

# Lecture 11

## Canonical Correlation Analysis

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

DSA 8070 Multivariate Analysis

Whitney Huang  
Clemson University

Canonical  
Correlation  
Analysis

CLEMSON  
UNIVERSITY

Background  
Canonical Variates  
& Canonical  
Correlations  
Sales Data  
Example

11.1

Notes

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

### Agenda

- 1 Background
- 2 Canonical Variates & Canonical Correlations
- 3 Sales Data Example

Canonical  
Correlation  
Analysis

CLEMSON  
UNIVERSITY

Background  
Canonical Variates  
& Canonical  
Correlations  
Sales Data  
Example

11.2

### Overview

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

RELATIONS BETWEEN TWO SETS OF VARIATES\*.

By HAROLD HOTELLING, Columbia University.

CONTENTS.	
SECT.	PAGE
1. The Correlation of Vectors. The Most Predictable Criterion and the Tetrad Difference . . . . .	291
2. Theorems on Determinants and Matrices . . . . .	295
3. Canonical Variates and Canonical Correlations. Applications to Algebra and Geometry . . . . .	329
4. Vector Correlation and Aliation Coefficients . . . . .	338
5. Standard Errors . . . . .	339
6. Examples, and an Iterative Method of Solution . . . . .	342
7. The Vector Correlation as a Product of Correlations or of Cosines . . . . .	349
8. An Exact Sampling Distribution of $q$ . . . . .	352
9. Moments of $q$ . The Distribution for Large Samples . . . . .	354
10. The Distribution for Small Samples. Form of the Frequency Curve . . . . .	359
11. Tests for Complete Independence . . . . .	363
12. Alternants of a Plane and of a Sample . . . . .	365
13. The Bivariate Distribution for Complete Independence ( $n=2$ , $n=4$ ) . . . . .	369
14. Theorem on Circularly Distributed Variates . . . . .	371
15. Generalization of Section 13 for Samples of Any Size . . . . .	372
16. Further Problems . . . . .	373

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

CLEMSON  
UNIVERSITY

Background  
Canonical Variates  
& Canonical  
Correlations  
Sales Data  
Example

11.3

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

Canonical  
Correlation  
Analysis

CLEMSON  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.4

Notes

---

---

---

---

---

---

---

---

Linear Combinations of Two Sets of Variables

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

Similar to PCA, we define a set of linear combinations

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

and

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

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

Canonical  
Correlation  
Analysis

CLEMSON  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.5

Notes

---

---

---

---

---

---

---

---

Defining Canonical Variates

We call  $(U_i, V_i)$  be the  $i^{th}$  **canonical variate** pair. One can compute the variance of  $U_i$  with the following expression:

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

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

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

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  $i^{th}$  canonical variate pair is simply the correlation between  $U_i$  and  $V_i$ :

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

Canonical  
Correlation  
Analysis

CLEMSON  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.6

Notes

---

---

---

---

---

---

---

---

## 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  $a_{11}, a_{12}, \dots, a_{1p}$  and  $b_{11}, b_{12}, \dots, b_{1q}$  are chosen to **maximize the canonical correlation**  $\rho_1^*$ . 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  $a_{21}, a_{22}, \dots, a_{2p}$  and  $b_{21}, b_{22}, \dots, b_{2q}$  that maximize  $\rho_2^*$  under the following constraints:

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

This procedure is repeated for each pair of canonical variates



Notes

---

---

---

---

---

---

---

---

## Finding Canonical Variates Cont'd

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

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

The  $i^{\text{th}}$  pair of canonical variates is given by

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

where

- $\mathbf{u}_i$  is the  $i^{\text{th}}$  eigenvector of  $\Sigma_{\mathbf{X}}^{-1/2} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{X}}^{-1/2}$
- $\mathbf{v}_i$  is the  $i^{\text{th}}$  eigenvector of  $\Sigma_{\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{X}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1/2}$
- The  $i^{\text{th}}$  canonical correlation is given by,  $\text{Cor}(U_i, V_i) = \rho_i^*$ , where  $\rho_i^{*2}$  is the  $i^{\text{th}}$  eigenvalue of  $\Sigma_{\mathbf{X}}^{-1/2} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{X}}^{-1/2}$



Notes

---

---

---

---

---

---

---

---

## Likelihood Ratio Test: Is CCA Worthwhile?

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

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

$$-2 \log(\Lambda) = n \log \left( \frac{|\hat{\Sigma}_{\mathbf{X}}| |\hat{\Sigma}_{\mathbf{Y}}|}{|\hat{\Sigma}|} \right) = -n \sum_{j=1}^p \log(1 - \hat{\rho}_j^{*2})$$

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

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

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



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

CLEMSON  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.10

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	$\sim 0$
$\rho_2^* = \rho_3^* = 0$	18.53	$8.25 \times 10^{-14}$
$\rho_3^* = 0$	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

CLEMSON  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.11

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 ( $\rho_i^*$ )	$\rho_i^{*2}$
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

CLEMSON  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.12

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  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.13

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  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.14

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  
UNIVERSITY

Background

Canonical Variates  
& Canonical  
Correlations

Sales Data  
Example

11.15

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

11.16

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---