**Canonical Correlation[[1]](#footnote-1)©**

In a canonical correlation (multiple multiple correlation) one has two or more X variables and two or more Y variables. The goal is to describe the relationships between the two sets of variables. You find the canonical weights (coefficients) *a1*, *a2*, *a3*, ... *ap* to be applied to the *p* X variables and *b1*, *b2*, *b3*, ... *bm* to be applied to the *m* Y variables in such a way that the correlation between CVX1 and CVY1 is maximized.

*CVX1* = *a1X1* + *a2X2* +...+ *apXp*. *CVY1* = *b1Y1* + b2 Y2 + ... + *bm Ym*. *CVX1* and *CVY1* are the first canonical variates, and their correlation is the sample canonical correlation coefficient for the first pair of canonical variates. The residuals are then analyzed in the same fashion to find a second pair of canonical variates, *CVX2* and *CVY2*, whose weights are chosen to maximize the correlation between *CVX2* and *CVY2*, using only the variance remaining after the variance due to the first pair of canonical variates has been removed from the original variables. This continues until a "significance" cutoff is reached or the maximum number of pairs (which equals the smaller of *m* and *p*) has been found.

You may think of the pairs of canonical variates as representing superordinate constructs. For each pair this construct is estimated as a linear combination of the variables, where the sole criterion for choosing one linear combination over another is maximizing the correlation between the two canonical variates. The resulting constructs may not be easily interpretable as representing an underlying dimension of interest. Since each pair of canonical variates is calculated from the residuals of the pair(s) extracted earlier, the resulting canonical variates are orthogonal. The underlying dimensions in which you are interested may, however, be related to one another.

To learn about canonical correlation, we shall reproduce the analysis reported by Patel, Long, McCammon, & Wuensch (*Journal of Interpersonal Violence*, 1995, *10*: 354-366, 1994). We had two sets of data on a group of male college students. The one set was personality variables from the MMPI. One of these was the PD (psychopathically deviant) scale, Scale 4, on which high scores are associated with general social maladjustment, rebelliousness, antisocial behavior, criminal behavior, impulsive acting out, insensitivity, hostility, and difficulties with interpersonal relationships (family, school, and authority figures). The second was the MF (masculinity/femininity) scale, Scale 5, on which low scores are associated with traditional masculinity ‑ being easy‑going, cheerful, practical, coarse, adventurous, lacking insight into own motives, preferring action to thought, overemphasizing strength and physical prowess, having a narrow range of interests, and harboring doubts about one's own masculinity and identity[[2]](#endnote-1). The third was the MA (hypomania) scale, Scale 9, on which high scores are associated with overactivity, emotional lability, flight of ideas, being easily bored, having low frustration tolerance, narcissism, difficulty inhibiting impulses, thrill‑seeking, irritability, restlessness, and aggressiveness. The fourth MMPI variable was Scale K, which is a validity scale on which high scores indicate that the subject is “clinically defensive,” attempting to present himself in a favorable light, and low scores indicate that the subject is unusually frank. The second set of variables was a pair of homonegativity variables. One was the IAH (Index of Attitudes Towards Homosexuals), designed to measure affective components of homophobia. The second was the SBS, (Self‑Report of Behavior Scale), designed to measure past aggressive behavior towards homosexuals, an instrument specifically developed for this study.

Download the following files:

* Sunita.dat from my [StatData Page](http://core.ecu.edu/psyc/wuenschk/StatData/StatData.htm).
* Canonical.sas from my [SAS Programs Page](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm).

Edit the program file so that it properly points to the location of Sunita.dat and then run the program. The IF statement is used to restrict the analysis to data from men.

The zero-order correlations show that aggression against gays is significantly correlated with hypomania and homophobia. Homophobia is significantly correlated with masculinity. Although not significantly correlated with either of the homonegativity variables, psychopathic deviance is significantly associated with hypomania and with the K scale (clinical defensiveness).

