MWSUG 2024 - Paper AL43

%SURVEYCANCORR Macro: Canonical Correlation for Complex Survey

Mark Pickering, Indiana University School of Public Health-Bloomington, Bloomington, IN
Raul Cruz, Indiana University School of Public Health-Bloomington, Bloomington, IN
Taylor Lewis, George Mason University, Fairfax, VA

Abstract

Classical canonical correlation analysis (CCA) is a statistical technique that is widely used when assessing associations between two sets of variables. Although methods exist in SAS® for conducting basic CCA using PROC CANCORR, these methods are limited in their capacities to account for elements in complex survey design (CSD), and no methods have yet been developed that incorporate crucial CSD elements such as clusters or strata and provide meaningful results. Therefore, the accuracy of statistical significance assessments is limited when studying correlations involving datasets that contain CSD elements. To address this limitation, we developed the %SURVEYCANCORR macro. This paper outlines the theoretical basis of our macro and provides a comprehensive description of its options. Applying our macro to a national survey dataset demonstrates the different conclusions that can be drawn depending on whether CSD elements are accounted for when assessing the statistical significance of canonical correlations.

Canonical Correlation Analysis and Incorporating CSD

Classical canonical correlation analysis (CCA) is an exploratory statistical method that enables the analysis of associations between two sets of variables. Compared with other techniques, CCA is advantageous, offering improved interpretability of analyses (Thompson, 1984). CCA attempts to maximize correlations among both sets of variables, offering a very direct and simple measure of both the strength of the relationship and its nature. CCA has been applied in fields as varied as biology (Sherry & Henson, 2005), psychology (Thompson, 1984), economics (Lütkepohl, 2005), education (Tatsuoka, 1988), and marketing research (Green et al., 2001), demonstrating its broad utility.

Although current methods exist for CCA inference and interval estimation, most of these methods assume the use of simple random sampling, which can lead to inaccurate results when data is collected using other sampling methods, such as complex survey design (CSD). Moreover, existing CCA methods may not align with published methodological guidelines for assessing studies that use CSD, leading to results without practical significance. To address these limitations, we developed an innovative CSD CCA method designed to account for intricacies in data collected using CSD. Our method generalizes the existing PROC CANCORR methodology to allow the incorporation of CSD elements into calculations, enabling the acquisition of more accurate results.

CCA attempts to identify linear combinations of the variables within two separate sets, $X = \{X_1, X_2, ..., X_p\}$ and $Y = \{Y_1, Y_2, ..., Y_q\}$ (Hotelling, 1936), aiming to maximize the correlation between these two sets, which provides information regarding the extent and nature of this correlation. The resulting canonical variates are examined to uncover meaningful associations between variables, with the strengths and directions of these associations indicated by the magnitudes and signs of the canonical coefficients. Although CCA is often stopped after the first canonical correlation has been determined, additional correlations can be computed to reveal any underlying relationships between variable sets that are not captured by the first canonical correlation.

When working with data collected using CSD methods rather than simple random sampling, analyses must consider two key factors: (1) survey weights, which are used to adjust the sample to more accurately reflect the population under study by correcting for unequal selection probabilities, ensuring that population parameter estimates are accurate (Hahs-Vaughn et al., 2011; Lewis, 2016); and (2) CSD elements, such as clusters, and strata, which must be considered to account for the non-independence of

sampling and are used to obtain accurate estimates of standard errors, confidence intervals, and *p*-values (Hahs-Vaughn et al., 2011).

Using methods intended for simple random samples for the analysis of stratified samples typically leads to the overestimation of standard errors, whereas applying these methods to clustered samples typically leads to the underestimation of standard errors. Ignoring survey weights can result in incorrect point estimates (Lumley, 2004b). However, updated versions of statistical software, including SAS (PROC CANCORR), Stata (canon command), and **R** (candisc::cancor() function), allow for the inclusion of survey weights when calculating weighted canonical correlations and corresponding *p*-values.

For example, the PROC CANCORR procedure includes methods, such as Wilks' Lambda, in its base functionality. However, these methods still do not allow for the inclusion of other key CSD elements, such as clusters or strata, and ignoring these elements may result in inaccurate *p*-value estimates in weighted canonical correlations. There exists a user written which addresses the CSD factors (Nelson et al., 2020) but, when used on national health surveys, generally yields nominally significant *p*-values for all canonical correlations, even when their magnitude and practical relevance are minimal.

