Canonical Correlation Analysis of Survey Data: The SurveyCC R Package

by Raul Cruz-Cano, Aaron Cohen, and Erin Mead-Morse

Abstract Classic Canonical Correlation Analysis (CCA) is a popular statistical method that allows for the analysis of the associations between two sets of variables. However, currently it cannot be applied following the published methodological documentation to data sets collected using complex survey design (CSD), which includes factors, such as replicate weights, clusters, and strata, that are critical for the accurate calculation of the statistical significance of any correlation. To close this gap, we have developed the Survey CC algorithm and implemented it in an R package. We describe the theoretical foundations of our algorithm and provide a detailed report of the options of the function that performs it. Moreover, the application of our newly developed method to several national survey data sets shows the differences in conclusions that can be reached if the CSD elements are not taken into consideration during the calculation of the statistical significance of the canonical correlations.

1 Introduction

Classical Canonical Correlation Analysis (CCA) is a popular exploratory statistical method that allows for the analysis of relationships between two sets of variables, offering several advantages over other statistical techniques, including the ability to limit the probability of committing Type I errors when examining more than two variables (Hair et al., 2019). CCA has been applied to diverse fields, including ecology (Gittins, 2012), neuroscience (Zhuang and Yang, 2020), biomedical imaging (Correa et al., 2008), human geography (Clark, 1975), effect of work place regulations (Borland et al., 1991), illicit drug use (Leo and Wulfert, 2013) and tobacco use (Morris et al., 2018).

Current methods for CCA inference and interval estimation, however, depend on simple random sampling, and are, therefore, inaccurate when data is collected using a complex survey design (CSD). Or, they are based on methods that cannot be applied following the published CSD methodological documentation, hence, leading to results without practical significance. To overcome these issues, we will employ our recently developed, innovative Survey CC method.

Using CCA, investigators study the extent and nature of the correlation between two sets of variables, $X = \{X_1, X_2, ..., X_p\}$ and $Y = \{Y_1, Y_2, \ldots, Y_q\}$ (Hotelling, 1936). As described by (Clark, 1975), without loss of generality, it is assumed that $p \geq q$ and that each and every variable in X and Y have been standardized to have a mean zero and a variance equal to one. CCA is used to explore sample correlations between X and Y observed on the same experimental units by analyzing the coefficients, $A_1 = (a_1, a_2, \ldots, a_p)^T$ and $B_1 = (b_1, b_2, \ldots, b_q)^T$, which maximize the correlation between linear combinations of the variables X and Y subject to $Var(XA_1) = Var(YB_1) = 1$.

The "classic" canonical correlation can be defined as (Hotelling, 1936):

$$\rho_1 = max_{A_1,B_1} Corr(XA_1, YB_1) = max_{A_1,B_1} \frac{Cov(XA_1, YB_1)}{\sqrt{Var(XA_1) \cdot Var(YB_1)}} \tag{1}$$

subject to $Var(XA_1) = A_1^T X^T XA_1 = B_1^T Y^T YB_1 = 1$, as described by (Hardoon and Shawe-Taylor, 2011). These linear combinations, $U_1 = XA_1$ and $V_1 = YB_1$, are called the first canonical variates. Their correlation, ρ_1, is the first canonical correlation, and the coefficients A_1 and B_1 are the first canonical coefficients.

The first canonical coefficients in each set will help to identify variables from X and Y that are associated with each other. Variables with a strong positive relationship will have coefficients with large magnitudes and similar signs (positive or negative), whereas inverse relationships will have coefficients with opposing signs. The coefficients' magnitudes are commonly examined to determine which variables are the most relevant as variables with small coefficients will have weak relationships.

The process can be duplicated to identify the coefficients, A_2 and B_2, that maximize the correlation, ρ_2, subject to $Var(XA_2) = Var(YB_2) = 1$, with the additional constraint that the new canonical variates, $U_2 = XA_2$ and $V_2 = YB_2$, are uncorrelated with the first pair of canonical variates. This process can be repeated q times. These secondary canonical correlations are important because not all the significant relationships that exist among variables can be expressed in the first canonical correlation.

Many researchers base their conclusions regarding the relationships among the variables examined in a CCA on the canonical coefficients associated with statistically significant canonical correlations

(Boedeker and Henson, 2020) (Hill and Lewicki, 1997), hence, tools able to accurately estimate their p-values are necessary. For example, (Travis and Lagrosen, 2014) limited their conclusions to the information inferred from the first set of canonical coefficients because only the first canonical correlation had a significant p-value. In (Stowe et al., 1980), three canonical correlations have statistically significant p-values, hence all three sets of canonical coefficients are studied.

