{\mu}} + \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \\
\Longrightarrow \underset{(p\times 1)}{\mathbf{X}} - \underset{(p\times 1)}{\boldsymbol{\mu}} &= \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \\
\mu_i &= mean\ of\ variable\ i, i = 1, \cdots, p, \\
\varepsilon_i &= i^{th}\ specific\ factor(errors), i = 1, \cdots, p, \\
F_j &= j^{th}\ common\ factor, j = 1, \cdots, m, \\
l_{ij} &= loading\ of\ the\ i^{th}\ variable\ on\ j^{th}\ factor. \\
F &= [F_1, \cdots, F_m]'; L = [l_{ij}]
\end{aligned}
$$

Covariance Structure for the Orthogonal Factor Model

- The unobservable random vectors $\mathbf{F}$ and ε satisfy the following conditions:
  - $\mathbf{F}$ and ε are independent.
  - $E(\mathbf{F}) = \mathbf{0}$, $Cov(\mathbf{F}) = \mathbf{I}$.
  - $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $Cov(\boldsymbol{\varepsilon}) = \Psi$, where Ψ is a **diagonal** matrix.
  - $Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = 0$
- $Cov(\mathbf{X}) = \mathbf{L}\mathbf{L}^{'} + \boldsymbol{\Psi}$
  or it can also be expressed as,
  $Var(X_i) = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2 + \psi_i$
  $Cov(X_i, X_k) = l_{i1}l_{k1} + ... + l_{im}l_{km}.$
- $Cov(\mathbf{X}, \mathbf{F}) = \mathbf{L}$
  or
  $Cov(X_i, F_j) = l_{ij}.$

Some Features

- The model $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$ is linear in common factors.
- The portion of the variance of the i^{th} variable (X_i) contributed by the m common factors is termed as the i^{th} ***communality***.

$$Var(X_i) = \sigma_{ii} = \underbrace{l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2}_{communality} + \underbrace{\psi_i}_{specific\ variance}$$

- Also, the portion of $Var(X_i) = \sigma_{ii}$ due to the specific factor (ϵ_i) is called the ***uniqueness***, or ***specific variance***.
- Also, it can be seen easily that $\sigma_{ii} = h_i^2 + \psi_i$, $i = 1, 2, ..., p$, where $h_i^2 = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2$.

- If $m > 1$, then there is always some **inherent ambiguity** associated with the factor model. Let T be any $m \times m$ orthogonal matrix, so

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} = \mathbf{LTT'F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^{*'} + \boldsymbol{\varepsilon},$$

  where $\mathbf{L}^* = \mathbf{LT}$ and $\mathbf{F}^{*'} = \mathbf{T'F}$

- Factor loadings $\mathbf{L}$ are determined only up to an orthogonal matrix $\mathbf{T}$. Thus the loadings,

$$\mathbf{L}^* = \mathbf{LT} \; and \; \mathbf{L}$$

  both give the same representation. The **communalities**, given by the diagonal elements of $\mathbf{LL}^{'} = (\mathbf{L}^*)(\mathbf{L}^*)^{'}$ are also **unaffected** by the choice of $\mathbf{T}$.

- The **analysis** of the factor model **proceeds by imposing conditions** that **allow** one to **uniquely estimate** $\mathbf{L}$ and Ψ. ($\mathbf{L}'\Psi^{-1}\mathbf{L} = \; diagonal \; matrix$).

Methods of Estimation

- The sample covariance matrix $\mathbf{S}$ is an estimator of the unknown population covariance matrix $\mathbf{\Sigma}$.
    - If the **off-diagonal elements** of matrix $\mathbf{S}$ are **small** or those of the sample correlation matrix $\mathbf{R}$ essentially zero, the variables are not related, and a **factor analysis will not be useful** in that scenario. In these circumstances, the specific factors play the dominant role, whereas the **major aim** of factor analysis is to **determine a few important common factors**.
    - If $\mathbf{\Sigma}$ appears to **deviate significantly from any diagonal matrix**, then a factor model can be fitted, With the initial problem being **estimating** the **factor loadings** l_{ij} and **specific variances** ψ_i.

- The two most popular methods for parameter estimation are:

    - Principal Component (or Principal Factor).

    - Maximum Likelihood (Self Study).

Principal Component Solution of the Factor Model

- The principal component factor analysis of the sample covariance matrix $\mathbf{S}$ is specified in terms of its eigenvalues-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), ..., (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 \geq ... \geq \hat{\lambda}_p$. Let $m < p$ be the number of common factors. Then the matrix of estimated factor loadings $\{\tilde{l}_{ij}\}$ is obtained as,

$$\tilde{\mathbf{L}} = \left[\begin{array}{c:c:c:c} \sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 & \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 & \cdots & \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m \end{array} \right]$$

- The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}'$, so,

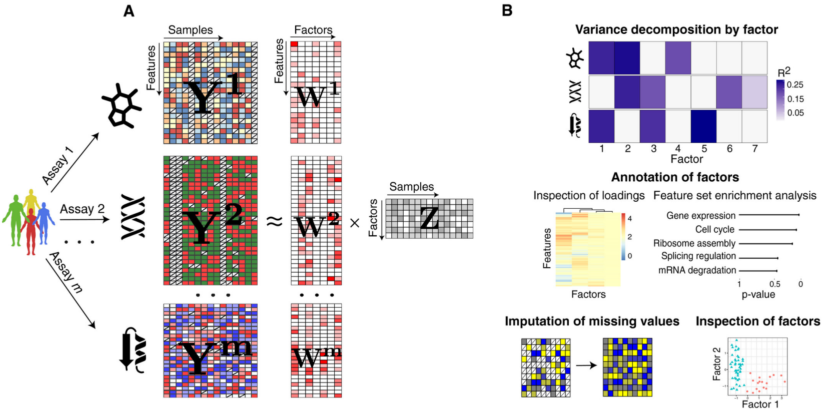
$$\tilde{\mathbf{\Psi}} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix}, \; with \; \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{l}_{ij}^2$$

- **Communalities** are estimated as,

$$\tilde{h}_i^2 = \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + ... + \tilde{l}_{im}^2.$$

- The principal component **factor analysis of the sample correlation matrix** is obtained by starting with $\mathbf{R}$ in place of $\mathbf{S}$.

Application of Factor Analysis...



- Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics data sets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation and the detection of outlier samples.
- Reference: Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., ... & Stegle, O. (2018). Multi-Omics Factor Analysis–a framework for unsupervised integration of multi-Omics data sets. Molecular systems biology, 14(6), e8124.