To account for these additional CSD elements, the **Survey CC** method was proposed (Cruz-Cano et al., 2024c). This method calculates degrees of freedom using methods in line with the recommendations for linear regression models in CSD methodological guidelines, ensuring more consistent and meaningful results when analyzing data collected using CSD methods. The **Survey CC** method was first implemented in the **R** package SurveyCC (Cruz-Cano et al., 2024b).

Methodology

The **Survey CC** method is based on three fundamental concepts underlying the data collected using CSD methods: (1) survey weights are sufficient for calculating parameter estimates, such as canonical coefficients; (2) linear combinations of variables in datasets collected using CSD methods have the same structures as the original variables; and (3) the *p*-values in simple linear regressions and correlations are the same even when accounting for CSD elements.

Survey CC Step-by-Step Method:

1. Select variable sets (*X* and *Y*) from the datasets.
2. Include survey weights when calculating point estimates of weighted canonical coefficients and correlations.
3. For each canonical correlation (1 to *q*):
    - Calculate weighted canonical variates, U_j and V_j
    - Perform a linear regression of U_j on V_j, while including all CSD elements (survey weights, clusters, and strata), and obtain the corresponding *p*-values.
    - Perform a linear regression of V_j and U_j, while including all CSD elements, and obtain the corresponding *p*-values.
    - Compare the *p*-values from the two regression models and select the larger value.

Existing CCA procedures that account for survey weights may be able to identify weighted canonical coefficients and weighted canonical variates, after which *p*-values can be calculated using complex survey linear regression techniques. Complex survey linear regression techniques implemented by modern statistical software, including Stata and R, allow for the inclusion of survey weights, clusters, and strata. The proposed macro, %SURVEYCANCORR, performs the necessary matrix operations to find the canonical variates U_j and V_j, after which regression analysis can be performed using PROC SURVEYREG, which allows for the inclusion of CSD elements (the values of these elements are typically available in user guides and manuals datasets collected using CSD methods). This macro allows users to include CSD elements without needing to perform these calculations themselves.

Advantages of Survey CC:

Traditional methods have been criticized for testing all correlations together (Harris, 1976), as this approach can lead to misleading conclusions. Unlike traditional methods, the **Survey CC** method allows the statistical significance of each canonical correlation to be assessed separately, providing individual *p*-values for each canonical correlation.

This approach allows separate conclusions to be drawn for each canonical correlation, which is particularly important when assessing datasets collected using CSD methods because the effective sample size can vary for each correlation when accounting for CSD factors. As a result of this variation, a smaller second canonical correlation might still be statistically significant, even if the first canonical correlation is not significant.

This also implies that the *p*-values obtained using **Survey CC** are not directly comparable to those obtained using classical CCA methods, as **Survey CC** uses a different set of hypotheses than classical CCA methods, a general scientist may interpret these results differently than they would when using traditional methods.

The %SURVEYCANCORR Macro

The developed %SURVEYCANCORR macro was designed to extend the functionality of PROC CANCORR by implementing the proposed **Survey CC** algorithm. As mentioned above, the **Survey CC** algorithm calculates both the correlations among canonical variates and their corresponding statistical significance via an equivalent sequence of univariate linear regressions. The %SURVEYCANCORR macro calculates the necessary test statistics, degrees of freedom (DF), and *p*-values for estimating and interpreting the statistical significance of canonical correlations when accounting for CSD elements. The results are displayed in tables as "%SURVEYCANCORR Canonical Correlations for U_j and V_j". This macro allows the user to leverage existing theoretical and computational resources to integrate CSD elements into regression models (Valliant and Dever [2018]). The **Survey CC** algorithm can integrate the same CSD elements as PROC SURVEYREG, hence the main function of the %SURVEYCANCORR macro requires the input of variables related to survey weights (weight_var), strata (strata_var), and clusters (cluster_var). The other values in the results table produced by the %SURVEYCANCORR macro, which include StdErr, tValue, DF, and Probt, are provided by SAS when using PROC SURVEYREG with CSD elements.

