Are You Considering All Possible Factors?

A User Friendly Discussion of Canonical Correlation and Commonality Analysis

R. Michael Haynes, PhD
Executive Director, Office of Institutional Research and Effectiveness
Tarleton State University

WHY WOULD STATISTICAL ANALYSES BE USEFUL IN HIGHER EDUCATION AND ENROLLMENT MANAGEMENT?

- Increasing calls for accountability in use of resources…want a measurable return on investment on academic outcomes (retention, graduation rates, SCH attainment, time to degree, etc…)
- What predictor could be a factor in identifying best possible students to achieve optimal outcomes…what predictor could be A FACTOR, not the single factor!

WHY ARE MULTIVARIATE TECHNIQUES SUPERIOR TO UNIVARIATE, SUCH AS BIVARIATE CORRELATION?

- As with independent variables, rarely do dependent variable exist in a silo…correlation with other related measures
- Multivariate allows analysis of multiple independent variables as well as multiple dependent variables (Kroff, 2002; Roberts, 1999; Si, 2001)
- Subsumes univariate measures such as t-tests, ANOVA, etc. in the ability to explain variance across multiple dependent variables
- Minimizes the likelihood of Type I error occurrences (Kane, 2006; Thompson, 1987), or rejecting the null hypothesis when it is true

CANONICAL CORRELATION ANALYSIS (CCA)

- An extension of the general linear model (GLM)
- Conducts correlation analysis of two sets of variables:
  - Independent set = high school rank, SAT score, family income
  - Dependent set = first-year GPA, SCH attempted, SCH completed
  - SPSS creates latent synthetic variables for each set then correlates

CANONICAL CORRELATION ANALYSIS

- Generates canonical roots that identify the variance accounted for between latent, synthetic variables; number of functions generated = number of variables in the smaller set of variable
- Easily performed in SPSS using MANOVA commands;
- MANOVA

  hsrank SAT faminc WITH fyGPA SCHa SCHc
  /PRINT=SIGNIF (MULTIV EIGEN DIMENR)
  /DISCRIM=(STAN ESTIM COR ALPHA (.999))
- In fact, SPSS v. 24 now has a canonical function under “Correlate”

WHAT DO CCA OUTPUTS TELL YOU?

- Canonical roots are analyzed for importance via statistical significance (*F* statistic) and unique variance accounted for (via Wilks Lambda)
- Canonical correlation coefficient for each function/root (Rc)…analogous to Pearson *R* (-1.0 to 1.0)
- Squared canonical correlation for each function (Rc^2)…effect size analogous to R^2 (0.00 to 1.0)
- Raw canonical coefficients which are analogous to *b* weights in regression; each variables weight in the linear equation of the model

- Standardized canonical coefficients which are analogous to *beta* weights in regression; each variables importance in creating the synthetic dependent variable
- Structure coefficients; each variables correlation with the dependent variable (-1.0 to 1.0)
- Squared structure coefficients; how well each variable explains the variance in the synthetic dependent variable (0.0 to 1.0)

- Like regression, CCA is great in identifying relationships between variables…in this case, two sets of variables, right?
- Gives us useful information like b and beta weights, correlation coefficients, effect sizes…but does little in identifying the MOST USEFUL variable in the set!

COMMONALITY ANALYSIS

- Uses an algorithm to decompose the effect size from a regression or canonical model
- Independent variables used with synthetic dependent variable to obtain an R^2…what's neat, you can do it the opposite way with the dependent variables on a synthetic independent variable!!
- Identifies each variable's unique and combine usefulness in explaining variance in the dependent variable

COMMONALITY ANALYSIS

- Formula for number of algorithms is $2^k - 1$, where k = the number of independent variables....if you have 2 IVs, you would have 3 algorithms; if you have 3, you would have 5!

$$Ua = -R^2 + R^2y.b$$
$$Ub = R^2 - R^2y.a$$
$$Cab = R^2 - Ua - Ub$$

- Using more than 4 variables makes it really complicated!!....but Excel spreadsheets can handle it if you can build it!!!

SO, IN CONCLUSION,
HERE'S WHAT I'M HOPING YOU TAKE FROM THIS SESSION…

- Univariate analyses are nice (SAT > College GPA; High School Rank > SCH's Completed, etc.), but multivariate techniques provide a more holistic picture of variables/factors at play

- Then after you conduct your multivariate analysis (MANOVA, CCA, ETC…), commonality analysis can decompose the overall effect size (R^2) to identify which variable(s) do the best job in explaining the variance!

- Tarleton's Office of Institutional Research and Effectiveness is happy to assist if you want to apply these techniques to your own data!

Thank You!

