Canonical Correlation

Today . . .

- How easily does the procedure capitalize on chance?
- How are canonical correlations cross-validated?
- Assumptions?

How easily does the procedure capitalize on chance?

Canonical correlation analysis has elements of two procedures—principal components analysis and multiple regression analysis—that are often criticized as susceptible to capitalizing on chance.

How worried should we be? An easy way to answer questions like this is through simulation with data generated to have particular properties.

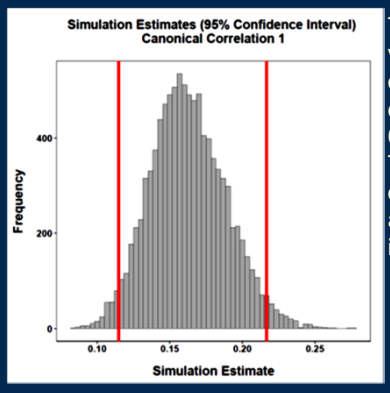
A total of 10,000 samples (N= 500) of 10 variables were drawn from a multivariate standard normal distribution with the following covariance matrix:

```
VC
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]  1.0  0.5  0.5  0.5  0.5  0.0  0.0  0.0  0.0   0.0
##  [2,]  0.5  1.0  0.5  0.5  0.5  0.0  0.0  0.0  0.0   0.0
##  [3,]  0.5  0.5  1.0  0.5  0.5  0.0  0.0  0.0  0.0   0.0
##  [4,]  0.5  0.5  0.5  1.0  0.5  0.0  0.0  0.0  0.0   0.0
##  [5,]  0.5  0.5  0.5  0.5  1.0  0.0  0.0  0.0  0.0   0.0
##  [6,]  0.0  0.0  0.0  0.0  0.0  1.0  0.5  0.5  0.5   0.5
##  [7,]  0.0  0.0  0.0  0.0  0.0  0.5  1.0  0.5  0.5   0.5
##  [8,]  0.0  0.0  0.0  0.0  0.0  0.5  0.5  1.0  0.5   0.5
##  [9,]  0.0  0.0  0.0  0.0  0.0  0.5  0.5  0.5  1.0   0.5
## [10,]  0.0  0.0  0.0  0.0  0.0  0.5  0.5  0.5  0.5   1.0
```

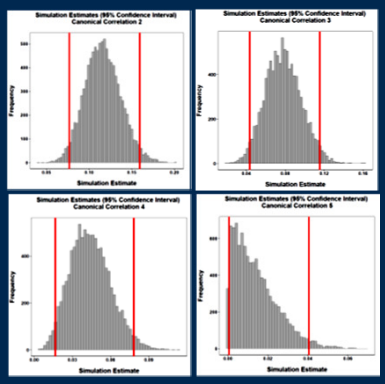
Two sets of 5 items, correlated .5 within sets but uncorrelated between sets.

Results from the 10,000 canonical correlation analyses can be examined to determine the expected behavior of the approach when the underlying data are consistent with the null hypothesis of no relation between sets.

- How large are the canonical correlations that are found?
- Because of the chance-capitalizing nature of the procedure, do we reject the null more often than we should?
- How large are loadings, weights, adequacy coefficients, communalities, and redundancies?



The expected value for the first canonical correlation is not 0 under the null. The procedure capitalizes on any relationship it can find.



The procedure will find any relationship it can, even when purely chance-determined as must be true here.

Despite the truth of the null hypothesis underlying the data being analyzed, non-zero canonical correlations are not at all unusual. Should we be worried? Not necessarily.

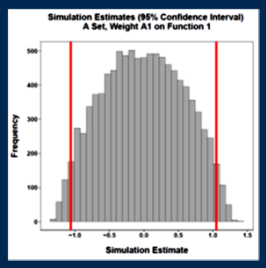
Canonical correlations must lie between 0 and 1 and because the procedure finds the maximum correlation at each step, correlations greater than 0 would be expected even if only capitalizing on chance. The key is whether the significance tests fairly take this into account. For the 10,000 analyses we can tally the proportion of tests that were significant.