In **PROC CANCORR**, ALL requests all optional statistics, VN gives a name to the variables specified in the VAR statement, WN gives a name to the variables specified in the WITH statement, VP gives a prefix to be used with names for the canonical variates constructed from the variables in the VAR statement, and WP gives a prefix to be used with names for the canonical variates constructed from the variables in the WITH statement.

Two pairs (the maximum) of canonical variates were constructed. The first has a canonical correlation of .38, the second .32. The first canonical correlation will always have a value at least as large as the largest *R* between one variable and the opposite set of variables, but that canonical *r* can be much larger than that largest *R*. The test of the significance reported in the row with the first eigenvalue is a test that as a set (both pairs of canonical variates tested simultaneously) CVX is independent of CVY. This null was rejected, *p* = .0099. One next tests all remaining pairs (as a set) with the first pair removed, then all remaining pairs with the first and second pairs removed, etc. For the *p* values from such a test to be valid, at least one of the two sets of variables should have an approximately normal multivariate distribution. For these data the second canonical correlation was significant (*p* = .0392) with the first removed. I generally do not interpret a canonical correlation that is less than .30, even if it is significant, since it is trivially small (the overlap, percentage of variance shared by the two canonical variates, is 9% or less), but if I found a small correlation to be meaningful, I just might share my interpretation of it.

For each root the **eigenvalue** is equal to the ratio of the squared canonical correlation (explained variance in the canonical variate) to one minus the squared canonical correlation (unexplained variance in the canonical variate). An eigenvalue of 1 would be obtained if the squared canonical correlation was .5 – the proportion of variance explained would be equal to the proportion of variance not explained. An eigenvalue of 1/3 would be obtained if the squared canonical correlation was .25 – the proportion of unexplained variance would be three times the proportion of explained variance. An eigenvalue of 3 would be obtained if the squared canonical correlation was .75 – the proportion of explained variance would be three times the proportion of unexplained variance.

CANCORR gave us the raw and the standardized coefficients (*a1*, *a2*, ... *b1*, *b2*, ...) for each pair of canonical variates. One generally interprets the canonical variates from their loadings rather than from their canonical coefficients, and SAS gives us those loadings under the descriptive title "Correlations Between The (set of variables) And Their Canonical Variables." For the Homonegativity variables, CV1 loads heavily on both IAH and SBS -- high scores on this CV indicate that the individual is homophobic and aggresses against gays. For the MMPI variables, CV1 loads well on all MMPI scales (negatively on MF and K) -- high scores on this CV indicate that the individual is hypomanic, masculine, psychopathically deviant, and unusually frank. The canonical correlation for the first pair of canonical variates indicates that stereotypically masculine, hypomanic, psychopathically deviant, frank men are homophobic and report aggressing against homosexuals.

The second pair of canonical variates show suppression. Look at the correlations and the standardized coefficients (beta weights) for the homonegativity CV2 and its variables. For each of the variables, the beta weights are higher than the correlations, indicating cooperative suppression (each variable suppresses irrelevant variance in the other). Individuals scoring high on this CV are not homophobic, but do aggress against gays. Perhaps these individuals are, in the words of one of my graduate students (Cliff Wall), “equal opportunity bullies” -- they aggress against everybody, not just against gays. Such nondiscriminatory aggression is associated with (look at the loadings for the second CV of the MMPI) hypomania and femininity (dare I call this CV ‘bitchiness’?).

One generally wants to know **how much variance a canonical variate extracts from its set of variables**. Look at the loadings for the Homonegativity variables. Homoneg\_1 captures .52192 of the variance in IAH and .99072 of the variance in SBS, so the proportion of the total variance in the Homonegativity variables captured by the Homoneg\_1 canonical variate is the mean of .52192 and .99072 = .6269. We can find the proportion of variance each canonical variate extracts from its own set of variables by simply finding the mean of the squared loadings between the canonical variate and the variables of its own set. SAS gives us these proportions as "Standardized Variance of the (set of variables) Explained by Their Own Canonical Variables." Homoneg\_2 captures the remaining 37.31% of the variance in the Homonegativity variables, for a total of 100% captured by the two Homonegativity canonical variates.

