Canonical Correlation Analysis

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Canonical Correlation Analysis

1. Introduction
2. Exploring Redundancy in Sets of Variables
   - An Example – Personality and Achievement
3. Basic Properties of Canonical Variates
4. Calculating Canonical Variates
   - The Fundamental Result
   - The Geometric View
   - Different Kinds of Canonical Weights
   - Partially Standardized Weights
   - Fully Standardized Weights
5. A Simple Example
   - The Data
   - Basic Calculations in R
   - Partially Standardized Weights
   - Fully Standardized Weights
6. A Canonical Correlation Function
7. Some Examples
   - UCLA Academics Data
   - Work Satisfaction Data
   - Health Club Data

Introduction

Previously, we studied factor analytic methods as an approach to understanding the key sources of variation within sets of variables.

There are situations in which we have several sets of variables, and we seek an understanding of key dimensions that are correlated across sets.

Canonical correlation analysis is the one of the oldest and best known methods for discovering and exploring dimensions that are correlated across sets, but uncorrelated within set.

The relationship between personality and achievement is of interest.

Suppose the **x** variables are a set of personality scale scores, and the **y** variables are a set of academic achievement scores.

Then the first canonical variate in each set will isolate dimensions of personality and achievement that predict each other well.

Basic Properties of Canonical Variates

Canonical Correlation Analysis (CCA) is, in a sense, a combination of the ideas of principal component analysis and multiple regression.

In CCA, we have two sets of variables, **x** and **y**, and we seek to understand what aspects of the two sets of variables are redundant.

The CCA approach seeks to find *canonical variates*, linear combinations of the variables in **x** and **y**.

There are different canonical variates within each set. If there are q_1 variables in **x** and q_2 variables in **y**, then there are at most $k = \min(q_1, q_2)$ canonical variates in either set. These are $u_i = \mathbf{a}_i'\mathbf{x}$, and $v_i = \mathbf{b}_i'\mathbf{y}$, with i ranging from 1 to k.

Basic Properties of Canonical Variates

Within each set, the k distinct canonical variates are uncorrelated. Across each set, $\mathbf{u}_i$ and $\mathbf{v}_j$ are uncorrelated, unless $i = j$.

The correlation between corresponding canonical variates $\mathbf{u}_i$ and $\mathbf{v}_i$ is the ith *canonical correlation*.

An alternate view of the *first* canonical variate is that it is the linear combination of variables in one set that has the highest possible multiple correlation with the variables in the other set.

Calculating Canonical Variates

Defining the canonical variates is tantamount to deriving expressions for $\mathbf{a}_i$ and $\mathbf{b}_i$.

Clearly, since correlations are invariant under linear transformations, there are infinitely many ways we might define canonical variates.

It is important to realize that textbooks, in general, are very confused (or at least very confusing) in their treatments of canonical correlation.

In particular, there are different meanings of the same term, depending on which book you read.

Calculating Canonical Variates

The Fundamental Result

A number of textbooks books derive the fact that the linear weights producing canonical variates with maximum possible correlation can be computed as an eigenvector problem.

Specifically, $\mathbf{a}_i$ may be computed as the ith eigenvector of $\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$.

The squared canonical correlation r_i^2 is the corresponding eigenvalue. Likewise, $\mathbf{b}_i$ is the ith eigenvector of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$.

Calculating Canonical Variates

The Geometric View

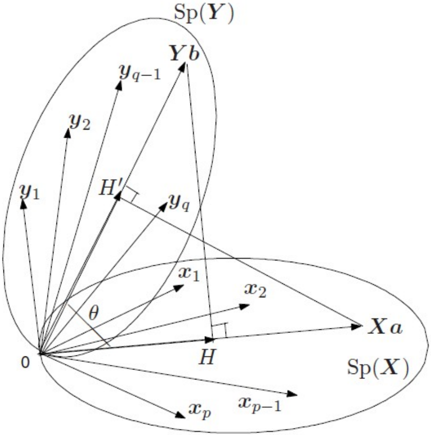

Figure 6.2: Vector representation of canonical correlation analysis. (The vectors $\overrightarrow{0H'}$ and $\overrightarrow{0H}$ are, respectively, $\boldsymbol{P}_Y\boldsymbol{Xa}$ and $\boldsymbol{P}_X\boldsymbol{Yb}$, and the angle between the two vectors is designated as θ.)

