Canonical Correlation

Today . . .

- The basic goal of canonical correlation analysis
- A simple example
- Interpretation

Canonical correlation finds the linear combinations of variables in two sets that are maximally correlated across sets but orthogonal within sets.

X_1 X_2 X_3 X_4 . . . X_q

What is the best way to understand how the variables in these two sets are related?

Y_1 Y_2 Y_3 Y_4 . . . Y_p

X_1 X_2 X_3 X_4 . . . X_q

What is the best way to understand how the variables in these two sets are related?

Y_1 Y_2 Y_3 Y_4 . . . Y_p

- Bivariate correlations across sets (pq)
- Multiple correlations across sets (p+q)
- Principal components within sets; correlations between principal components across sets

X_1 X_2 X_3 X_4 . . . X_q

What linear combinations of the X variables (**u**) and the Y variables (**t**) will maximize their correlation?

Y_1 Y_2 Y_3 Y_4 . . . Y_p

$$\left.\begin{array}{c} b_1X_1 \\ + \\ b_2X_2 \\ + \\ b_3X_3 \\ + \\ b_4X_4 \\ + \\ . \\ + \\ b_pX_p \end{array}\right\} \quad \left\{\begin{array}{c} a_1Y_1 \\ + \\ a_2Y_2 \\ + \\ a_3Y_3 \\ + \\ a_4Y_4 \\ + \\ . \\ + \\ a_qY_q \end{array}\right.$$

What linear combinations of the X variables (**u**) and the Y variables (**t**) will maximize their correlation?

= **u** = **t**

If X and Y are in standard score form, and

u = **Xb**

t = **Ya**

then find **a** and **b** to maximize $r_{t,u}$:

$$r_{tu} = \frac{1}{N-1} a'Y'Xb$$

while

$$\frac{1}{N-1} a'Y'Ya = 1 \qquad \frac{1}{N-1} b'X'Xb = 1$$

If X and Y are in standard score form, and

u = **Xb**

t = **Ya**

then find **a** and **b** to maximize $r_{t,u}$:

$$r_{tu} = a'R_{YX}b$$

while

$$a'R_{YY}a = 1 \qquad b'R_{XX}b = 1$$

	X_1	X_2	X_3	Y_1	Y_2	Y_3
X_1	1.00	$r_{X_1X_2}$	$r_{X_1X_3}$	$r_{X_1Y_1}$	$r_{X_1Y_2}$	$r_{X_1Y_3}$
X_2	$r_{X_2X_1}$	1.00	$r_{X_2X_3}$	$r_{X_2Y_1}$	$r_{X_2Y_2}$	$r_{X_2Y_3}$
X_3	$r_{X_3X_1}$	$r_{X_3X_2}$	1.00	$r_{X_3Y_1}$	$r_{X_3Y_2}$	$r_{X_3Y_3}$
Y_1	$r_{Y_1X_1}$	$r_{Y_1X_2}$	$r_{Y_1X_3}$	1.00	$r_{Y_1Y_2}$	$r_{Y_1Y_3}$
Y_2	$r_{Y_2X_1}$	$r_{Y_2X_2}$	$r_{Y_2X_3}$	$r_{Y_2Y_1}$	1.00	$r_{Y_2Y_3}$
Y_3	$r_{Y_3X_1}$	$r_{Y_3X_2}$	$r_{Y_3X_3}$	$r_{Y_3Y_1}$	$r_{Y_3Y_2}$	1.00