MMPI\_1 captures 22.70% of the variance in the MMPI variables and MMPI\_2 another 28.92%, for a total of 51.62% captured by the two MMPI canonical variates. Note that with respect to capturing variance in MMPI, the second MMPI canonical variate captured more than did the first - recall that canonical variates are ordered by the size of their canonical correlations, not by how much variance they capture in their variables. Here, the first weighted linear combination of the MMPI is constructed from that variance in the MMPI that is best related to the first linear combination of the Homonegativity, but that variance of the first MMPI is less than that of the second MMPI. It is always true that 100% of the variance will be captured from the smaller set of variables and less from the larger set. If both sets have the same number of variables, all the variance in each set will be captured by its own *p* = *m* canonical variates.

SAS also gives us the **correlations between each variable and each opposite canonical variate**, under the title "Correlations Between the (set of variables) and the Canonical Variables of the (opposite set of variables). Look at these cross-set loadings for the Homonegativity variables. The MMPI\_1 canonical variate "explains" .19822 of the variance in IAH and .37622 of the variance in SBS, for a total of  of the total variance in the Homonegativity variables. Note that this percentage was computed as a mean of the squared (cross-set) loadings, just as we previously did to find the proportions of variance a canonical variate captured from its own set of variables. This percentage is called a **redundancy**. The redundancy Xj is the proportion of the total variance in the set of X variables that is redundant with (predicted from, "explained" by) the jth canonical variate of the (opposite set) Y variables.

Such a redundancy can also be computed from own-set proportions of variance captured and the canonical correlations. Homoneg\_1 captures .6269 of the variance in its own set of variables. The canonical *r2* (the square of the *r* between MMPI\_1 and Homoneg\_1) is .1442, = indicating that 14.42% of the variance in the Homoneg\_1 canonical variate is explained by its correlation with the MMPI\_1 canonical variate. Well, if .6269 of the variance in the Homonegativity variables is captured by the Homoneg\_1 canonical variate and .1442 of the variance in Homoneg\_1 is explained by MMPI\_1, then (.6269)(.1442) = .0904 of the variance in the Homonegativity variables is explained by the MMPI\_1 canonical variate, which is exactly what SAS tells us under the title "Standardized Variance of the Delinquency Explained by the Opposite Canonical Variables."

In general, you multiply (the proportion of variance that one X canonical variate captures from its own set of X variables) times (the squared canonical correlation between that X canonical variate and the corresponding Y canonical variate) to obtain the amount of variance in the X variables explained by the canonical variate from the (opposite) Y set of variables.

For our data, the MMPI canonical variates explain a total of 12.95% of the variance in the Homonegativity variables (.0904 explained by MMPI\_1, .0391 explained by MMPI\_2). The Homonegativity canonical variates explain a total of 6.30% of the variance in the MMPI variables (.0327 explained by Homoneg\_1, .0303 explained by Homoneg\_2).

SAS completes the canonical redundancy analysis by giving for each variable the *R2* for predicting that variable from the first canonical variate from the opposite set of variables, the first and second canonical variates from the opposite set, the first three canonical variates from the opposite set, etc. For our data, SBS is predicted moderately well by MMPI\_1 (*R2* = .1415) and only very slightly better by the combination of MMPI\_1 and MMPI\_2 (*R2* = .1434). IAH is not as well predicted by MMPI\_1 (*R2* = .0393), but adding MMPI\_2 helps a lot (*R2* = .1155). Note that averaging the IAH-MMPI\_1 and SBS-MMPI\_1 *R2* values gives the redundancy for Homonegativity from MMPI\_1 (proportion of variance in the Homonegativity variables explained by the MMPI\_1 canonical variate). Likewise, averaging the *R2* values for predicting IAH and SBS from MMPI\_1 and MMPI\_2, (.1155 + .1434) / 2 = .1295, gives the total redundancy for the Homonegativity variables predicted from the MMPI canonical variates. The redundancies of the MMPI variables predicted from the Homonegativity canonical variates can likewise be obtained by averaging the *R2* values between the MMPI variables and the "First M Canonical Variables of the Homonegativity."