Calculating Canonical Variates

Different Kinds of Canonical Weights

You don't have to look at many textbook presentations of canonical correlation to realize that the canonical weights presented do not necessarily agree with those produced by various computer programs.

In some cases, the discrepancies are the result of error, but you should also be aware that there are several different kinds of canonical weights:

- *Completely Raw*. These weights are, in fact, the eigenvectors described on the previous slide, computed from the covariance matrices.
- *Partially Standardized.* These weights are multiplied by a constant, so the the resulting canonical variates have unit variance.
- *Fully Standardized.* These weights are computed on standardized variables (i.e., correlation matrices), then multiplied by a constant so that the resulting canonical variates have unit variance.

Calculating Canonical Variates

Partially Standardized Weights

Let $\mathbf{A}$ and $\mathbf{B}$ contain the raw canonical weights obtained via eigenvector decompositions.

Then the canonical variates are $\mathbf{U} = \mathbf{XA}$ and $\mathbf{V} = \mathbf{YB}$. To standardize the canonical variates, we recall that $\text{Var}(\mathbf{U}) = \mathbf{A}'\mathbf{S}_{xx}\mathbf{A}$, and $\text{Var}(\mathbf{V}) = \mathbf{B}'\mathbf{S}_{yy}\mathbf{B}$.

Consequently, we need only postmultiply $\mathbf{U}$ and $\mathbf{V}$ by the symmetric inverse square root of their covariance matrices.

Calculating Canonical Variates

Partially Standardized Weights

Thus, we have

$$\begin{aligned}\mathbf{U}^* &= \mathbf{XA}(\mathbf{A}'\mathbf{S}_{xx}\mathbf{A})^{-1/2}\\ \mathbf{V}^* &= \mathbf{YB}(\mathbf{B}'\mathbf{S}_{yy}\mathbf{B})^{-1/2}\end{aligned}$$

which may be expressed as $\mathbf{U}^* = \mathbf{XA}^*$, $\mathbf{V}^* = \mathbf{YB}^*$, with

$$\begin{align}\mathbf{A}^* &= \mathbf{A}(\mathbf{A}'\mathbf{S}_{xx}\mathbf{A})^{-1/2} \notag\\ \mathbf{B}^* &= \mathbf{B}(\mathbf{B}'\mathbf{S}_{yy}\mathbf{B})^{-1/2} \tag{1}\\ & \tag{2}\end{align}$$

To add to the confusion, SAS refers to these partially standardized weights as "raw canonical weights."

Calculating Canonical Variates

Fully Standardized Weights

In fully standardized canonical correlation analysis, we operate on Z scores instead of raw scores for both $\mathbf{x}$ and $\mathbf{y}$ variables.

In score notation, the canonical weights $\mathbf{A}_s$ and $\mathbf{B}_s$ are the first k eigenvectors of $\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$ and $\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$, respectively, restandardized as in the previous slide.

The canonical variate scores themselves are obtained by applying the canonical weights to $\mathbf{Z}_x$ and $\mathbf{Z}_y$, the sample Z-scores. SAS refers to these weights as the "standardized weights."

A Simple Example

The Data

Suppose we have an **X** and **Y** given by

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 3 & 2 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \\ 2 & 2 & 3 \\ 3 & 3 & 2 \\ 1 & 3 & 2 \\ 4 & 3 & 5 \\ 5 & 5 & 5 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 4 & 4 & -1.07846 \\ 3 & 3 & 1.214359 \\ 2 & 2 & 0.307180 \\ 2 & 3 & -0.385641 \\ 2 & 1 & -0.078461 \\ 1 & 1 & 1.61436 \\ 1 & 2 & 0.814359 \\ 2 & 1 & -0.0641016 \\ 1 & 2 & 1.535900 \end{pmatrix} \tag{3}$$

A Simple Example

The Data

In this highly artificial example, I constructed the third column of **Y** from the columns of **X** with the linear weights $\mathbf{a}_1' = [.4, .6, -\sqrt{.48}]$.