Unlike the **SurveyCC** package implemented in **R** (Cruz-Cano et al., 2024a), which can deal with cases that include replicate weights, the proposed %SURVEYCANCORR macro can not handle replicate weights. However, it can handle the generalized case where none of the CSD elements are present, thus meaning that it would give the same output as someone running PROC CANCORR with the same variables.

For more information on how to obtain the macro, contact the authors of this paper.

Syntax

The %SURVEYCANCORR macro has one user-facing function, %SURVEYCANCORR(), with the following arguments:

- **`input_dataset`**: This argument specifies the dataset containing the variables to be analyzed, which should be accessible in the SAS environment.
- **`x_vars, y_vars`**: These arguments define the names of the variables in the first and second variable sets. Both arguments should be space-separated lists of the variable names present in the dataset. x_vars represents one set of related measurements (e.g., physiological variables), whereas y_vars represents another set of related measurements (e.g., demographic variables).

- **`weight_var`**: This required argument specifies the name of the weight variable used in the survey analysis. If the dataset involves data collected using CSD methods, the weights ensure that each observation is appropriately represented.
- **`strata_var`**: This required argument specifies the name of the strata variable in the dataset, which accounts for the survey's stratified sampling design.
- **`cluster_var`**: This required argument specifies the name of the cluster variable, which indicates how the data are grouped.
- **`id_var`**: This required argument is used to identify individual observations uniquely and helps keep track of each unit when interpreting analysis results.
- **`x_name, y_name`**: These required arguments specify labels for the sets of variables defined by x_vars and y_vars. For example, if x_vars represents physiological variables, x_name could be set to "Physiological." Similarly, y_name could be set to "Demographic" to describe the demographic variables in y_vars.

Output

The %SURVEYCANCORR() macro output in SAS is a set of tables with the following structure:

- **Canonical Correlation Results**:
  - For each canonical variate pair (e.g., V_1, V_2...), the output includes parameter estimates, standard errors, t-values, DFs, and *p*-values (Probt), which are displayed in tabular form, allowing relationships between variable sets to be assessed.
- **Canonical Correlation Coefficients**:
  - The output also includes canonical variate coefficient estimates for each set of variables, which can be used to interpret the contributions of each original variable to the canonical variates.

This very simple and concise output is designed to provide the most relevant information needed to assess the relationships between the variable sets under the CSD framework, providing key statistical measures such as parameter estimates, significance tests, and standard errors.

Example

We present an example illustrating the different aspects and uses of the %SURVEYCANCORR macro. We use the publicly available 2007–2010 National Health and Nutrition Examination Survey (NHANES), which includes a complete set of CSD factors (e.g., survey weights, clusters, and strata) to demonstrate the application of **Survey CC**. The data set is available at https://wwwn.cdc.gov/nchs/nhanes/Default.aspx (Curtin LR, 2013), and we have also included these data in a supplementary file for easy reader access. This dataset was generated by merging the 2007–2008 and 2009–2010 surveys in accordance with the specifications provided in the "National Health and Nutrition Examination Survey: Estimation Procedures, 2007–2010 Data Evaluation and Methods Research" from the CDC's National Center for Health Statistics.

We applied the **Survey CC** method to this dataset to evaluate the relationship between demographic factors (i.e., RIDAGEYR: Age in years at the time of the screening interview, and INDHHIN2: Estimated total household income) and obesity-related body measurements (i.e., BPXPLS: Resting pulse per minute, BPXSY1: Systolic blood pressure, and BPXDI1: Diastolic blood pressure) among individuals 45–64 years of age. This analysis considers appropriate CSD factors, including survey weights, clusters, and strata. Although additional demographic factors, such as race/ethnicity and education, could also be included in a more comprehensive study, this example focuses on a relatively straightforward analysis while incorporating CSD factors.