You should note that it is quite possible for a pair of canonical variates that have a large squared correlation not to explain much of the variance in the variables, that is, the canonical analysis may produce a pair of highly correlated weighted combinations of the variables that extract only a very small amount of the variance in the original variables. There are ways to produce weighted combinations of variables that maximize redundancies rather than canonical correlations, but they will not be presented here.

Do note that redundancies are not symmetrical - an X canonical variate may explain much more (or less) of the variance in the Y variables than does the Y canonical variate in the X variables. For our example, the MMPI canonical variates explain 13% of the variance in the Homonegativity variables, but the Homonegativity canonical variates explain only 6% of the variance in the MMPI variables. Much of the unexplained variance in the MMPI variables is in the PD variable - the *R2* for predicting PD from both Homonegativity canonical variates was only .02.

Note that I created an output data set (Sol) with the four canonical variates. PROC CORR on those data shows that the for each set of variables the correlation between the first canonical variate and the second is absolutely zero.

**SPSS**

Download the following files

* Sunita.sav from my [SPSS Data Page](http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Data.htm).
* Canonical.sps from my [SPSS Programs Page](http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Programs.htm).

Bring Canonical.sps into SPSS. Edit Canonical.sps so that it properly points at Sunita.sav. From the SPSS Syntax Editor, click Run, All.

In SPSS one does canonical correlation by using the MANOVA routine, calling one set of variables DEPENDENT VARIABLES and the other set COVARIATES. Under "EFFECT..WITHIN+RESIDUAL Regression" SPSS gives you essentially the same output that SAS does, formatted differently. Ignore the "EFFECT..CONSTANT" output.

MANOVA IAH SBS with MA MF PD K

/discrim stan corr alpha(1)

/print signif(mult univ eigen dimenr)

/noprint param(estim)

/method=unique

/error within+residual

/design .

IAH and SBS are the “Dependents” and MA, MF, PD, and K are the “Covariates.”

The "ALPHA(1)" statement in the DISCRIM subcommand is needed to force MANOVA to calculate all possible canonical functions, regardless of whether of not they are significant.

The "Univariate F-tests" give us the squared multiple correlation coefficients (SMC's) for predicting IAH and SBS from the MMPI variables. Note that these match the SMC's that SAS gave us for predicting IAH and SBS from MMPI canonical variates 1 and 2. We would get the same SMC’s if we just did two multiple regressions, one predicting IAH from the MMPI variables and the other predicting SBS from the MMPI variables.

Under the title “Variance in dependent variables explained by canonical variables” is the redundancy analysis for the homonegativity variables. In the “Pct Var Dep” column are the percentages of standardized variance in the homonegativity variables explained by their own canonical variates, and in the “Pct Var CO” column are the percentages of standardized variance in the homonegativity variables explained by the opposite (MMPI) canonical variates.

Under the title “Variance in covariates explained by canonical variables” is the redundancy analysis for the MMPI variables. In the “Pct Var DE” column are the percentages of standardized variance in the MMPI variables explained by the opposite (homonegativity) canonical variates, and in the “Pct Var CO” column are the percentages of standardized variance in the MMPI variables explained by their own canonical variates.

I find the SAS output to be easier to read than the SPSS output, especially for the redundancy analysis. The SPSS output used to be even more confusing. In some earlier versions of SPSS the proportions of variance in the X variables explained by the Y canonical variates was mistakenly described as being the proportions of variance in the Y variables explained by the X canonical variates. I explained this problem to the folks at SPSS, and, eventually, they fixed it. I also explained it to the authors of a multivariate statistics text (Tabachnik and Fidell) who made exactly the same error. They never responded to my letter.