Here are some questions:

- What should the first vector of canonical weights for the **Y** variates be?
- What should the first canonical correlation be?

A Simple Example

The Data

To answer the two questions on the preceding slide, recall that the purpose of canonical correlation analysis is to (a) *find* and (b) *characterize* the linear redundancy between two sets of variates.

In our simple example, one of the variates in **Y** can be reproduced exactly as a linear combination of the three variates in **X**.

Canonical correlation analysis (if it is working properly) will simply select y_3 as the first canonical variate in the **Y** set, with canonical weights $\mathbf{b}_1' = [001]$, and recover the linear combination of the variables in the first group that was used to generate $\mathbf{y}_3$ by giving $\mathbf{a}_1' = [.4, .6, -\sqrt{.48}]$ as the canonical weights for the **X** set.

The first canonical correlation will, of course, be 1.

A Simple Example

Basic Calculations in R

We have discussed three different ways of performing canonical correlation analysis:

- *Completely Raw*.
- *Partially Standardized.*
- *Fully Standardized.*

Let's perform the calculations in R.

We'll start with the "Completely Raw" calculation.

A Simple Example

Basic Calculations in R

First, we download necessary data and utility routines, which establish variable sets X and Y for further analysis.

```
> source("http://www.statpower.net/R312/Steiger R Library Functions.txt")
> source("http://www.statpower.net/R312/Data 1.txt")
> X

      [,1] [,2] [,3]
 [1,]    1    1    3
 [2,]    2    3    2
 [3,]    1    1    1
 [4,]    1    1    2
 [5,]    2    2    3
 [6,]    3    3    2
 [7,]    1    3    2
 [8,]    4    3    5
 [9,]    5    5    5

> Y

      [,1] [,2]     [,3]
 [1,]    4    4 -1.07846
 [2,]    3    3  1.21436
 [3,]    2    2  0.30718
 [4,]    2    3 -0.38564
 [5,]    2    1 -0.07846
 [6,]    1    1  1.61436
 [7,]    1    2  0.81436
 [8,]    2    1 -0.06410
 [9,]    1    2  1.53590
```

A Simple Example

Basic Calculations in R

To calculate the completely raw weights, we need the variance-covariance matrices for **X** and **Y**, as well as the cross-covariance matrices.

```
> S.xy <- cov(X, Y)
> S.xx <- var(X)
> S.yx <- cov(Y, X)
> S.yy <- var(Y)
```

Now that we have these matrices, it is easy to calculate the “completely raw” canonical weights and canonical correlations in R.

```
> A <- eigen(solve(S.xx) %*% S.xy %*% solve(S.yy) %*% S.yx)$vectors
> B <- eigen(solve(S.yy) %*% S.yx %*% solve(S.xx) %*% S.xy)$vectors
> R <- sqrt(eigen(solve(S.yy) %*% S.yx %*% solve(S.xx) %*%
+     S.xy)$values)
```

A Simple Example

Basic Calculations in R

The resulting weights for the first canonical variates are what we expected, and the first canonical correlation is 1.

```
> A

        [,1]    [,2]    [,3]
[1,]  0.4000  0.7961 -0.5776
[2,]  0.6000 -0.5838  0.4286
[3,] -0.6928 -0.1597  0.6947

> B

              [,1]     [,2]    [,3]
[1,]  0.0000001941  0.53653  0.8348
[2,] -0.0000004336 -0.84377 -0.1386
[3,]  1.0000000000 -0.01364  0.5329

> R

[1] 1.00000 0.51938 0.09103
```

A Simple Example

Partially Standardized Weights

To standardize the weights so that the canonical variances have variances of 1, we need to apply the correction shown earlier.

```
> ## Singly standardized weights (SAS 'raw')
> A.single <- A %*% solve(sqrt(diag(diag(var(X %*% A)))))
> B.single <- B %*% solve(sqrt(diag(diag(var(Y %*% B)))))
> A.single

        [,1]    [,2]    [,3]
[1,]  0.4324  1.4468 -0.8180
[2,]  0.6485 -1.0610  0.6070
[3,] -0.7489 -0.2902  0.9838

> B.single

              [,1]     [,2]    [,3]
[1,]  0.0000002098  0.84865  1.5200
[2,] -0.0000004686 -1.33462 -0.2524
[3,]  1.0809120704 -0.02158  0.9702
```

A Simple Example

Fully Standardized Weights

To compute fully standardized weights, we need to calculate Z-scores for our data.