Do you have any questions?

R. Michael Haynes, PhD
(254) 968 -9354
rhaynes@tarleton.edu

Running Head: COMMONALITY ANALYSIS AND ADEQUACY/REDUNDANCY

Commonality Analysis and Adequacy/Redundancy Coefficients: Partitioning Out the Variance in Canonical Correlation Analysis

R. Michael Haynes

University of North Texas

(95)

Abstract

Canonical correlation analysis is useful in determining the variance between synthetic variates created by sets of two or more variables. It is easily conducted through the use of statistical packages such as SPSS or SAS. However, these programs do not provide researchers with summaries related to the variance accounted for by the individual canonical variables. Commonality analysis utilizes algorithms to decompose the R^2 between a variate and set of variables so that their unique and combined usefulness can be determined.

Canonical correlation analysis (CCA) subsumes regression, ANOVA, and t-tests in the general linear model as it allows researchers to measure variance between multiple predictor as well as multiple criterion variables (Kroff, 2002; Roberts, 1999; Si, 2001). This makes CCA an attractive and practical parametric tool as in reality, criterion variables are rarely uncorrelated or independent of other criterion variables (Capraro, 2000; Thompson, 1987). Furthermore, as a multivariate analysis, CCA minimizes the likelihood of Type I error occurrences (Kane, 2006; Thompson, 1987). However, as with multiple regression, CCA does not provide researchers with information pertaining to the unique and combined usefulness of the individual predictor variables.

Stepwise regression provides researchers a methodology to determine a predictor's individual meaningfulness as it is introduced into the regression model (Pedhazur, 1997). However, stepwise regression can lead to serious Type I errors and the selection/entry order into the model can misrepresent a variable's usefulness (Thompson, B., Smith, Miller, and Thompson, W.A.; as cited by Rowell, 1991).

Commonality analysis (CA) provides an effective alternative to determining the variance accounted for by respective predictor variables (Si, 2001). Multiple regression is used to obtain R^2 values between the variate and each individual predictor, then in unique combinations with the other predictors. Through the use of algorithms, a calculated R^2 can be partitioned into 2^k-1 (where k=number of predictors) components representing the unique and combined usefulness of the variables under analysis. For example, for 2 independent variables a and b, the total number of unique and common combinations equals 3. Si demonstrates the required calculation in mathematical format as follows:

$$Ua = -R^2 + R^2y.b$$

$$Ub = R^2-R^2y.a$$

$$Cab = R^2 - Ua - Ub$$

CA was developed in the 1960s for use with regression where one criterion variable is being analyzed (Capraro, 2000). However, many researchers have documented CA's utility in conjunction with CCA (Capraro, 2000; Kroff, 2002; Si, 2001) where multiple criterion variables are considered. By simply collapsing the set of dependent variables into a synthetic criterion variable, the process is identical to that associated with regression (R. K. Henson, personal communication, April 26, 2006). Much like CCA, the process is reversible in that the R^2 being partitioned can represent either the predictor or criterion set of canonical variables.

However, it should be noted that the complexity of commonality analysis increases exponentially with the number of variables considered. For example, in conducting CA with 4 independent variables, 15 unique and combinations of variance accounted for are generated. With 5 or 6 independent variables, the number increases to 31 and 63 respectively. Even utilizing a spreadsheet application to calculate the various coefficients, it is possible the variance will become so disbursed that no variable will surface as a clearly powerful predictor.

Some researchers have suggested factor or cluster analysis as a method of collapsing myriad variables into fewer, more manageable groups (Mood, 1969; Seibold & McPhee, 1979; Wisler, 1972; as cited by Rowell, 1991). However, Rowell also notes that this action defeats the purpose of CA in that the ability to identify the most useful

individual variable is lost. Therefore, it is more practical and parsimonious to limit the number of predictor variables analyzed to four or fewer (Si, 2004).

The present paper will utilize a hypothetical data set and SPSS commands to demonstrate the steps and calculations involved in CCA and a subsequent CA. Summary tables as well as SPSS syntax used to perform the various statistical computations will be provided. Additionally, a brief explanation and discussion of adequacy and redundancy coefficients as they relate to variance accounted for is included.

Canonical Commonality Analysis

A data set provided by SPSS 14.0 (N=406) containing 6 variables associated with automobile design and performance will be used in this illustration. Three predictor variables contain information associated with vehicle design: engine displacement size as measured in cubic inches (engine); automobile weight in pounds (weight); and year of production (year). The criterion variables are associated with subsequent automobile performance: miles per gallon of gasoline consumed (mpg); acceleration as measured in seconds from 0 to 60 miles per hour (accel); and horsepower (horse). Descriptive statistics for the data set are provided in Table 1.