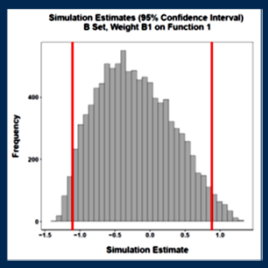
Null Rejection Summary

Canonical Correlation	*p*	*Lower 95% CL*	*Upper 95% CL*
1	*0.0468*	*0.0428*	*0.0511*
2	*0.0256*	*0.0147*	*0.0443*
3	*0.0833*	*0.0149*	*0.3539*
4	*0*	*0*	*0.7935*
5	-	-	-

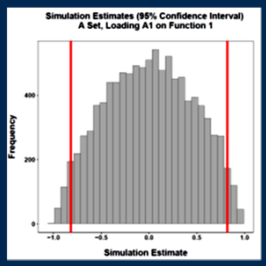
The tests of canonical correlations are sequential so a test at one step is only interpreted if tests at previous steps are significant: Given that the first test is significant, how often is the second test also significant? 2.56% The first step tests if one or more of the canonical correlations is significant. The results are clearly consistent with the null hypothesis (only 4.68% rejections).

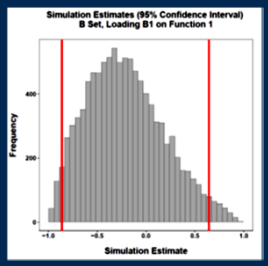
Although the canonical correlations are relatively low, individual weights can be sizeable. Why?





That is true for the loadings (correlations) too.



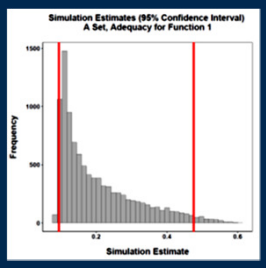


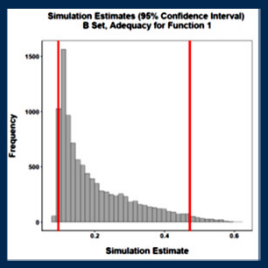
The linear combinations will faithfully and completely reproduce the variance of the variables within sets when the number of canonical functions equals the number of variables. The communality for each variable in each analysis in the simulation is 1.00.

This is a reminder that canonical variates are simply transformations of the original variables and can represent those variables completely.

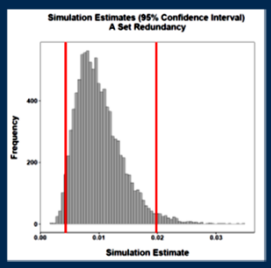
The adequacy coefficients will likewise accurately reflect the proportion of variance within a set that is accounted for by the canonical variates.

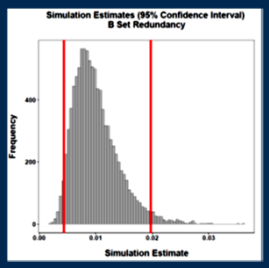
With 5 functions and 5 variables for each set, any given function should account for about 20% of its own set's variance given the null hypothesis.





Given that the null hypothesis is true in the simulation, the canonical variates should account for trivial amounts of variance in the other set.





Provided the significance tests are used to guide decisions about the presence of canonical correlations, the procedure will not "unfairly" capitalize on chance.

Nonetheless, like other statistical procedures used in a largely exploratory way, our faith in the conclusions is likely to be bolstered considerably with cross-validation.

The most convincing approach to cross-validation requires a calibration sample and a separate hold-out sample.

The calibration sample is analyzed and the canonical coefficients derived. Those coefficients are then applied to the hold-out sample. A standard canonical correlation analysis is also conducted on the hold-out sample.

The correlations among the actual and estimated canonical variates are computed. Corresponding canonical variates should be highly correlated.

A sample of 400 participants completed the NEO (The Big Five, Set 1) and measures of self-esteem, optimism, life satisfaction, and happiness (Set 2).

The sample was split randomly to produce a calibration sample and a hold-out sample.

A canonical correlation analysis was conducted on the calibration sample to determine the nature of the relations between sets and to obtain standardized canonical coefficients to use in the hold-out sample.

```
CCA_Calibration

##
## Canonical correlation analysis of:
##   5   Personality  variables:  NEO_N, NEO_E, NEO_O, NEO_A, NEO_C
##   with  4   Well-Being  variables:  SE, Opt, Life_Sat, Happy
##
##     CanR  CanRSQ   Eigen percent    cum
## 1 0.8768 0.76882 3.32566 68.9370  68.94
## 2 0.7190 0.51692 1.07003 22.1806  91.12
## 3 0.5347 0.28585 0.40027  8.2972  99.41
## 4 0.1657 0.02746 0.02823  0.5852 100.00
##                        scree
## 1 ************************
## 2 ********
## 3 ***
## 4
Wilks' Lambda, using F-approximation (Rao's F):
           stat approx df1  df2  p.value
1 to 4:  0.07757 75.374  20 1298 0.000000
2 to 4:  0.33552 44.176  12 1037 0.000000
3 to 4:  0.69454 26.189   6  786 0.000000
4 to 4:  0.97254  5.562   2  394 0.004149
```

All four functions are significant, but the first two account for nearly all of the relations across sets.

