Lecture 10
Canonical Correlation Analysis

Reading: Johnson & Wichern 2007, Chapter 10;
Zelterman Chapter 13.2; Izenman Chapter 7.3

DSA 8070 Multivariate Analysis

Canonical
Correlation
Analysis

CLEMS⬛N
U N I V E R S I T Y

Background

Canonical Variates
& Canonical
Correlations

Sales Data
Example

10.1

Whitney Huang
Clemson University

Notes

Agenda

1 **Background**

2 **Canonical Variates & Canonical Correlations**

3 **Sales Data Example**

Canonical
Correlation
Analysis

CLEMS⬛N
U N I V E R S I T Y

Background

Canonical Variates
& Canonical
Correlations

Sales Data
Example

10.2

Notes

Overview

Canonical correlation analysis (CCA, Hotelling, 1936) is a method for exploring the relationships between two sets of multivariate variables $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^T$ and $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_q)^T$

RELATIONS BETWEEN TWO SETS OF VARIATES*.

By HAROLD HOTELLING, Columbia University.

CONTENTS.

1. *The Correlation of Vectors. The Most Predictable Criterion and the Tetrad Difference.* Concepts of correlation and regression may be applied not only to ordinary one-dimensional variates but also to variates of two or more dimensions.

Canonical
Correlation
Analysis

CLEMS⬛N
U N I V E R S I T Y

Background

Canonical Variates
& Canonical
Correlations

Sales Data
Example

10.3

Notes

Relating Two Random Vectors

Examples:

- $X = (X_1, X_2)$ represents two **reading** test scores, and $Y = (Y_1, Y_2)$ represents two **arithmetic** test scores

- X is a vector of variables associated with **environmental health**: species diversity, total biomass, and environmental productivity, while Y represents concentrations of heavy metals, pesticides, and dioxin, which measure **environmental toxins**

Goal: CCA relates two sets of variables X and Y by finding linear combinations of variables that maximally correlated

Motivation: relates X and Y using a small number of linear combinations for ease of interpretation

---

Linear Combinations of Two Sets of Variables

Recall we have $X = (X_1, X_2, \cdots, X_p)^T$ and $Y = (Y_1, Y_2, \cdots, Y_q)^T$. Without loss of generality, let's assume $p \leq q$.

Similar to PCA, we define a set of linear combinations

$$U_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$
$$U_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$
$$\vdots = \cdots$$
$$U_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

and

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q$$
$$V_2 = b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q$$
$$\vdots = \cdots$$
$$V_p = b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pq}Y_q$$

We want to find linear combinations that maximize the correlation of $(U_i, V_i), \quad i = 1, \cdots, p$

Defining Canonical Variates

We call (U_i, V_i) be the ith canonical variate pair. One can compute the variance of U_i with the following expression:

$$\text{Var}(U_i) = \sum_{k=1}^{p} \sum_{\ell=1}^{p} a_{ik} a_{i\ell} \text{Cov}(X_k, X_\ell), \quad i = 1, \cdots, p.$$

Similarly, we compute the variance of V_j with the following expression:

$$\text{Var}(V_j) = \sum_{k=1}^{q} \sum_{\ell=1}^{q} b_{jk} b_{j\ell} \text{Cov}(Y_k, Y_\ell), j = 1, \cdots, q.$$

The covariance between U_i and V_j is:

$$\text{Cov}(U_i, V_j) = \sum_{k=1}^{p} \sum_{\ell=1}^{q} a_{ik} b_{j\ell} \text{Cov}(X_k, Y_\ell).$$

The canonical correlation for the ith canonical variate pair is simply the correlation between U_i and V_i:

$$\rho_i^* = \frac{\text{Cov}(U_i, V_i)}{\sqrt{\text{Var}(U_i)\text{Var}(V_i)}}$$

Finding Canonical Variates

Let us look at each of the p canonical variates pair one by one.

First canonical variable pair (U_1, V_1): The coefficients $a_{11}, a_{12}, \cdots, a_{1p}$ and $b_{11}, b_{12}, \cdots, b_{1q}$ are chosen to maximize the canonical correlation ρ_1^*. As in PCA, this is subject to the constraint that $\text{Var}(U_1) = \text{Var}(V_1) = 1$

Second canonical variable pair (U_2, V_2): Similarly we want to find $a_{21}, a_{22}, \cdots, a_{2p}$ and $b_{21}, b_{22}, \cdots, b_{2q}$ that maximize ρ_2^* under the following constraints:

$$\text{Var}(U_2) = \text{Var}(V_2) = 1,$$
$$\text{Cov}(U_1, U_2) = \text{Cov}(V_1, V_2) = 0,$$
$$\text{Cov}(U_1, V_2) = \text{Cov}(U_2, V_1) = 0.$$