We begin by using the $\mathbf{Q}$ operator to convert the scores into deviation scores.

Recall that we learned that $\mathbf{Q}_1$, the complementary orthogonal projector for a vector of 1's, will convert a column of scores into deviation score form. The R library functions include a `UnitVector` function and a `Q` function that make this easy.

```
> ## Deviation score X,Y
> X.dev <- Q(UnitVector(9)) %*% X
> Y.dev <- Q(UnitVector(9)) %*% Y
```

A Simple Example

Fully Standardized Weights

To convert the deviation scores to Z-scores, we multiply each column by the inverse standard deviation of the scores in that column.

There are lots of ways we can do this. I'm using the matrix algebra approach of post-multiplying by a diagonal matrix with diagonal entries equal to the inverse standard deviation.

```
> ## Z-score X,Y Create diagonal matrices with standard
> ## deviations Then invert using solve
> D.x <- solve(sqrt(diag(diag(var(X)))))
> D.y <- solve(sqrt(diag(diag(var(Y)))))
> ## Postmultiply the deviation score matrix to create
> ## Z-scores
> Z.x <- X.dev %*% D.x
> Z.y <- Y.dev %*% D.y
```

A Simple Example

Fully Standardized Weights

Finally, we apply the identical method used to compute the singly standardized ("SAS Raw") canonical variates, except that we use Z-scores and correlation matrices instead of raw scores and covariance matrices.

```
> R.xy <- cor(X, Y)
> R.xx <- cor(X)
> R.yx <- cor(Y, X)
> R.yy <- cor(Y)
> A.s <- eigen(solve(R.xx) %*% R.xy %*% solve(R.yy) %*% R.yx)$vectors
> B.s <- eigen(solve(R.yy) %*% R.yx %*% solve(R.xx) %*% R.xy)$vectors
> A.fully <- A.s %*% solve(sqrt(diag(diag(var(Z.x %*% A.s)))))
> B.fully <- B.s %*% solve(sqrt(diag(diag(var(Z.y %*% B.s)))))
```

A Simple Example

Fully Standardized Weights

```
> A.fully

        [,1]    [,2]    [,3]
[1,]  0.6405  2.1432 -1.2118
[2,]  0.8647 -1.4146  0.8093
[3,] -1.0443 -0.4046  1.3719

> B.fully

              [,1]     [,2]    [,3]
[1,]  0.0000002098  0.84865  1.5200
[2,] -0.0000004940 -1.40682 -0.2660
[3,]  0.9999999345 -0.01996  0.8976
```

A Canonical Correlation Function

I put together the calculations for canonical correlation in a library function called `CanCorr.r`. Let's load it in and try it on the X and Y data. I store the output in an object called `output` so that I can examine the results piece-by-piece.

```
> source("http://www.statpower.net/R312/CanCorr.r")
> ## Analyze
> output <- canonical.cor(X, Y)
```

A Canonical Correlation Function

Let's start by examining the canonical correlations and the significance tests that accompany them.

```
> output[1]

$`Canonical Correlations`
     Canonical R Wilk's Lambda            F df1   df2
[1,]     1.00000     2.026e-13 136016.33779   9 7.452
[2,]     0.51938     7.242e-01      0.35019   4 8.000
[3,]     0.09103     9.917e-01      0.04178   1 5.000
       p value
[1,] 1.580e-18
[2,] 8.370e-01
[3,] 8.461e-01
```

In this case, the first canonical correlation is overwhelmingly significant, but neither of the additional two canonical correlations is significant.