I should caution you that sometimes SPSS/PASW will construct a canonical variate that is exactly the opposite of that created by SAS. For example, look at these loadings from SAS, paying special attention to the second canonical variate

Correlations Between the Homonegativity and Their Canonical Variables

Homoneg\_1 Homoneg\_2

IAH 0.5219 -0.8530

SBS 0.9907 0.1361

* High Scores on Homoneg\_1 reflect high scores on the SBS and, to a lesser degree, on the IAH.
* High Scores on Homoneg\_2 reflect high scores on the SBS and low scores on the IAH (notice the negative sign on the loading for IAH)

MMPI\_1 MMPI\_2

MA 0.5326 0.7224

MF -0.4937 0.7638

PD 0.3241 0.2129

K -0.5251 -0.0788

* High scores on MMPI\_1 reflect high scores on MA, high scores on PD, low scores on MF, and low scores on K.
* High scores on MMPI\_2 reflect from high scores on MA and MF.

Now look at the loadings produced by SPSS

Correlations between DEPENDENT and canonical variables

Function No.

Variable 1 2

iah .52190 .85301

sbs .99069 -.13613

Correlations between COVARIATES and canonical variables

CAN. VAR.

Covariate 1 2

ma .53259 -.72236

mf -.49368 -.76384

pd .32410 -.21288

k -.52513 .07876

Note that the second canonical variate of the homonegativity variables is just the opposite of that created by SAS and the same is true of the second canonical variate of the MMPI variables. The interpretation of the correlation is still the same, it is just that instead of saying A is positively related to B we are now saying that NOT A is positively related to NOT B.

This is an annoyance, and can be confusing, but it is otherwise not a problem. Consider this simple example. You have three X variables based on responses to a survey. You expect X1 to be positively associated with a conservative attitude about government spending and X2 and X3 to be negatively associated with that attitude. You also have three Y variables, based on an audit of the subjects’ personal finances. You expect Y1 and Y2 to be positively associated with frugality and Y3 to be negatively associated with frugality.

SAS creates a canonical variate CVx1 = w1X1 – w2X2 – w3X3. High scores on this CV indicate that the subject has conservative attitudes about government spending.

SAS creates a canonical variate CVy1 = -w1Y1 + w2X2 + w3X3. High scores on this CV indicate that the subject is frugal.

The canonical correlation shows that frugality is associated with conservative attitudes about government spending. [This is speculative, on my part, as I know lots of people with such conservative attitudes who are spendthrifts with their own finances).

SPSS creates a canonical variate CVx1 = -w1X1 + w2X2 + w3X3. High scores on this CV indicate that the subject has liberal attitudes about government spending.

SPSS creates a canonical variate CVy1 = w1Y1 - w2X2 - w3X3. High scores on this CV indicate that the subject is wasteful.

The canonical correlation shows that wastefulness is associated with liberal attitudes about government spending. This is really the same conclusion reached upon interpretation of the weights provided by SAS. In the one case we conclude that A is associated with B, and in the other case we conclude that not A is associated with not B.

At any time, if you find it makes the interpretation easier, you should feel free to multiply all of the weights of both canonical variates by minus one.

* [Fair Use of this Document](http://core.ecu.edu/psyc/WuenschK/Fair-Use.htm)
* [SAS Listing](http://core.ecu.edu/psyc/wuenschk/SAS/Canonical-Results.pdf)
* [Return to my statistics lessons page](http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm).

Copyright 2018, Karl L. Wuensch - All rights reserved.

Endnote

1. © Copyright 2013, Karl L. Wuensch - All rights reserved. [↑](#footnote-ref-1)
2. A high Scale 5 score indicates that the individual is more like members of the other gender than are most people. A man with a high Scale 5 score lacks stereotypical masculine interests, and a woman with a high Scale 5 score has interests that are stereotypically masculine. Low Scale 5 scores indicate stereotypical masculinity in men and stereotypical femininity in women. MMPI Scale scores are “T-scores” – that is, they have been standardized to mean 50, standard deviation 10. The normative group was residents of Minnesota in the 1930’s. The MMPI-2 was normed on what should be a group more representative of US residents. [↑](#endnote-ref-1)