Canonical Correlation Analysis (CCA)

Alireza Sheikh-Zadeh, Ph.D.

Principal components analysis considers interrelationships *within a set of variables*. But there are situations where the researcher may be interested in assessing the relationships *between two sets of variables*.

For example, in psychology, an investigator may measure a number of aptitude variables and a number of achievement variables on a sample of students and wish to say something about the relationship between "aptitude" and "achievement". One technique for addressing such questions is *canonical correlation analysis*. A typical use for canonical correlation in the experimental context is to take two sets of variables and see what is common amongst the two sets.

In canonical correlation analysis where there is more than a single variable in each of the two sets, the objective is to find the linear functions of the variables in one set that maximally correlate with linear functions of variables in the other set.

In a multiple regression analysis, a single variable Y is related to two or more variables $X_1, X_2, \ldots, X_q$, to see how Y is related to the X variables. From this point of view canonical correlation is a generalization of multiple regression in which several Y variables are simultaneously related to several X variables.

In practice, more than one pair of canonical variables can be calculated from a set of data. If there are q variables $X_1, X_2, \ldots, X_q$ and p variables $Y_1, Y_2, \ldots, Y_p$, then CCA finds up to $min(q, p)$ pairs of variables.

$$U_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1q}X_q$$

$$U_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2q}X_q$$

.

.

$$U_q = a_{q1}X_1 + a_{q2}X_2 + \ldots + a_{qq}X_q$$

and

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1p}Y_p$$

$$V_2 = b_{21}Y_1 + b_{22}Y_2 + \ldots + b_{2p}Y_p$$

.

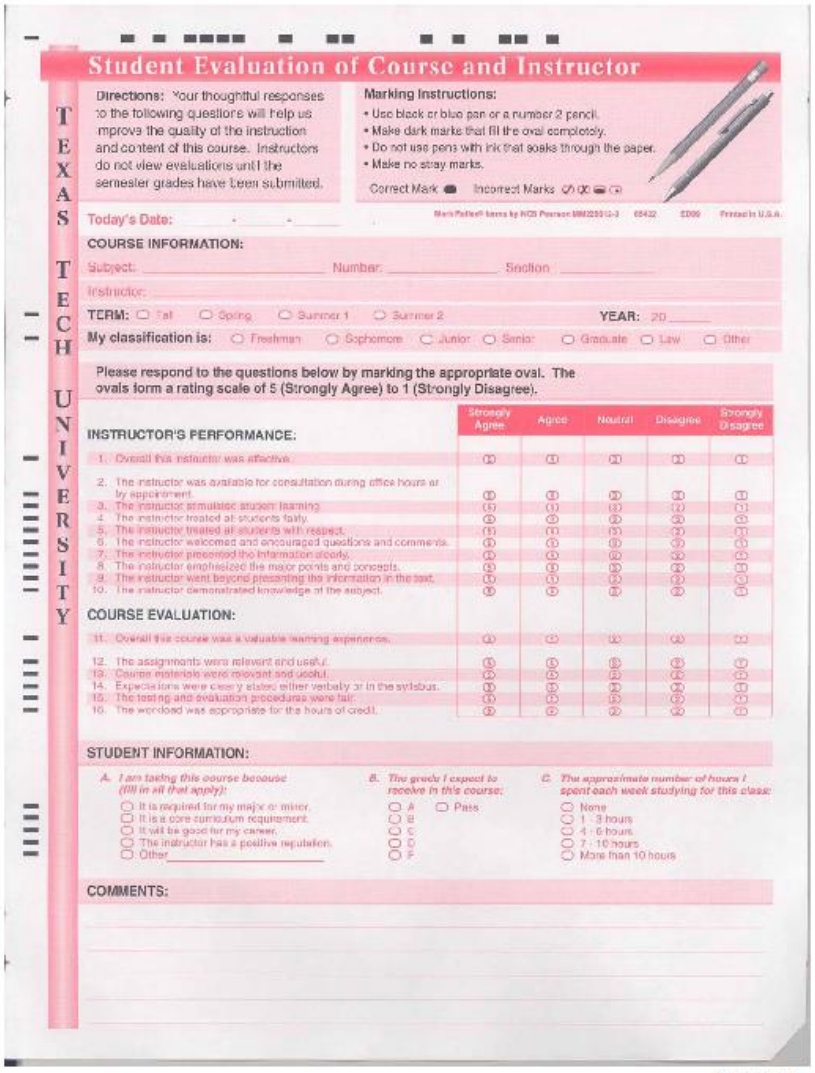
.