A Canonical Correlation Function

We print the singly standardized (SAS "Raw") canonical weights. These can be interpreted much like the factor loadings from a factor analysis of a covariance matrix. We see, in particular, is that the first canonical variate on the Y side is almost precisely colinear with Y_3.

```
> output[2:3]

$`X (SAS) Raw Weights`
        [,1]    [,2]    [,3]
[1,]  0.4324  1.4468  0.8180
[2,]  0.6485 -1.0610 -0.6070
[3,] -0.7489 -0.2902 -0.9838

$`Y (SAS) Raw Weights`
              [,1]     [,2]    [,3]
[1,]  0.0000002098  0.84865  1.5200
[2,] -0.0000004686 -1.33462 -0.2524
[3,]  1.0809120704 -0.02158  0.9702
```

A Canonical Correlation Function

Next come the fully standardized weights

```
> output[4:5]

$`X Fully Standardized Weights`
        [,1]    [,2]    [,3]
[1,]  0.6405  2.1432  1.2118
[2,]  0.8647 -1.4146 -0.8093
[3,] -1.0443 -0.4046 -1.3719

$`Y Fully Standardized Weights`
              [,1]     [,2]    [,3]
[1,]  0.0000002098  0.84865  1.5200
[2,] -0.0000004940 -1.40682 -0.2660
[3,]  0.9999999345 -0.01996  0.8976
```

A Canonical Correlation Function

For comparison to other software, the `canonical.cor` function also prints Canonical Loadings, the correlations between the observed variables and the canonical variables.

```
> output[6:7]

$`X Canonical Loadings`
          [,1]   [,2]    [,3]
[1,]  0.508428 0.6402 -0.5758
[2,]  0.772114 0.1219 -0.6237
[3,] -0.006404 0.4936 -0.8696

$`Y Canonical Loadings`
        [,1]          [,2]          [,3]
[1,] -0.6630 -0.1390795978 0.73556069686
[2,] -0.4142 -0.7947228961 0.44372120723
[3,]  1.0000 -0.0000003634 0.00000006489
```

Rencher (his section 11.5.2) argues against using the loadings as an aid to interpretation.

Some Examples

UCLA Academics Data

Next, we examine an example from the UCLA Statistics website.

```
> ## grab UCLA data
>
> mm <- read.csv("http://www.statpower.net/R312/UCLACCData.txt")
> attach(mm)
> X <- mm[, 1:3]
> Y <- mm[, 4:8]
>
> ## Analyze
> output <- canonical.cor(X, Y)
```

Some Examples

UCLA Academics Data

```
> output[1]

$`Canonical Correlations`
     Canonical R Wilk's Lambda      F df1  df2
[1,]      0.4641        0.7544 11.716  15 1635
[2,]      0.1675        0.9614  2.944   8 1186
[3,]      0.1040        0.9892  2.165   3  594
       p value
[1,] 7.498e-28
[2,] 2.905e-03
[3,] 9.109e-02
```

Some Examples

UCLA Academics Data

```
> output[4:5]

$`X Fully Standardized Weights`
                    [,1]    [,2]    [,3]
locus_of_control  0.8404  0.4166  0.4435
self_concept     -0.2479  0.8379 -0.5833
motivation        0.4327 -0.6948 -0.6855

$`Y Fully Standardized Weights`
           [,1]     [,2]     [,3]
read    0.45080  0.04961 -0.21601
write   0.34896 -0.40921 -0.88810
math    0.22047 -0.03982 -0.08848
science 0.04878  0.82660  1.06608
female  0.31504 -0.54057  0.89443
```

Some Examples

Work Satisfaction Data

Here's another!

```
> ## grab Work Satisfaction data
> worksat <- read.csv("http://www.statpower.net/R312/worksat.csv")
> names(worksat)

 [1] "ID"
 [2] "SupervisorSatisfaction.Y1."
 [3] "CareerFutureSatisfaction.Y2."
 [4] "FinancialSatisfaction.Y3."
 [5] "WorkloadSatisfaction.Y4."
 [6] "CompanyIdentification.Y5."
 [7] "WorkTypeSatisfaction.Y6."
 [8] "GeneralSatisfaction.Y7."
 [9] "FeedbackQuality.X1."
[10] "TaskSignificance.X2."
[11] "TaskVariety.X3."
[12] "TaskIdentity.X4."
[13] "Autonomy.X5."
```

Here's another example. You try it!

```
> ## grab Work Satisfaction data
> health <- read.csv("http://www.statpower.net/R312/HealthClub.csv")
> names(health)

[1] "Weight" "Waist"  "Pulse"  "Chins"  "Situps"
[6] "Jumps"
```