$$r_{tu} = a' R_{YX} b$$

$$b' R_{XX} b = 1$$

	X_1	X_2	X_3	Y_1	Y_2	Y_3
X_1	1.00	$r_{X_1X_2}$	$r_{X_1X_3}$	$r_{X_1Y_1}$	$r_{X_1Y_2}$	$r_{X_1Y_3}$
X_2	$r_{X_2X_1}$	1.00	$r_{X_2X_3}$	$r_{X_2Y_1}$	$r_{X_2Y_2}$	$r_{X_2Y_3}$
X_3	$r_{X_3X_1}$	$r_{X_3X_2}$	1.00	$r_{X_3Y_1}$	$r_{X_3Y_2}$	$r_{X_3Y_3}$
Y_1	$r_{Y_1X_1}$	$r_{Y_1X_2}$	$r_{Y_1X_3}$	1.00	$r_{Y_1Y_2}$	$r_{Y_1Y_3}$
Y_2	$r_{Y_2X_1}$	$r_{Y_2X_2}$	$r_{Y_2X_3}$	$r_{Y_2Y_1}$	1.00	$r_{Y_2Y_3}$
Y_3	$r_{Y_3X_1}$	$r_{Y_3X_2}$	$r_{Y_3X_3}$	$r_{Y_3Y_1}$	$r_{Y_3Y_2}$	1.00

	X_1	X_2	X_3	Y_1	Y_2	Y_3
X_1	1.00	$r_{X_1X_2}$	$r_{X_1X_3}$	$r_{X_1Y_1}$	$r_{X_1Y_2}$	$r_{X_1Y_3}$
X_2	$r_{X_2X_1}$	1.00	$r_{X_2X_3}$	$r_{X_2Y_1}$	$r_{X_2Y_2}$	$r_{X_2Y_3}$
X_3	$r_{X_3X_1}$	$r_{X_3X_2}$	1.00	$r_{X_3Y_1}$	$r_{X_3Y_2}$	$r_{X_3Y_3}$
Y_1	$r_{Y_1X_1}$	$r_{Y_1X_2}$	$r_{Y_1X_3}$	1.00	$r_{Y_1Y_2}$	$r_{Y_1Y_3}$
Y_2	$r_{Y_2X_1}$	$r_{Y_2X_2}$	$r_{Y_2X_3}$	$r_{Y_2Y_1}$	1.00	$r_{Y_2Y_3}$
Y_3	$r_{Y_3X_1}$	$r_{Y_3X_2}$	$r_{Y_3X_3}$	$r_{Y_3Y_1}$	$r_{Y_3Y_2}$	1.00

$$a' R_{YY} a = 1$$

The correlation between the two sets is called the ***canonical correlation*** and is the largest possible correlation that can be found between linear combinations.

The weights (**a** and **b**) that are used to create the linear combinations are called the ***standardized canonical coefficients***. The linear combinations created (**t** and **u**) are called the ***canonical variates***.

Additional canonical variates and their correlations can be found provided they satisfy:

$$max(r_{tu}) = \frac{1}{N-1} a_2' Y' X b_2$$

$$\frac{1}{N-1} a_2' Y' Y a_2 = 1 \qquad \frac{1}{N-1} b_2' X' X b_2 = 1$$

$$r_{t_1 t_2} = 0 \qquad r_{u_1 u_2} = 0$$

Additional canonical variates and their correlations can be found provided they satisfy:

$$max(r_{tu}) = a_2' R_{YX} b_2$$

$$a_2' R_{YY} a_2 = 1 \qquad b_2' R_{XX} b_2 = 1$$

$$r_{t_1 t_2} = 0 \qquad r_{u_1 u_2} = 0$$

The extraction of canonical variates can continue up to a maximum defined by the number of measures in the smaller of the two sets.

The standardized canonical coefficients (**a** and **b**) are interpreted in the same way as standardized regression coefficients in multiple regression—they indicate the unique contribution of a variable to the linear combination.

It is also possible to derive the correlations between each variable and the linear combination. These are called ***canonical loadings*** or ***canonical structure coefficients*** and are interpreted the same way as loadings in principal components and discriminant analysis.

These loadings can be calculated as:

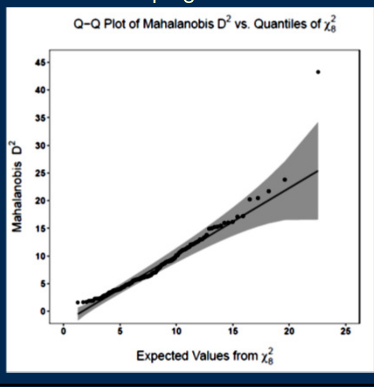
$$R_{x_u} = \frac{1}{N-1} X' u = \frac{1}{N-1} X'(Xb) = R_{xx} b$$

The loadings can assist in understanding the nature of the linear combinations in each set.