For the sample cross-correlation matrix, $\left(\begin{smallmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{smallmatrix}\right)$, canonical correlations can be calculated by finding the descending-ordered square roots of the matrix eigenvalues $\mathbf{M} = S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY}$. The canonical coefficients for Y are the corresponding q eigenvectors, and the elements of eigenvector i are the canonical coefficients B_i. Correlations between canonical variates and original variables (called canonical factor loadings) offer information that is complementary to canonical coefficients and, hence, are preferred by some researchers (Meredith, 1964).

2 Weighted CCA

Broadly speaking, when data sets have been collected using CSD methods, two main types of factors are used to account for the fact that the sample is not a simple random sample:

- Survey weights are used to create a synthetic sample for which the distribution of covariates is similar to the population under study, i.e., they help to account for unequal selection probability (Hahs-Vaughn et al., 2011), leading to correct population parameter estimates (Lewis, 2016).
- Repetition weights, clusters, and strata are CSD elements essential to consider the nonindependence among the sampling units derived from the complex sampling design utilized to gather the data (Hahs-Vaughn et al., 2011). Hence, they are used during the calculation of standard errors to accurately estimate confidence intervals and p-values for any hypothesis tested.

Standard statistical methods that ignore structures, such as clustering or stratification, can give seriously misleading results (Holt et al., 1980). Analyzing a stratified sample as if it were a simple random sample tends to overestimate the standard errors, whereas analyzing a clustered sample as if it were a simple random sample tends to underestimate the standard errors, and analyzing the data without taking into consideration the survey weights will lead to incorrect point estimates (Lumley, 2004). Examples of how intracluster correlations can severely affect the effective sample size can be found in (Killip et al., 2004).

From (Winship and Radbill, 1994), we know that the formulas for weighted covariance and variance for two variables, X_1 and X_2, are:

$$Weighted\ Cov(X_1, X_2) = \frac{\sum_{i=1}^{n} W_i (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sum_{i=1}^{n} W_i} \tag{2}$$

$$Weighted\ Var(X_1) = \frac{\sum_{i=1}^{n} W_i (X_{i1} - \bar{X}_1)^2}{\sum_{i=1}^{n} W_i} \tag{3}$$

where W is the vector that stores the survey weight for each of the n subjects in the sample.

If we choose A_j and B_j, such that $WeightedVar(U_j) = WeightedVar(V_j) = 1$, then substituting (2) and (3) into (1) reduces the weighted canonical correlation to

$$Weighted\ \rho_j = max_{A_j,B_j} \frac{\sum_{i=1}^{n} W_i ((XA_j)_i (YB_j)_i)}{\sum_{i=1}^{n} W_i} \tag{4}$$

It can be proven that the q canonical coefficients A_j are the q eigenvectors of:

$$|(X^T diag(W) Y) \cdot (Y^T diag(W) Y)^{-1} \cdot (Y^T diag(W) X) \cdot (X^T diag(W) X)^{-1} - \lambda^2 I| = 0 \tag{5}$$

The q canonical coefficients B_j can be found similarly. As in the classic CCA, the q eigenvalues, λ, are the weighted canonical correlations; hence, parameters A and B are not needed to find the weighted canonical variates and correlations, as this calculation derives their values.

Modern statistical software, including SAS 9.4 software's PROC CANCORR (SAS, 2012), Stata's canon command (StataCorp, 2023b), and R's `cancor()` function in the **candisc** package, apply these formulas and are able to consider survey weights in their calculations of weighted canonical correlations and their corresponding p-values according to the methods Wilks' lambda, Pillai's trace, and Hotelling-Lawley trace (Caliński et al., 2006) and Roy's largest root (Johnstone, 2009). However, these programs are unable to account for other factors, such as clusters and strata. As mentioned previously, ignoring these complex survey design elements will lead to different estimates of p-values for weighted canonical correlations.

The SAS macro presented in (Nelson et al., 2020) accounts for the CSD factors by calculating a design-effect adjusted weight for each variable. The way in which these calibrated weights are calculated tends to result in design-effect adjusted sample sizes large enough to consistently provide trivially significant results, i.e., statistically significant p-values for canonical correlations regardless of their magnitude and practical meaning. The authors of this algorithm provide a recommendation to ameliorate this problem but their advice requires the user to choose a random variable as the dependent variable, hence leading to inconsistent results that are influenced by this selection. In contrast, our proposed algorithm leads to the same degrees of freedom recommended for linear regression models outlined in the complex surveys' guidelines.