NHANES Example Setup

To use the %SURVEYCANCORR macro, the user inputs the dataset, variables, and names, which are then passed into the main function, %SURVEYCANCORR**.**

```
%SURVEYCANCORR(
  input_dataset=mydata2.nhanes,
  x_vars=BPXPLS BPXDI1 BPXSY1,
  y_vars=RIDAGEYR INDHHIN2,
  x_name=Physiological,
  y_name=Demographic,
  weight_var=WTMEC4YR,
  strata_var=SDMVSTRA,
  cluster_var=SDMVPSU,
  id_var=SEQN
);
```

NHANES Example Results

The %SURVEYCANCORR macro estimates the canonical correlation *p*-values for the NHANES example. Figure 1 provides detailed information regarding canonical correlation estimates and *p*-values when using PROC CANCORR with the weights statement.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Test of H0: The canonical correlations in the current row and all that follow are zero				
					Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.547687	0.547332	0.010039	0.299962	0.69918692	317.33	6	9718	<.0001
2	0.034878	0.031954	0.014323	0.001216	0.99878354	2.96	2	4860	0.0519

Figure 1. Abbreviated output using PROC CANCORR

When using PROC CANCORR, which only includes the weight statement, the second canonical correlation is not significant at the 0.05 level. Figure 2 provides detailed information regarding canonical correlation estimates, *p*-values, standard error, *t*-statistic, and DF when using the %SURVEYCANCORR macro while including all CSD factors (e.g., survey weights, clusters, and strata).

SURVEYCANCORR Canonical Correlations for Uj and Vj

Parameter	Estimate	StdErr	tValue	df	Probt
V1	0.54769	0.016018	34.1913	32	0.000000
V2	0.03488	0.014844	2.3497	32	0.025123

Figure 2. Output using %SURVEYCANCORR macro

When implementing the **Survey CC** algorithm with the proposed %SURVEYCANCORR macro, the *p*-value for the second canonical correlation is significant at the 0.05 level, indicating that the inclusion of

all CSD elements alters the calculation of *p*-values and potentially changes the conclusions that can be drawn from the analysis. In the PROC CANCORR example, which only considers the weight statement, the second canonical correlation does not appear to provide a statistically significant additional dimension of association, which may lead to the conclusion that the first canonical correlation captures the only strong relationship between the two variable sets. However, when using the proposed macro, the second correlation appears to provide an additional, independent association between the two sets of variables, which may lead to a different conclusion.

This difference is critical. Although the first canonical correlation is generally thought to capture the strongest linear association between two variable sets, the significance of additional (e.g., second, third, fourth) canonical correlations can have major implications. In the above example, a user using PROC CANCORR might decide that the first canonical correlation is the only major association, allocating the majority of resources toward attempting to identify which variables contribute most to the first pair of canonical variates. However, a user using the **Survey CC** algorithm might instead reach the conclusion that the presence of multiple significant canonical pairs indicates that the relationship between variable sets is multifaceted, involving several underlying patterns, which may eventually lead to richer insights regarding how the variable sets interact.

Conclusion

The accurate calculation of canonical correlations when assessing datasets obtained using CSD methods requires the consideration of CSD elements (i.e., survey weights, clusters, strata). Incorporating CSD elements is particularly critical for CCA, a multivariate method that is particularly useful when examining the associations among sets of variables. Our proposed %SURVEYCANCORR macro allows a user to both account for CSD elements and separately test the statistical significance of each canonical correlation rather than assessing the statistical significance of a collection of canonical correlations. Comparing the results of the %SURVEYCANCORR macro with those of classic statistical tests demonstrates that using the macro enables the user to completely and accurately assess the statistical significance of each canonical correlation, which may lead to different conclusions when examining the canonical correlation structure.

A primary limitation of the %SURVEYCANCORR macro in its current form is that the output remains rudimentary and basic. Although the output provides useful information for drawing conclusions, especially in cases involving CSD elements, other software, such as the SurveyCC package in **R**, is able to provide more detailed information such as the *p*-values for secondary canonical correlations beyond the second one, provide *p*-values based on several classical statistical methods and the chance to have the units and variables graphs drawn. Users of our macro should be mindful of this limitation when assessing the statistical significance of canonical correlations.

To our knowledge, %SURVEYCANCORR macro represents the first method that enables the incorporation of CSD factors when calculating the statistical significance of all canonical correlations using the exact adjustments as they are specified in the methodological documentation of the surveys under study. However, other approaches may be able to combine or modify existing statistical procedures to achieve this objective.