Sample 1: Calibration Sample

```
CCA_Calibration$coef
## $X
##              P_1      P_2      P_3       P_4
## NEO_N -0.22308 -0.80648  0.8251 -0.393287
## NEO_E  0.17813 -0.49624 -0.1667 -0.007479
## NEO_O  0.63299  0.30450  0.3715  0.714660
## NEO_A  0.04131 -0.91893 -0.1351  0.041384
## NEO_C  0.73568  0.02147  0.3811 -0.689296
##
## $Y
##              WB_1    WB_2    WB_3    WB_4
## SE         0.1773  0.5026 -0.6913 -0.5772
## Opt        0.6268  0.5786  0.3356  0.7041
## Life_Sat   0.5725 -0.5111  0.4903 -0.8200
## Happy     -0.1364 -0.6474 -0.8329  0.6122
```

```
CCA_Calibration$structure

## $X.xscores
##              P_1      P_2      P_3       P_4
## NEO_N -0.4230 -0.23239  0.86784  0.005039
## NEO_E  0.3182 -0.46109 -0.39102  0.009468
## NEO_O  0.4602 -0.01180  0.47155  0.752108
## NEO_A  0.3692 -0.63714 -0.39483  0.224249
## NEO_C  0.7372 -0.08711 -0.02572 -0.660487
## $Y.yscores
##              WB_1    WB_2     WB_3     WB_4
## SE         0.4162  0.4331 -0.69513 -0.3950
## Opt        0.8572  0.2375  0.01090  0.4569
## Life_Sat   0.7904 -0.5315  0.06485 -0.2977
## Happy      0.4652 -0.5766 -0.58109  0.3367
```

The coefficients, especially the structure coefficients, suggest that the first pair of functions represent the link between healthy personality and an optimistic and satisfying life. The second pair represent a link between agreeableness and extraversion with high life satisfaction and happiness but low self-esteem.

The canonical correlation analysis is repeated on the hold-out sample.

Similarity in the patterns of correlations, coefficients, and loadings would suggest stability in the solution but this involves a very large number of comparisons.

An alternative approach uses the standardized coefficients from the calibration sample along with the standardized data from the hold-out sample to estimate canonical variate scores in the hold-out sample.

If the canonical coefficients from the first sample are stable and not subject to wild chance fluctuation, they should produce estimated canonical variate scores in the hold-out sample that are highly correlated with canonical variate scores calculated entirely from the hold-out sample.

Four sets of canonical scores need to be calculated:

Actual Set 1 Canonical Variate Scores:

$Z_{\text{Set 1 Sample 2}}$ $W_{\text{Set 1 Sample 2}}$

Actual Set 2 Canonical Variate Scores:

$Z_{\text{Set 2 Sample 2}}$ $W_{\text{Set 2 Sample 2}}$

Estimated Set 1 Canonical Variate Scores:

$Z_{\text{Set 1 Sample 2}}$ $W_{\text{Set 1 Sample 1}}$

Estimated Set 2 Canonical Variate Scores:

$Z_{\text{Set 2 Sample 2}}$ $W_{\text{Set 2 Sample 1}}$

	A_P_1	A_P_2	A_P_3	A_P_4	A_WB_1	A_WB_2	A_WB_3	A_WB_4
A_P_1	1.00	0.000	0.000	0.000	0.88	0.000	0.000	0.000
A_P_2	0.00	1.000	0.000	0.000	0.00	0.664	0.000	0.000
A_P_3	0.00	0.000	1.000	0.000	0.00	0.000	0.403	0.000
A_P_4	0.00	0.000	0.000	1.000	0.00	0.000	0.000	0.164
A_WB_1	0.88	0.000	0.000	0.000	1.00	0.000	0.000	0.000
A_WB_2	0.00	0.664	0.000	0.000	0.00	1.000	0.000	0.000
A_WB_3	0.00	0.000	0.403	0.000	0.00	0.000	1.000	0.000
A_WB_4	0.00	0.000	0.000	0.164	0.00	0.000	0.000	1.000

These are the correlations among the actual (from the hold-out sample) canonical variate scores for the four possible variates in each set (personality and well-being). The gray highlighted values are the canonical correlations in the hold-out set. These are similar to the values from the calibration set:

1	0.877
2	0.719
3	0.535
4	0.166

The variates are also independent within sets (yellow highlighted blocks).