3 Complex survey design CCA

The proposed Survey CC method is based on three well-known ideas related to data collected using CSD: (1) the survey weights are sufficient to calculate correct parameter estimates, e.g., the canonical coefficients; (2) the linear combination of variables in a complex survey data set follows the same CSD structure as the original variables; and (3) the p-values for a simple linear regression coefficient and a correlation using the same two variables are equal.

In the case of (3), due to the linearization of the variance estimator used by `survey::svyglm()`, the p-value for the slope of the regression of U_j on V_j will not be equal to the p-value for the regression of V_j on U_j, as would be the case if there were no CSD elements. We addressed this issue by calculating both p-values and then selecting the largest of them, i.e., we took a conservative approach for the test of the hypothesis $H_0 : \rho_j = 0$ (Nelson et al., 2020).

Survey CC step-by-step methodology:

1. Select variable sets, X and Y, from existing data sets.
2. Include survey weights in the calculation of the weighted canonical coefficients and correlations to find their point estimates.
3. For each j canonical correlation where $1 \leq j \leq q$:
   (a) Use weighted canonical coefficients to calculate weighted canonical variates, U_j and V_j.
   (b) Perform a simple linear regression of U_j on V_j that includes all complex survey design elements (e.g., survey weights, cluster, and strata) to find the p-values for their corresponding r_j canonical correlation
   (c) Perform a simple linear regression of V_j on U_j that includes all complex survey design elements (e.g., survey weights, cluster, and strata) to find the p-values for their corresponding r_j canonical correlation
   (d) Compare the p-values of the linear regression models in the previous two steps and select the largest.

In practice, existing computational CCA procedures that consider survey weights can be used to identify weighed canonical coefficients, which, in turn, can be used to calculate the weighted canonical variates. Once these canonical variates have been identified, given that they are just linear combinations of the original variables, p-values for canonical correlations can be estimated using complex survey linear regression procedures, as linear regression survey procedures in modern statistical software can include not only survey weights but also information regarding the repetition weights, clusters, and strata from the original data set. The values required for these CSD factors are provided in the user guides, codebooks, and manuals of the surveys that would be examined by the users of our package. In other words, the users of our package will not need to determine these values; they will be able to find these values using the help materials provided by the institutions that conduct the surveys and make them available to the public.

A further advantage of our Survey CC algorithm is that it will allow the assessment of the statistical significance of each canonical correlation, and hence the conclusions that might be drawn from the associated canonical coefficients or loadings, to be done separately. The reason is that a byproduct of the proposed algorithm are the p-values for each individual r_j canonical correlation (i.e., $H_0 : \rho_j = 0$), whereas current software tests the hypothesis $H_0 : \rho_j = \rho_{j+1} = \cdots = \rho_{q-1} = \rho_q = 0$. This sequential testing approach followed by the classic statistical methods has been criticized **?**. Because the effective sample size might change for each canonical correlation depending on the canonical coefficients and the CSD factors, a larger correlation magnitude might not imply a smaller p-value, therefore secondary canonical correlations might be statistically significant even if the first one is not.

This difference on the hypotheses being tested also implies that Survey CC should be considered an alternative to the existing CCA methods, not an improvement on them. Moreover, this also means that the p-values produced by Survey CC are not directly comparable to those provided by the classic statistical methods. Therefore, when it is mentioned that these statistical tests lead to different conclusions, it refers to the interpretation that a general scientist with basic training in statistics would give to the p-values, not to the formal interpretation that should be strictly given to them.

4 The R package

The package **SurveyCC** extends the functionality of `candisc::cancor()` by calculating the test statistics, degrees of freedom and p-values necessary to estimate and interpret the statistical significance of the canonical correlations according to the methods of Wilks' lambda, Pillai's trace, and Hotelling-Lawley trace (Caliński et al., 2006). The statistics for these secondary canonical correlations were, up until now, unavailable to the users of Stata's canon and SAS PROC CANCORR. In R, the repeated application of the function `p.asym()` in the **CCP** package, once for each of these classic methods, is needed as `candisc::cancor()` provides by the default only the results for the Wilks' lambda test. It is worth noting that the existing statistical software uses an F-distribution to calculate the p-values for these methods, while those provided by **SurveyCC** are based on a χ^2 distribution. This difference leads to slightly different results from those provided by the existing CCA commands. Using a different distribution makes the information provided by **SurveyCC** complementary, instead of redundant, to that provided by the existing CCA commands.