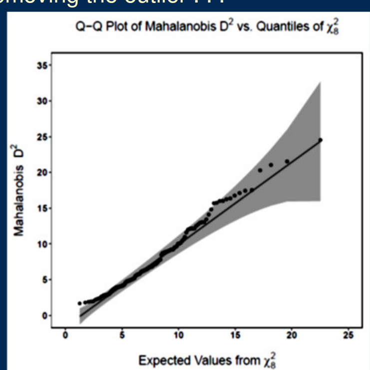
In this study, 126 college students completed a measure of the Big Five (the NEO), a measure of trait self-esteem (the Rosenberg), and measures of positive and negative affect (the PANAS).

A canonical correlation analysis can be used to determine if personality is related to feelings about the self and general affect.

First, a little housekeeping . . .



After removing the outlier . . .




```
cor(CCA_Trimmed[, 5:9])

##          neo_n    neo_e    neo_o   neo_a   neo_c
## neo_n  1.0000 -0.38685  0.21904 -0.1611 -0.2719
## neo_e -0.3869  1.00000  0.05385  0.2711  0.1508
## neo_o  0.2190  0.05385  1.00000  0.2680 -0.3069
## neo_a -0.1611  0.27112  0.26800  1.0000  0.1331
## neo_c -0.2719  0.15075 -0.30690  0.1331  1.0000

cor(CCA_Trimmed[, 2:4])

##          rosen_se   panas_n   panas_p
## rosen_se   1.0000 -0.336316  0.188864
## panas_n   -0.3363  1.000000 -0.008335
## panas_p    0.1889 -0.008335  1.000000

cor(CCA_Trimmed[, 2:4], CCA_Trimmed[, 5:9])

##             neo_n   neo_e    neo_o    neo_a   neo_c
## rosen_se -0.6966  0.3371 -0.20468  0.05339  0.2999
## panas_n   0.4093 -0.3656  0.07355 -0.15647 -0.2283
## panas_p  -0.2875  0.1332  0.14284  0.11280  0.2043
```

A simple canonical correlation analysis can be conducted using the cancor( ) function from the candisc package.

```
CCA_1 <- cancor(cbind(rosen_se, panas_n, panas_p) ~ neo_n + neo_e +
    neo_o + neo_a + neo_c, data = CCA_Trimmed, prefix = c("NEO", "Well_Being"),
    standardize = TRUE)
```

```
Test of H0: The canonical correlations in the
current row and all that follow are zero

   CanR WilksL    F df1 df2 p.value
1 0.771  0.357 9.75  15 323  0.0000
2 0.285  0.882 1.91   8 236  0.0593
3 0.200  0.960 1.65   3 119  0.1820
```

```
    CanR  CanRSQ   Eigen percent    cum
1 0.7715 0.59514 1.46997  91.875  91.87
2 0.2851 0.08127 0.08846   5.529  97.40
3 0.1997 0.03989 0.04155   2.597 100.00
```

Procedures for testing the significance of the canonical correlations are applied sequentially.

At each step, the test indicates whether there is any remaining significant relationship between the sets.

In this case, one pair of linear combinations can be formed, accounting for 92% of the relationship between the sets.