The process begins by conducting a CCA where three canonical functions and their respective Rc^2 coefficients are generated. The SPSS syntax utilized to conduct CCA is provided in the Appendix A. Upon review of the output, a decision regarding the number of functions to analyze further must be made. While all three functions are statistically significant at α=.05, Function 1 accounts for 90.3% of the variance between the synthetic predictor and criterion variables. Furthermore, Function 1 represents 95.9%

of the variance accounted for by the complete canonical model. Therefore, we will limit our CA to the coefficients associated with Function 1.

Table 1

Descriptive Statistics for CCA Dataset

Variable	N	Mean	Standard Deviation
mpg	398	23.51	7.816
accel	406	15.50	2.821
horse	400	104.83	38.522
engine	406	194.04	105.207
weight	406	2,969.56	849.827
year	405	75.94	3.742

Next, a synthetic dependent variable must be created for use in the individual regression models. This requires utilizing SPSS commands (see Appendix A) to obtain standardized scores for all cases of mpg, accel, and horsepower. These scores are then multiplied by the variate standardized canonical coefficients of Function 1. Finally the products are summed to obtain the synthetic variable, which in this example will be labeled CRIT1.

Regression is now conducted on CRIT1 utilizing the 3 predictor variables, individually and in unique combination with each other. Recalling our formula 2^k-1, it is determined that 7 (2 to the 3rd-1) separate regressions will be performed on the synthetic variate CRIT1. SPSS syntax for the unique and combined regressions is provided in

Appendix B. The unique and combined R^2s from the regression models are presented in Table 2. For ease of interpretation, each correlation coefficient has been labeled by its respective code in the CA algorithm.

Table 2

R^2 Coefficients for the Predictor Variables: Uniquely and in Progressive Combination

Variable	R^2
engine (a)	.783
weight (b)	.798
year (c)	.236
engine/weight (a,b)	.818
engine/year (a,c)	.810
weight/year (b,c)	.847
engine/weight/year (a,b,c)	.855

The predictors engine displacement size, vehicle weight, and year of production account for 78.3%, 79.8%, and 23.6% of the variance in the composite criterion variable CRIT1. However, CA allows us to decompose these variances components that will identify unique and common predictive power.

An Excel spreadsheet with the various algorithm formulae was created to aid in the completion of the CA summary table. By simply entering the R^2 for each regression into a specified cell of the worksheet, the calculation of unique and common variance can be calculated simply and instantaneously. The results of the CA are presented in Table 3.

Table 3

Unique and Common Variance Accounted for by the Predictor Variables, Engine Displacement, Vehicle Weight, and Year of Production

Grouping	engine	weight	year
Ua	.008		
Ub		.045	
Uc			.037
Ca,b	.566	.566	
Ca,c	.012		.012
Cb,c		-.010	-.010
Ca,b,c	.197	.197	.197
TOTAL	.783	.798	.236
Unique	.008	.045	.037
Common	.775	.753	.199

Engine, weight, and year uniquely account for .08%, 4.5%, and 3.7% respectively of the variance of CRIT 1. These percentages are relatively small when compared to their corresponding common predictive powers of 77.5%, 75.3%, and 19.9%. Furthermore, in total, engine accounts for only 2.2% more total variance than weight. This is evidenced by the combination Ca,b accounting for 56.6% of variance in CRIT1.

From this summary table, we can conclude that engine displacement size or vehicle weight could perform equally well in predicting vehicle performance. In addition, both are better predictors of vehicle performance than year of production. This

is the benefit of CA in that predictor variables can be analyzed for usefulness and retained or discarded from consideration based on a pre-determined level of variance accounted for. It should be mentioned that the negative coefficient of combination Cb,c is most likely due to a suppressor effect between the variables weight and year (Pedhazur, 1997).

Adequacy and Redundancy Coefficients

Additional information about the sets of criterion and predictor variables can be gained through exploring their adequacy and redundancy coefficients. Adequacy coefficients are the average squared structure coefficient of the canonical variable set. They indicate, on average, how much of the variance within the canonical variables is contained in the synthetic variate (Thompson, 1987). SPSS calculates adequacy coefficients for each canonical function and includes them as part of the standard output. They can be located under the "Variance in dependent variable/co-variates explained by canonical variables" section of the output and labeled "Pct Var DE".

The redundancy coefficient was created by Stewart and Love in 1968 (Stephens, 2002) and is the product of each variable set's adequacy coefficient and the Rc^2. Each variable set's index represents the amount of variance accounted for in the corresponding variable set. Adequacy and redundancy coefficients for our data set as well as structure coefficients are presented in Table 4.