We presented an example in which the relationships that exist between two groups of variables from a dataset collected using CSD methods were examined using the %SURVEYCANCORR macro, demonstrating its versatility and usefulness of the macro, and indicating that it can be used with various national datasets that include CSD elements. These surveys contain a wealth of information, and proper examination of the resulting data that accounts for CSD methodology is essential for determining the significance of relationships within these datasets. Our macro represents an important tool, facilitating the accurate assessment of the statistical significance of canonical correlations. In addition to NHANES, we expect that our macro will be useful when analyzing other national and international surveys that collect data using CSD methodologies, such as the Behavioral Risk Factor Surveillance System, the Medical

Expenditures Panel Survey, the Health Information National Trends Survey, and the Tobacco Use Supplement to the Current Population Survey.

In conclusion, the %SURVEYCANCORR macro can be used to extract accurate information regarding the statistical significance of relationships between sets of variables in datasets generated using CSD methods, with the potential to support researchers using these datasets to address important scientific questions.

References

1. Centers for Disease Control and Prevention (CDC). n.d.. *National Health and Nutrition Examination Survey (NHANES).* National Center for Health Statistics. https://wwwn.cdc.gov/nchs/nhanes/Default.aspx (accessed October 12, 2024).
2. Cruz-Cano, R. 2024a. "SurveyCC: Canonical Correlation for Survey Data." *R package Version 0.2.1*. https://CRAN.R-project.org/package=SurveyCC (accessed October 12, 2024).
3. Cruz-Cano, R., Cohen, A., & Mead-Morse, E. 2024b. "Canonical Correlation Analysis of Survey Data: The SurveyCC R package." Manuscript accepted for publication in *The R Journal*.
4. Cruz-Cano, R., et al. 2024c. "A proposed algorithm for handling canonical correlation analysis with complex survey design (CSD) elements." *Journal of Survey Statistics*.
5. Curtin, L. R. 2013. *National Health and Nutrition Examination Survey: Estimation procedures, 2007–2010 data evaluation and methods research.* National Center for Health Statistics, Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nchs/nhanes/Default.aspx (accessed October 12, 2024).
6. Nelson, D. R., Wong-Jacobson, S. H., & Lilly, E. 2020. "%surveycorrcov macro: Complex survey data correlations for multivariate analysis and model building." https://api.semanticscholar.org/CorpusID:215747405 (accessed October 12, 2024).
7. Green, P. E., Tull, D. S., & Albaum, G. 2001. *Research for marketing decisions* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
8. Harris, R. J. 1976. "Tests of significance in canonical correlation analysis." *Psychological Bulletin*, 83(4), 846–856.
9. Hahs-Vaughn, D. L., McWayne, C. M., Bulotsky-Shearer, R. J., Wen, X., & Faria, A. M. 2011. "Variable selection in multilevel models: Using survey weights in regression analysis with complex survey data." *Educational and Psychological Measurement*, 71(5), 849–880.
10. Lewis, T. 2016. "Survey weights and the correction of unequal selection probabilities." *Journal of Official Statistics*, 32(3), 569–589.
11. Lütkepohl, H. 2005. *New introduction to multiple time series analysis.* Berlin, Germany: Springer-Verlag.
12. Lumley, T. 2004. "Analysis of complex survey samples." *Journal of Statistical Software*, 9(1), 1–19.
13. National Center for Health Statistics (NCHS). 2013. *National Health and Nutrition Examination Survey: Estimation Procedures, 2007–2010* (DHHS Publication No. 2013-1350). U.S. Department of Health and Human Services.
14. Sherry, A., & Henson, R. K. 2005. "Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer." *Journal of Personality Assessment*, 84(1), 37–48.
15. Tatsuoka, M. M. 1988. *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York, NY: Macmillan Publishing Company.
16. Thompson, B. 1984. *Canonical correlation analysis: Uses and interpretations.* Beverly Hills, CA: SAGE Publications.
17. Valliant, R., & Dever, J. A. 2018. *Survey weights: A step-by-step guide to calculation.* Cambridge, UK: Cambridge University Press.

Contact Information

Your comments and questions are valued, encouraged, and welcomed. Please contact the author at:

Mark Pickering
Indiana University School of Public Health-Bloomington

(779)-279-4370
markpick@iu.edu

Raul Cruz-Cano
Indiana University School of Public Health-Bloomington
(812)-855-2235
raulcruz@iu.edu