	A_P_1	A_P_2	A_P_3	A_P_4	E_P_1	E_P_2	E_P_3	E_P_4
A_P_1	1.000	0.000	0.000	0.000	0.989	0.164	-0.007	0.048
A_P_2	0.000	1.000	0.000	0.000	-0.125	0.882	-0.348	0.285
A_P_3	0.000	0.000	1.000	0.000	0.069	-0.309	-0.933	0.004
A_P_4	0.000	0.000	0.000	1.000	-0.012	0.202	-0.056	-0.900
E_P_1	0.989	-0.125	0.069	-0.012	1.000	0.039	-0.031	0.038
E_P_2	0.164	0.882	-0.309	0.202	0.039	1.000	-0.047	0.156
E_P_3	-0.007	-0.348	-0.933	-0.056	-0.031	-0.047	1.000	-0.075
E_P_4	0.048	0.285	0.004	-0.900	0.038	0.156	-0.075	1.000

The canonical variate scores based on Sample 2 should be highly related to corresponding canonical variate scores that are estimated from Sample 1 weights. This is true for the personality canonical variates. Likewise, the estimated scores should show the independence that is true for the actual canonical variate scores.

	A_WB_1	A_WB_2	A_WB_3	A_WB_4	E_WB_1	E_WB_2	E_WB_3	E_WB_4
A_WB_1	1.000	0.000	0.000	0.000	0.991	0.164	0.004	-0.038
A_WB_2	0.000	1.000	0.000	0.000	-0.067	0.934	-0.254	0.292
A_WB_3	0.000	0.000	1.000	0.000	0.118	-0.211	-0.963	-0.064
A_WB_4	0.000	0.000	0.000	1.000	-0.007	0.238	0.096	-0.954
E_WB_1	0.991	-0.067	0.118	-0.007	1.000	0.073	-0.093	-0.058
E_WB_2	0.164	0.934	-0.211	0.238	0.073	1.000	-0.010	0.052
E_WB_3	0.004	-0.254	-0.963	0.096	-0.093	-0.010	1.000	-0.104
E_WB_4	-0.038	0.292	-0.064	-0.954	-0.058	0.052	-0.104	1.000

The same pattern is replicated for the well-being canonical variate scores.

Failure in cross-validation can occur when expected high correlations do not emerge, or, when high correlations appear where they are not expected. Neither is a problem for this data set.

	E_P_1	E_P_2	E_P_3	E_P_4	E_WB_1	E_WB_2	E_WB_3	E_WB_4
E_P_1	1.000	0.039	-0.031	0.038	0.871	0.059	-0.003	-0.057
E_P_2	0.039	1.000	-0.047	0.156	0.089	0.604	-0.025	0.142
E_P_3	-0.031	-0.047	1.000	-0.075	-0.035	-0.139	0.420	-0.034
E_P_4	0.038	0.156	-0.075	1.000	0.031	0.148	-0.063	0.194
E_WB_1	0.871	0.089	-0.035	0.031	1.000	0.073	-0.093	-0.058
E_WB_2	0.059	0.604	-0.139	0.148	0.073	1.000	-0.010	0.052
E_WB_3	-0.003	-0.025	0.420	-0.063	-0.093	-0.010	1.000	-0.104
E_WB_4	-0.057	0.142	-0.034	0.194	-0.058	0.052	-0.104	1.000

The most stringent test uses the data from Sample 2 and the weights from Sample 1. The estimated canonical correlations are remarkably close to those based entirely on Sample 2:

	CanR
1	0.880
2	0.664
3	0.403
4	0.164

Assumptions and other odds and ends:

- Interval level data
- Linear relations
- Homoskedasticity
- Low measurement error
- Unrestricted variances
- Low multicollinearity

Assumptions and other odds and ends:

- Similar distributions for all measures
- Multivariate normality for significance tests
- Sufficient sample size (10-20 times as many cases as variables to interpret the first canonical correlation; 40-60 times as many cases as variables for more than one canonical correlation)
- No outliers

The End