$$V_p = b_{p1}Y_1 + b_{p2}Y_2 + \ldots + b_{pp}Y_p$$

(U_1, V_1), (U_2, V_2), …, $(U_{min\{p,q\}}, V_{min\{p,q\}})$ are chosen as pairs of canonical variables so that the correlation between U_1 and V_1 is a maximum; the correlation between U_2 and V_2 is a maximum, subject to these variables being uncorrelated with U_1 and V_1; the correlation between U_3 and V_3 is a maximum, subject to these variables being uncorrelated with U_1, V_1, U_2, and V_2; and so on.

Example: TTU student evaluations

See here for an explanation of the 16 variables.



Student Evaluation of Course and Instructor

TEXAS TECH UNIVERSITY

Directions: Your thoughtful responses to the following questions will help us improve the quality of the instruction and content of this course. Instructors do not view evaluations until the semester grades have been submitted.

Marking Instructions:
- Use black or blue pen or a number 2 pencil.
- Make dark marks that fill the oval completely.
- Do not use pens with ink that soaks through the paper.
- Make no stray marks.

Correct Mark ● Incorrect Marks ⊘ ⊘ ⊘

Today's Date: ____ - ____ - ____

COURSE INFORMATION:

Subject: _____ Number: _____ Section _____

Instructor: _____

TERM: ○ Fall ○ Spring ○ Summer 1 ○ Summer 2 YEAR: 20____

My classification is: ○ Freshman ○ Sophomore ○ Junior ○ Senior ○ Graduate ○ Law ○ Other

Please respond to the questions below by marking the appropriate oval. The ovals form a rating scale of 5 (Strongly Agree) to 1 (Strongly Disagree).

INSTRUCTOR'S PERFORMANCE:	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
1. Overall this instructor was effective.	⑤	④	③	②	①
2. The instructor was available for consultation during office hours or by appointment.	⑤	④	③	②	①
3. The instructor stimulated student learning.	⑤	④	③	②	①
4. The instructor treated all students fairly.	⑤	④	③	②	①
5. The instructor treated all students with respect.	⑤	④	③	②	①
6. The instructor welcomed and encouraged questions and comments.	⑤	④	③	②	①
7. The instructor presented the information clearly.	⑤	④	③	②	①
8. The instructor emphasized the major points and concepts.	⑤	④	③	②	①
9. The instructor went beyond presenting the information in the text.	⑤	④	③	②	①
10. The instructor demonstrated knowledge of the subject.	⑤	④	③	②	①
COURSE EVALUATION:					
11. Overall this course was a valuable learning experience.	⑤	④	③	②	①
12. The assignments were relevant and useful.	⑤	④	③	②	①
13. Course materials were relevant and useful.	⑤	④	③	②	①
14. Expectations were clearly stated either verbally or in the syllabus.	⑤	④	③	②	①
15. The testing and evaluation procedures were fair.	⑤	④	③	②	①
16. The workload was appropriate for the hours of credit.	⑤	④	③	②	①

STUDENT INFORMATION:

A. *I am taking this course because (fill in all that apply):*
- ☐ It is required for my major or minor.
- ☐ It is a core curriculum requirement.
- ☐ It will be good for my career.
- ☐ The instructor has a positive reputation.
- ☐ Other _____

B. *The grade I expect to receive in this course:*
- ○ A ○ Pass
- ○ B
- ○ C
- ○ D
- ○ F

C. *The approximate number of hours I spent each week studying for this class:*
- ○ None
- ○ 1 - 3 hours
- ○ 4 - 6 hours
- ○ 7 - 10 hours
- ○ More than 10 hours

COMMENTS:

Attachment B
OP 32.32
8/29/07

The first ten variables are associated with an instructor's performance and the next six variables are related to course evaluation.

```
data <- read.csv("http://tiny.cc/isqs6350_evals")
evals <- data[,3:18]    # select variables to use
evals <- na.omit(evals) # do listwise cleaning for missing values
```

Use canonical correlation analysis to find the linear combinations for two sets of variables (`X` and `Y`) that maximize the correlation and interpret them.

```
#install.packages("CCA")

require(CCA)

X <- evals[, 1:10]  # instructor's performance

Y <- evals[, 11:16] # course evaluation

cca <- cc(X,Y) #CCA function
```

Correlation between the best six pair are as follow.

```
round(cca$cor, 3)

## [1] 0.867 0.312 0.176 0.095 0.094 0.035
```

For instance, $Cor(U_1, V_1) = 0.867$, and $Cor(U_2, V_2) = 0.312$.