This procedure is repeated for each pair of canonical variates

Canonical Correlation Analysis

CLEMS❀N
U N I V E R S I T Y

Background

Canonical Variates & Canonical Correlations

Sales Data Example

10.7

Notes

Finding Canonical Variates Cont'd

Let $\text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma_X}$ and $\text{Var}(\boldsymbol{Y}) = \boldsymbol{\Sigma_Y}$ and let $\boldsymbol{Z}^T = (\boldsymbol{X}^T, \boldsymbol{Y}^T)$. Then the covariance matrix of $\boldsymbol{Z}$ is

$$\begin{bmatrix} \text{Var}(\boldsymbol{X}) & \text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) \\ \text{Cov}(\boldsymbol{Y}, \boldsymbol{X}) & \text{Var}(\boldsymbol{Y}) \end{bmatrix} = \begin{bmatrix} \underbrace{\boldsymbol{\Sigma_X}}_{p \times p} & \underbrace{\boldsymbol{\Sigma_{XY}}}_{p \times q} \\ \underbrace{\boldsymbol{\Sigma_{YX}}}_{q \times p} & \underbrace{\boldsymbol{\Sigma_Y}}_{q \times q} \end{bmatrix}$$

The ith pair of canonical variates is given by

$$U_i = \underbrace{\boldsymbol{u}_i^T \boldsymbol{\Sigma}_X^{-1/2}}_{\boldsymbol{a}_i^T} \boldsymbol{X} \text{ and } V_i = \underbrace{\boldsymbol{v}_i^T \boldsymbol{\Sigma}_Y^{-1/2}}_{\boldsymbol{b}_i^T} \boldsymbol{Y},$$

where

- $\boldsymbol{u}_i$ is the ith eigenvector of $\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma_{XY}} \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\Sigma_{YX}} \boldsymbol{\Sigma}_X^{-1/2}$
- $\boldsymbol{v}_i$ is the ith eigenvector of $\boldsymbol{\Sigma}_Y^{-1/2} \boldsymbol{\Sigma_{YX}} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma_{XY}} \boldsymbol{\Sigma}_Y^{-1/2}$
- The ith canonical correlation is given by, $\text{Cor}(U_i, V_i) = \rho_i^*$, where ρ_i^{*2} is the ith eigenvalue of $\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma_{XY}} \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\Sigma_{YX}} \boldsymbol{\Sigma}_X^{-1/2}$

Canonical Correlation Analysis

CLEMS❀N
U N I V E R S I T Y

Background

Canonical Variates & Canonical Correlations

Sales Data Example

10.8

Notes

Likelihood Ratio Test: Is CCA Worthwhile?

Note that if $\boldsymbol{\Sigma_{XY}} = \boldsymbol{0}$, then $\text{Cov}(\boldsymbol{U}, \boldsymbol{V}) = \boldsymbol{a}^T \boldsymbol{\Sigma_{XY}} \boldsymbol{b} = 0$ for all $\boldsymbol{a}$ and $\boldsymbol{b} \Rightarrow$ all canonical correlations must be zero and there is no point in pursuing CCA.

For large n, we reject $H_0 : \boldsymbol{\Sigma_{XY}} = \boldsymbol{0}$ in favor of $H_1 : \boldsymbol{\Sigma_{XY}} \neq \boldsymbol{0}$ if

$$-2\log(\Lambda) = n \log\left(\frac{|\hat{\boldsymbol{\Sigma}}_X||\hat{\boldsymbol{\Sigma}}_Y|}{|\hat{\boldsymbol{\Sigma}}|}\right) = -n \sum_{j=1}^{p} \log(1 - \hat{\rho}_j^{*2})$$

is larger than $\chi_{pq}^2(\alpha)$

For an improvement to the χ^2 approximation, Bartlett suggested using the following test statistic:

$$-2\log(\Lambda) = -[n - 1 - \frac{1}{2}(p + q + 1)] \sum_{j=1}^{p} \log(1 - \hat{\rho}_j^{*2})$$

Canonical Correlation Analysis

CLEMS❀N
U N I V E R S I T Y

Background

Canonical Variates & Canonical Correlations

Sales Data Example

10.9

Notes

Example: Sales Data [Source: PSU STAT 505]

The example data comes from a firm that surveyed a random sample of $n = 50$ of its employees in an attempt to determine which factors influence sales performance. Two collections of variables were measured:

- Sales Performance: `Sales Growth`, `Sales Profitability`, `New Account Sales`
  $\Rightarrow p = 3$

- Intelligence Test Scores: `Creativity`, `Mechanical Reasoning`, `Abstract Reasoning`, `Mathematics`
  $\Rightarrow q = 4$

We are going to carry out a canonical correlation analysis using `R`

Canonical Correlation Analysis

CLEMS☙N
U N I V E R S I T Y

Background

Canonical Variates & Canonical Correlations

Sales Data Example

10.10

Notes

Likelihood Ratio Test: Is CCA Worthwhile?

