Canonical Correlation Analysis

Dipankar Sutradhar (194102305)

15 June 2020

1 Aim

Analysing Canonical Correlation Analysis that lets us identify associations among groups of variables at a time.

2 Introduction

Canonical correlation analysis is a method for exploring the relationships between two multivariate sets of variables (vectors), all measured on the same individual.

Consider, as an example, variables related to exercise and health. On one hand, you have variables associated with exercise, observations such as the climbing rate on a stair stepper, how fast you can run a certain distance, the amount of weight lifted on bench press, the number of push-ups per minute, etc. On the other hand, you have variables that attempt to measure overall health, such as blood pressure, cholesterol levels, glucose levels, body mass index, etc. Two types of variables are measured and the relationships between the exercise variables and the health variables are of interest.
One approach to studying relationships between the two sets of variables is to use canonical correlation analysis which describes the relationship between the first set of variables and the second set of variables. We do not necessarily think of one set of variables as independent and the other as dependent, though that may potentially be another approach.

3 Advantages of CCA

By using CCA we can :
• Find out whether two sets of variables are independent or, measure the magnitude of their relationship if there is one.
• Interpret the nature of their relationship by assessing each variable's contribution to the canonical variates (i.e. components) and find what dimensions are common between the two sets.
• Summarise relationships into a lesser number of statistics..
• Conduct a dimensionality reduction that takes into account the existence of certain variable groups.

4 Canonical Variates

Given two sets of variables:

$$\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, ...\boldsymbol{X_p})$$

$$\boldsymbol{Y} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, ..., \boldsymbol{Y_q})$$

We select X and Y based on the number of variables that exist in each set so that $p \leqslant q$.
. This is done for computational convenience.

We look at linear combinations of the data, similar to principal components analysis. We define a set of linear combinations named U and V. U corresponds to the linear combinations from the first set of variables, X, and V corresponds to the second set of variables, Y. Each member of U is paired with a member of V. For example, U_1 below is a linear combination of the p X variables and V_1 is the corresponding linear combination of the q Y variables. Similarly, U_2 is a linear combination of the p X variables, and V_2 is the corresponding linear combination of the q Y variables. And, so on....

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

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

Thus define

$$U_i, Vi$$

as the i_{th} canonical variate pair (U_1, V_1) is the first canonical variate pair, similarly (U_2, V_2) would be the second canonical variate pair and so on. With $p \leqslant q$. there are p canonical covariate pairs.
We compute the variance of U_i variables with the following expression:

$$\mathbf{var}(\boldsymbol{U_i}) = \sum_{\boldsymbol{k=1}}^{\boldsymbol{p}} \sum_{\boldsymbol{l=1}}^{\boldsymbol{p}} \boldsymbol{a_{ik}a_{il}cov(X_k, X_l)}$$

The coeffcients a_{i1} through a_{ip} that appear in the double sum are the same

coefficients that appear in the definition of U_i . The covariances between the k_{th} and l_{th} X-variables are multiplied by the corresponding coefficients a_{ik} and a_{il} for the variate U_i .
Similar calculations can be made for the variance of V_j as shown below:

$$\mathbf{var}(V_j) = \sum_{k=1}^{p} \sum_{l=1}^{q} b_{jk} b_{jl} \mathbf{cov}(Y_k, Y_l)$$

The covariance between U_i and V_j is:

$$\mathbf{cov}(U_i, V_j) = \sum_{k=1}^{p} \sum_{l=1}^{q} a_{ik} b_{jl} \mathbf{cov}(X_k, Y_l)$$

The correlation between U_i and V_j is calculated using the usual formula. We take the covariance between the two variables and divide it by the square root of the product of the variances:

$$\frac{\mathbf{cov}(U_i, V_j)}{\sqrt{\mathbf{var}(U_i)\mathbf{var}(V_j)}}$$

The canonical correlation is a specific type of correlation. The canonical correlation for the i_{th} canonical variate pair is simply the correlation between U_i and V_i :

$$\rho_i^* = \frac{\mathbf{cov}(U_i, V_i)}{\sqrt{\mathbf{var}(U_i)\mathbf{var}(V_i)}}$$

This is the quantity to maximize. We want to find linear combinations of the X's and linear combinations of the Y's that maximize the above correlation.