First we want to see what are the coefficients of X_1, X_2, ..., X_{10} variables for the vectors U_1, U_2, ..., U_6 (because min{10, 6} = 6).

```
a <- cca$xcoef
round(a, 3)

##           [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
## RESP_1  -0.302  1.169  0.255 -0.878  0.720 -0.288
## RESP_2  -0.116 -0.553  0.921  0.388 -0.744 -0.248
## RESP_3  -0.130  0.891  0.151  0.726 -0.319 -0.983
## RESP_4  -0.151 -1.079 -0.282 -0.254  1.293 -1.048
## RESP_5  -0.009  0.163 -1.127  0.971 -1.249  0.118
## RESP_6  -0.044 -0.269 -0.179 -1.286 -0.262  1.060
## RESP_7  -0.147 -0.127 -0.755 -0.486 -0.244 -0.203
## RESP_8  -0.230 -0.764  0.480  0.891  1.408  0.608
## RESP_9  -0.086  0.303 -0.180  0.716 -0.359  1.229
## RESP_10 -0.192 -0.287  0.928 -1.133 -0.627 -0.264
```

For instance

$$U_1 = -0.3025X_1 - 0.1156X_2 - \ldots - 0.1919X_{10}$$

U_1 is a new variable like PC1 in PCA analysis.

```
U <- cca$scores$xscores
# U1 scores is the first column of xscores
head(U)
```

```
##         [,1]        [,2]        [,3]      [,4]        [,5]        [,6]
## 1  0.5357973  0.01113195  0.75806943 0.8517536 -0.44567430 -0.25915984
## 2 -0.3075895 -1.03042213  0.54983194 1.4821331 -0.56086046  0.46013053
## 3 -0.4542650 -1.15707677 -0.20480351 0.9963568 -0.80461046  0.25704770
## 4 -0.7567588  0.01225422  0.05016306 0.1188287 -0.08505256 -0.03085868
## 5 -0.7567588  0.01225422  0.05016306 0.1188287 -0.08505256 -0.03085868
## 6  0.6513564  0.56425863 -0.16323286 0.4635944  0.29806634 -0.01107966
```

Then, what are the coefficients of $Y_1, Y_2, \ldots, Y_6$ variables for the vectors $V_1, V_2, \ldots, V_6$.

```
b <- cca$ycoef
round(b,3)
```

```
##           [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
## RESP_11 -0.428  1.391 -0.191 -1.085  0.337  0.058
## RESP_12 -0.095 -0.103  0.953  1.068 -0.017  1.882
## RESP_13 -0.074  0.117  0.293  0.729 -0.579 -2.062
## RESP_14 -0.327 -1.098  0.615 -1.367 -0.828  0.155
## RESP_15 -0.210 -0.538 -0.188  0.368  1.827 -0.348
## RESP_16 -0.142 -0.112 -1.440  0.282 -0.795  0.251
```

For instance

$$V_1 = -0.428Y_1 - 0.095Y_2 - \ldots - 0.142Y_6$$

$$V_1 = 1.39Y_1 - 0.10Y_2 + \ldots - 0.11Y_6$$

V_1 is also a new variable. U_1 is a surrogate variable for set 1 and $V_1$is a surrogate variable for set 2. U_1 and V_1 have the highest possible correlation compared to other possible surrogate variables.

Note: Remember X and Y variables are all scaled.

```
V <- cca$scores$yscores
# V1 and V2 scores are the first column of yscores
head(V)
```

```
##         [,1]       [,2]         [,3]       [,4]         [,5]        [,6]
## 1  0.2101380 1.86259709  1.211523839 -1.2041591  1.140076186 -0.14329019
## 2 -0.7796588 0.01717667  0.004229182  0.1584865 -0.047604897 -0.01511567
## 3 -0.4530626 1.11554578 -0.610584537  1.5253886  0.780278578 -0.16968766
## 4 -0.7796588 0.01717667  0.004229182  0.1584865 -0.047604897 -0.01511567
## 5 -0.7796588 0.01717667  0.004229182  0.1584865 -0.047604897 -0.01511567
## 6  0.4959445 0.36014905 -0.037693261  0.1624909  0.008153618  0.04909489
```

CCA Interpretation

How to interpret the results? We usually focus on the first two pair of canonical variables (U_1, V_1), (U_2, V_2).