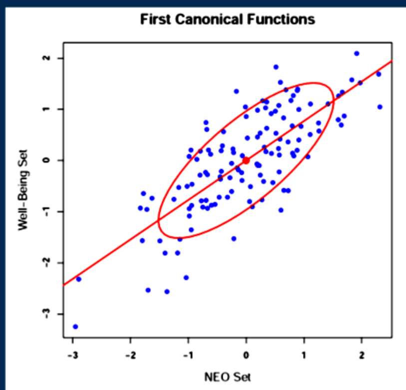
```
Canonical correlation analysis of:
  5   X  variables:  neo_n, neo_e, neo_o, neo_a, neo_c
  with  3   Y  variables:  rosen_se, panas_n, panas_p

    CanR  CanRSQ   Eigen percent    cum
1 0.7130 0.50835 1.03395  86.536  86.54
2 0.3223 0.10391 0.11596   9.705  96.24
3 0.2073 0.04298 0.04491   3.759 100.00
                           scree
1 ************************
2 ***
3 *

Test of H0: The canonical correlations in the
current row and all that follow are zero

   CanR WilksL    F df1 df2 p.value
1 0.713  0.422 7.99  15 326  0.0000
2 0.322  0.858 2.38   8 238  0.0177
3 0.207  0.957 1.80   3 120  0.1515
```

Had we not screened the data, we would have thought a second pair of linear combinations should be interpreted.




```
coef(CCA_1, standardize = TRUE)

##          NEO1     NEO2    NEO3
## neo_n -0.85535  0.06982  0.6239
## neo_e  0.18407 -0.06990  0.9143
## neo_o  0.09491  0.97106 -0.3711
## neo_a -0.08610  0.09887  0.4329
## neo_c  0.24388  0.56869 -0.0560

coef(CCA_2, standardize = TRUE)

##          Well-Being1 Well-Being2 Well-Being3
## rosen_se      0.7531     -0.6001      -0.496
## panas_n      -0.3311     -0.2633      -0.976
## panas_p       0.2751      0.9413      -0.281
```

Typically the linear combinations are formed after the variables have been standardized. The weights are then interpreted as standardized regression coefficients.

```
$X.xscores
         NEO1    NEO2     NEO3
neo_n -0.9582 0.13900  0.13441
neo_e  0.5335 0.06792  0.76190
neo_o -0.1804 0.83456 -0.05204
neo_a  0.1595 0.40462  0.47332
neo_c  0.4636 0.25432  0.08372
```

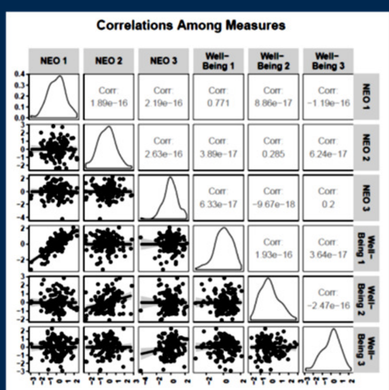
```
$Y.yscores
          Well_Being1 Well_Being2 Well_Being3
rosen_se       0.9164    -0.33373     -0.2209
panas_n       -0.5867    -0.06938     -0.8068
panas_p        0.4201     0.83018     -0.3665
```

The loadings or structure coefficients are the bivariate relationships within sets, between variables and linear combinations.

```
$Y.xscores
            NEO1     NEO2     NEO3
rosen_se  0.7070 -0.09514 -0.04411
panas_n  -0.4526 -0.01978 -0.16115
panas_p   0.3241  0.23666 -0.07320
```

```
$X.yscores
      Well_Being1 Well_Being2 Well_Being3
neo_n     -0.7392     0.03963     0.02685
neo_e      0.4116     0.01936     0.15217
neo_o     -0.1392     0.23791    -0.01039
neo_a      0.1230     0.11535     0.09454
neo_c      0.3576     0.07250     0.01672
```

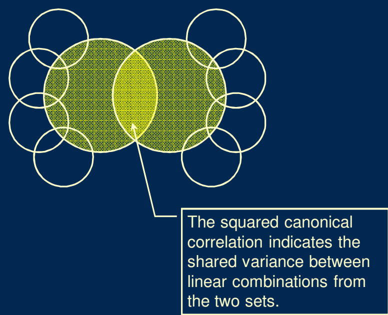
The cross loadings are the bivariate relationships between variables in one set and the linear combinations for the other set.

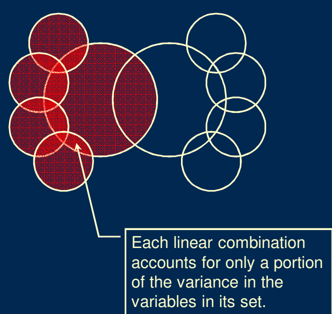


How much variance is
really accounted for?