Canonical Variates Defined:

Let us look at each of the p canonical variates pair one by one.
First canonical variate pair:(U_1, V_1) :
The coefficients $a_{11}, a_{12}, ..., a_{1p}$ and $b_{11}, b_{12}, ..., b_{1q}$ are selected to maximize the canonical correlation p_1^* of the first canonical variate pair. This is subject to the constraint that variances of the two canonical variates in that pair are equal to one.
var(U_1)=var(V_1)=1
This is required to obtain unique values for the coefficients.
Second canonical variate pair:(U_2, V_2) :
Similarly we want to find the coefficients $a_{21}, a_{22}, ..., a_{2p}$ and $b_{21}, b_{22}, ..., b_{2q}$ that maximize the canonical correlation p_2^* of the second canonical variate pair, (U_2, V_2). Again, we will maximize this canonical correlation subject to the constraints that the variances of the individual canonical variates are both equal to one. Furthermore, we require the additional constraints that (U_1, V_2) and

(U_2, V_1) are uncorrelated. In addition, the combinations (U_1, V_2) and (U_2, V_1) must be uncorrelated. In summary, our constraints are:
var(U_2)=var(V_2)=1
cov(U_1, U_2)=cov(V_1, V_2)=0
cov(U_1, V_2)=cov(U_2, V_1)=0
Basically, we require that all of the remaining correlations equal zero.
This procedure is repeated for each pair of canonical variates. In general, ...
i_th canonical variate pair:(U_i, V_i) :
The coefficients $a_{i1}, a_{i2}, ..., a_{ip}$ and $b_{i1}, b_{i2}, ..., b_{iq}$ are selected to maximize the canonical correlation p_i^* of the i_th canonical variate pair that
var(U_i)=var(V_i)=1
cov(U_1, U_i)=cov(V_1, V_i)=0
cov(U_2, U_i)=cov(V_2, V_i)=0
.
.
.
cov(U_{i-1}, V_i)=cov(U_i, V_{i-1})=0
Again, requiring all of the remaining correlations to be equal to zero.

5 Example:An Introduction to Canonical Correlation Analysis with Python

we want to find out how a school's ambience affects its students' academic success. On one hand, we have variables about the level of support, trust and collaboration in their learning environment. On the other hand, we have students' academic records and test results. CCA lets us explore associations between these two sets of variables as a whole, rather than considering them on an individual basis. Loosely speaking, we come up with a collective representation (a latent variable called canonical variate) for each of these variable sets in a way that the correlation between those variates is maximised.
We construct the first pair of Canonical Variates as linear combinations of the variables in each group:
cv_{1X}=a_1X_1 + a_2X_2+...+a_pX_p
cv_{1Y}=b_1Y_1 + b_2Y_2+...+b_pY_p
where the weights $(a_1, a_2, ..., a_p)$and $(b_1, b_2, ..., b_q)$ are chosen in a way that the correlation between the two variates is maximised.
We compute min(p,q) pairs in a similar fashion and end up with min(p,q) components ready to explore. the number of variables in each set doesn't have to be the same.
X=(Collaborative Teachers , Supportive Environment , Family-Community Ties , Trust)
Y=(Average ELA Proficiency, Average Math Proficiency)

6 Observation:

```
C:/Users/Dipankar/Desktop/194102305_ML_PROJECT_2/project_py/
project_py.py:29: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/indexing.html#indexing-view-versus-copy
  for col in X.columns:
Training CCA, kernel = None, regularization = 0.0000, 2 components
Canonical Correlation Per Component Pair: [0.44568176 0.14310912]
% Shared Variance: [0.19863223 0.02048022]
```



```
In [13]:
```

For our two pairs of canonical variates, we have canonical correlations of 0.44 and 0.18 respectively. So, latent representations of school ambience and students' performance do have a positive correlation of 0.46 and share 21 percent of variance.
weights assigned to our standardised variables is given as :

```
In [13]: cca.ws
Out[13]:
[array([[-9.62369999e-03, -2.93476761e-02],
        [-2.05346262e-02,  9.39019148e-03],
        [-1.47326031e-05,  1.11415592e-02],
        [ 1.04640581e-02,  6.42884184e-03]]),
 array([[ 0.0106447 , -0.05657155],
        [-0.02988275,  0.04920032]])]

In [14]:
```