Adequacy coefficients for the predictor and criterion groups are .639 and .672 respectively. This would indicate that the synthetic variables account for high levels of variance in each canonical set. However, it is noteworthy to remind readers these are merely averages and the coefficients should come as no surprise upon review of the

individual squared structure coefficients. Ultimately, adequacy coefficients are simply a measure of central tendency and this should be kept in mind when attempting to interpret their meaning.

Table 4

Structure and Squared Structure Coefficients of Predictor and Criterion Variables

Variable	Struct Coef	Sq Struct Coef
Mpg	-.932	.868
Accel	-.516	.266
Horse	.939	.882
Adequacy (criterion)		.672
Rd (criterion)		.607
Rc^2 (Function 1)		.903
Rd (predictors)		.577
Adequacy (predictors)		.639
Engine	.955	.912
Weight	.964	.929
Year	-.530	.281

The resulting redundancy coefficients indicate the predictor group accounts for .577 of the variance in the criterion group, while the criterion group accounts for .607 of the variance in the predictor group. This is an example of the inherent problem of utilizing redundancy coefficients in a CCA. Roberts (1999) notes that unlike the Rc^2 of a canonical function, redundancy coefficients are rarely symmetrical (e.g. Rc^2 criterion =

Rc^2 predictor and vice a versa) and therefore inappropriate for inclusion with CCA. He further notes that redundancy coefficients are univariate in nature and should not be used to interpret multivariate data.

Discussion

CCA allows researchers to analyze phenomenon while considering the interactions of myriad criterion and predictor variables. However, CCA does little in identifying the usefulness of the individual variables in predicting outcomes or explaining variance. CA provides a methodology for partitioning unique and combined variance accounted for by either predictor or criterion variables. Much like factor analysis, this provides for more proficient models with better prediction incorporating fewer variables. While adequacy and redundancy coefficients attempt to add to the information extracted from the canonical variable set, researchers should be mindful that these are univariate statistics making inferences about multivariate results.

References

Capraro, R. M. (2000, November). *Commonality and the common man: Understanding variance contribution to overall canonical correlation effects*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY. (ERIC Document Reproduction Service No. ED449232)

Kane, Jr., R. K. (2006, February). *Making sense of the array of coefficients in canonical correlation analysis.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Kroff, M. W. (2002, February). *Commonality analysis: A method of analyzing unique and common variance proportions.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED463309)

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). United States: Wadsworth.

Roberts, J. K. (1999, January). *Canonical redundancy (Rd) coefficients: They should (almost never) be computed and interpreted.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED428077)

Rowell, R. K. (1991, January). *Partitioning predicted variance into constituent parts: How to conduct commonality analysis.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED328589)

cite Alberg & Hansen (2005) for CCA advice

Si, CF. B. (2001, February). *Understanding variance contributions to overall canonical correlation effects: Canonical commonality analysis.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED451209)

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwaw, NJ: Lawrence Erlbaum Associates.

Thompson, B. (1987, April). *Fundamentals of canonical correlation analysis: Basics and three common fallacies in interpretation.* Paper presented at the annual meeting of the Society for Multivariatet Experimental Psychology, Southwestern Division, New Orleans, LA. (ERIC Document Reproduction Service No. ED282904)

Appendix A

SPSS Syntax for Conducting Canonical Correlation Analysis and Creating Synthetic Criterion (CRIT1) Variable

```
DESCRIPTIVES
  VARIABLES=mpg accel horse engine weight year /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX.

MANOVA
    mpg accel horse WITH engine weight year
    /PRINT=SIGNIF (MULTIV EIGEN DIMENR)
    /DISCRIM=(STAN ESTIM COR ALPHA (.999)).

LIST VARIABLES=zmpg zaccel zhorse zengine zweight zyear.

COMPUTE crit1 = (-.460*zmpg) + (.177*zaccel) + (.706*zhorse).
EXECUTE.
```

Note: -.460, .177, and .706 = standardized canonical coefficients of the criterion variables in Function 1

Appendix B

SPSS Syntax for Obtaining Unique and Combined R^2s for Commonality Analysis

```
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT crit1
  /METHOD=ENTER engine.
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT crit1
  /METHOD=ENTER weight.
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT crit1
  /METHOD=ENTER year.
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT crit1
  /METHOD=ENTER engine weight.
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT crit1
  /METHOD=ENTER engine year.
REGRESSION
  /MISSING MEANSUB
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT crit1
  /METHOD=ENTER weight year.
```

Appendix B *(continued)*.

```
REGRESSION
 /MISSING MEANSUB
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT crit1
 /METHOD=ENTER weight year engine.
```