Reliance on the canonical correlations for evidence of variance accounted for across sets of variables can be misleading.

Each linear combination only captures a portion of the variance in its own set. That needs to be taken into account when judging the variance accounted for across sets.



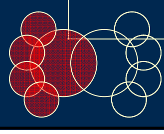


Redundancy coefficients indicate the proportion of variance in the variables of Set B that is accounted for by the linear combination of variables in Set A.

These coefficients are a function of the adequacy coefficients and the squared canonical correlations.

Canonical Loadings

	Variate 1	Variate 2	
X_1	$L_{1,1}$	$L_{1,2}$	$\Sigma L^2_{1,j}$
X_2	$L_{2,1}$	$L_{2,2}$	$\Sigma L^2_{2,j}$
X_3	$L_{3,1}$	$L_{3,2}$	$\Sigma L^2_{3,j}$
X_4	$L_{4,1}$	$L_{4,2}$	$\Sigma L^2_{4,j}$
	$(\Sigma L^2_{i,1})/4$	$(\Sigma L^2_{i,2})/4$	



Adequacy Coefficients

Canonical communality coefficients

```
$X.xscores
          NEO1     NEO2      NEO3
neo_n -0.9582 0.13900  0.13441
neo_e  0.5335 0.06792  0.76190
neo_o -0.1804 0.83456 -0.05204
neo_a  0.1595 0.40462  0.47332
neo_c  0.4636 0.25432  0.08372
```

The sum of these elements squared, divided by the number of these elements is the adequacy . . .

```
(NEO_Adequacy <- colSums((NEO_Loadings)^2)/length(NEO_Loadings[, 1]))
##   NEO1   NEO2   NEO3
## 0.2951 0.1898 0.1665
```

The adequacy multiplied by the squared canonical correlation gives the redundancy.

```
Redundancies for the X variables

    NEO1      NEO2      NEO3  total X|Y
 0.17565   0.01542   0.00664    0.19771
```

```
    CanR  CanRSQ
1 0.7130 0.50835
2 0.3223 0.10391
3 0.2073 0.04298
```

```
(Well_Being_Adequacy <- colSums((Well_Being_Loadings)^2)/
                       length(Well_Being_Loadings[, 1]))
## Well_Being1 Well_Being2 Well_Being3
##      0.4535      0.2685      0.2780
(NEO_Adequacy <- colSums((NEO_Loadings)^2)/length(NEO_Loadings[, 1]))
##   NEO1   NEO2   NEO3
## 0.2951 0.1898 0.1665
```

Canonical correlations will decrease in size but adequacy coefficients do not necessarily decrease in size. Why?

```
    CanR  CanRSQ
1 0.7130 0.50835
2 0.3223 0.10391
3 0.2073 0.04298
```

These canonical community coefficients are equal to 1.00, but . . .

```
(Well_Being_Communalities <- rowSums((Well_Being_Loadings)^2))

## rosen_se  panas_n  panas_p
##        1        1        1

(NEO_Communalities <- rowSums((NEO_Loadings)^2))

##   neo_n  neo_e  neo_o  neo_a  neo_c
## 0.9555 0.8697 0.7318 0.4132 0.2866
```

these canonical community coefficients are less than 1.00. Why?

- Any given loading can be squared to indicate the proportion of the variance in that variable that is accounted for by that canonical variate.
- The sum of the squared loadings for a given variable indicates the total proportion of variance accounted for by the collection of canonical variates.

- The average of the squared loadings for a canonical variate is the adequacy coefficient and indicates the proportion of variance in the collection of variables that is accounted for by the canonical variate.
- The redundancy coefficient is the proportion of variance in a set of variables that is accounted for by a linear combination from the other set.
- The sum of the redundancy coefficients gives the total proportion of variance in one set that is accounted for by the other set. These will usually be different values for each set.

Next time . . .

Assumptions, cross-validation, capitalization on chance.