First, we interpret U_1 and V_1. For easier interpretation, we can transform all coefficients to something between 0 and 1 (This doesn't hurt the $Cor(U_1, V_1)$ at all.)

```
a <- cca$xcoef

b <- cca$ycoef

a[,1]
```

```
##       RESP_1       RESP_2       RESP_3       RESP_4       RESP_5
## -0.302493818 -0.115559032 -0.130112528 -0.151133136 -0.009419085
##       RESP_6       RESP_7       RESP_8       RESP_9      RESP_10
## -0.044457196 -0.146675487 -0.230007944 -0.086369475 -0.191887447
```

```
a1 <- a[,1]/min(a[,1]) # Just a trick for easier interpretation
round(a1, 3)
```

```
##  RESP_1  RESP_2  RESP_3  RESP_4  RESP_5  RESP_6  RESP_7  RESP_8  RESP_9
##   1.000   0.382   0.430   0.500   0.031   0.147   0.485   0.760   0.286
## RESP_10
##   0.634
```

We can rewrite the transformed U_1:

$$U_1 = 1X_1 + 0.382X_2 + \ldots + 0.634X_{10}$$

This linear combination represents the overall effectiveness of the instructor. If we line up the students based on low to high U_1 score on this combination, the students at the top perceived the instructor very effective. On the other hand, students at the low end will have opposite perceptions, not an effective instructor. By looking at `a1`, scores are mostly dominated by the response to Q1, Q8, Q10. What are these questions? You can interpret U_1 as a representative of those questions. That's how we can attach a name to U_1.

```
b[,1]
```

```
##     RESP_11     RESP_12     RESP_13     RESP_14     RESP_15     RESP_16
## -0.42788627 -0.09535630 -0.07387761 -0.32659624 -0.20980710 -0.14207977
```

```
b1 <- b[,1]/min(b[,1])
round(b1, 3)
```

```
## RESP_11 RESP_12 RESP_13 RESP_14 RESP_15 RESP_16
##   1.000   0.223   0.173   0.763   0.490   0.332
```

$$V_1 = 1Y_1 + 0.22Y_2 + \ldots + Y_6$$

V_1 measures the student's perception of the overall course evaluation. If we line up the students in descending order of this combination score, the front portion of the line perceived the course as an effective course and the back portion of the line perceived it as ineffective. Scores are mostly dominated by the response of the two questions, Q11 and Q14, which focuses on the overall learning experience and clear statement of expectation respectively.

In the result, we see that the Cor(U_1, V_1)= 0.87, which is very high.

```
cca$cor[1]
```

```
## [1] 0.866684
```

This makes sense. Because a more effective instructor is correlated to the overall course evaluation.

Practice: How can we interpret the second pair of canonical variables (U_2, V_2)? Hint: Focus on the bellow highlighted coefficients.

```
a[,2]
```

```
##     RESP_1     RESP_2     RESP_3     RESP_4     RESP_5     RESP_6
##  1.1693310 -0.5531267  0.8906871 -1.0788263  0.1634011 -0.2689470
##     RESP_7     RESP_8     RESP_9    RESP_10
## -0.1266546 -0.7637556  0.3028432 -0.2869566
```

```
a2 <- a[,2]/max(a[,2]) # Just a trick for easier interpretation
round(a2, 3)
```

```
##  RESP_1  RESP_2  RESP_3  RESP_4  RESP_5  RESP_6  RESP_7  RESP_8  RESP_9
##   1.000  -0.473   0.762  -0.923   0.140  -0.230  -0.108  -0.653   0.259
## RESP_10
##  -0.245
```

```
b[,2]
```

```
##    RESP_11    RESP_12    RESP_13    RESP_14    RESP_15    RESP_16
##  1.3908958 -0.1025619  0.1166288 -1.0983691 -0.5380137 -0.1115523
```

```
b2 <- b[,2]/max(b[,2])
round(b2, 3)
```

```
## RESP_11 RESP_12 RESP_13 RESP_14 RESP_15 RESP_16
##   1.000  -0.074   0.084  -0.790  -0.387  -0.080
```