Let's first determine if there is any relationship between the two sets of variables at all.

```
rho <- cc(sales, intelligence)$cor
n <- dim(sales)[1]
p <- length(sales); q <- length(intelligence)
## Calculate p-values using the F-approximations
library(CCP)
p.asym(rho, n, p, q, tstat = "Wilks")
```

H_0	Approximate F value	p-value
$\rho_1^* = \rho_2^* = \rho_3^* = 0$	87.39	~ 0
$\rho_2^* = \rho_3^* = 0$	18.53	8.25×10^{-14}
$\rho_3^* = 0$	3.88	0.028

All three canonical variate pairs are significantly correlated and dependent on one another. This suggests that we may summarize all three pairs.

Canonical Correlation Analysis

CLEMS☙N
U N I V E R S I T Y

Background

Canonical Variates & Canonical Correlations

Sales Data Example

10.11

Notes

Estimates of Canonical Correlation

Since we rejected the hypotheses of independence, the next step is to obtain estimates of canonical correlation

```
cc1 <- cc(sales, intelligence)
cc1$cor
```

i	Canonical Correlation (ρ_i^*)	ρ_i^{*2}
1	0.9945	0.9890
2	0.8781	0.7711
3	0.3836	0.1472

98.9% of the variation in U_1 is explained by the variation in V_1, 77.11% of the variation in U_2 is explained by V_2, only 14.72% of the variation in U_3 is explained by V_3

Canonical Correlation Analysis

CLEMS☙N
U N I V E R S I T Y

Background

Canonical Variates & Canonical Correlations

Sales Data Example

10.12

Notes

Obtain the Canonical Coefficients

	U_1	U_2	U_3
Growth	0.0624	-0.1741	-0.3772
Profit	0.0209	0.2422	0.1035
New	0.0783	-0.2383	0.3834

The first canonical variable for sales is

$$U_1 = 0.0624X_{growth} + 0.0209X_{profit} + 0.0783X_{new}$$

	V_1	V_2	V_3
Creativity	0.0697	-0.1924	0.2466
Mechanical	0.0307	0.2016	-0.1419
Abstract	0.08956	-0.4958	-0.2802
Math	0.0628	0.0683	-0.0113

The first canonical variable for test scores is

$$V_1 = 0.0697Y_{create} + 0.0307Y_{mech} + 0.0896Y_{abstract} + 0.0628Y_{math}$$

Canonical
Correlation
Analysis

CLEMSON
U N I V E R S I T Y

Background

Canonical Variates
& Canonical
Correlations

Sales Data
Example

10.13

Notes

Correlations Between Each Variable and The Corresponding Canonical Variate

Correlations Between X's and U's

	U_1	U_2	U_3
Growth	0.9799	0.0006	-0.1996
Profit	0.9464	0.3229	0.0075
New	0.9519	-0.1863	0.2434

Correlations Between Y's and V's

	V_1	V_2	V_3
Creativity	0.6383	-0.2157	0.6514
Mechanical	0.7212	0.2376	-0.0677
Abstract	0.6472	-0.5013	-0.5742
Math	0.9441	0.1975	-0.0942

Canonical
Correlation
Analysis

CLEMSON
U N I V E R S I T Y

Background

Canonical Variates
& Canonical
Correlations

Sales Data
Example

10.14

Notes

Correlations Between Each Set of Variables and The Opposite Group of Canonical Variates

Correlations Between X's and V's

	V_1	V_2	V_3
Growth	0.9745	0.0006	-0.0766
Profit	0.9412	0.2835	0.0029
New	0.9466	-0.1636	0.0934

Correlations Between Y's and U's

	U_1	U_2	U_3
Creativity	0.6348	-0.1894	0.2499
Mechanical	0.7172	0.2086	-0.0260
Abstract	0.6437	-0.4402	-0.2203
Math	0.9389	0.1735	-0.0361

Canonical
Correlation
Analysis

CLEMSON
U N I V E R S I T Y

Background

Canonical Variates
& Canonical
Correlations

Sales Data
Example

10.15

Notes

Summary

Concepts to know:

- The main idea of canonical correlation analysis (CCA)

- How to compute the canonical variates from the data

- How to determine the number of significant canonical variate pairs

- How to use the results of CCA to describe the relationships between two sets of variables

R functions to know

- `cc` from the `CCA` library
- `p.asym` from the `CCP` library

In the next lecture, we will learn about Classification

Canonical Correlation Analysis

CLEMSON
UNIVERSITY

Background

Canonical Variates & Canonical Correlations

Sales Data Example

10.16

Notes

Notes

Notes