Roy's largest root test statistics are also provided by **SurveyCC**. Our implementation is based on the idea described in (Johnstone, 2009), which posits the use of the first canonical correlation as a test statistic for all secondary canonical correlations and simply changing the degrees of freedom of the F-distribution used to assess it. Other existing CCA functions provide the Roy's largest root test statistics only for the first canonical correlation. Additionally, we found in the examples included in this paper and countless preliminary experiments, that Roy's largest root p-value was smaller than the rest of those calculated by **SurveyCC**. We hypothesize that this difference might be due to the upper-bound method used by Johnstone (2009), which is the base of our programming work for this test. Although this might limit the usefulness of this statistic, we decided to include it in our calculations and leave it to the user to determine how much importance to give to the results of this method.

Moreover, **SurveyCC** implements the proposed Survey CC algorithm described in the previous section. The results of the Survey CC algorithm appears in the tables displayed by `surveycc()` as `Complex Survey CC`. As mentioned previously, the core idea of the algorithm is to calculate the correlations among the canonical variates and their corresponding statistical significance via an equivalent sequence of univariate linear regressions. This transformation allows the user to take advantage of the existing theoretical and computational resources that integrate the CSD elements into these regression models (Valliant and Dever, 2018). Hence, **SurveyCC** can include the same complex design elements as `survey::svyglm()`. Indeed, for the main function `surveycc()`, one of the main required inputs is a `survey::svydesign()` or `survey::svrepdesign()` object (see section 4.1 below). The `Statistic` for our algorithm is the coefficient for independent variable in the univariate linear regression model and, given that the variances of the canonical variates have been setup to be 1, it is equal to the corresponding canonical correlation. The other values in the results table for **SurveyCC**, i.e. `df1`, `df2`, `Chi-Sq/F` and `p-val`, are those provided by R for the coefficient estimate in simple linear regression model (for more information see the help material for `survey::svyglm()` in the **survey** package).

Beside the p-value corresponding to the Survey CC, `surveycc()` also provides statistics for the case in which only the survey weights are taken into consideration, while the rest of the complex survey design elements are ignored. This additional set of results, which are listed as `Weighted Survey CC`, allows the user to appreciate the effect that the elements, which are dispensable to obtain correct point estimates, affect the calculation of the statistical significance of the correlations.

The units and variables graphs (Gittins, 2012) can be drawn by `candisc::cancor()`, further complementing the information listed by the existing canonical correlation commands. The variables graph uses the first and second canonical coefficients for each the variables in the original data as coordinates in a 2-dimensional graph. This graph allows the user to see the associations that might exist among the variables in the original data set by studying their positions around the unit circle. Variables in opposite sides are expected to have inverse relationships, i.e., when one increases the other decreases, while variables close to each other should have positive relationships, i.e., when one increases/decreases, the other also increases/decreases. The distance from the origin can be interpreted as a direct proportional measure of the strength of these relationships. The units graph is created by multiplying the values of the original variables of each observation by the canonical coefficients of the two canonical correlations selected by the user and then using the resulting values as coordinates in a two-dimensional plane described by the corresponding canonical variates. The units graphs serves to identify possible clusters among the observations in the data set that share characteristics which might not have been apparent in the original variables (Cruz-Cano and Lee, 2014) and identify observations that might be considered outliers. For example, in a units graph, samples from grassland vegetation in a small area in North Wales formed three distinct clusters based on the community (limestone, neutral, and fertile grassland) to which they belonged (Gittins, 2012).

The **SurveyCC** package is currently available in the Comprehensive R Archive Network (CRAN): `https://cran.r-project.org/web/packages/SurveyCC/index.html`.

4.1 Syntax

The **SurveyCC** package has one main user-facing function, `surveycc()`. The arguments for this function are as follows:

- `design_object`
- `var.x, var.y`
- `howmany`
- `dim1, dim2`
- `selection`

The first item, the design object, needs to be created by using the **survey** functions `svydesign()` or `svrepdesign()`. This object will contain the actual data set, but will also keep track of all CSD characteristics, such as cluster, strata, etc. Note that, even if the data does not have a complex survey design structure, the design object still needs to be created and passed in (see "Simple Example" below).

The `var.x` and `var.y` arguments specify the names of the first and second sets of variables, respectively. These both should be character vectors, and the names should exactly match the desired variable names in the data set.

The `howmany` argument is a positive integer specifying how many canonical correlations one wants to test for. Note that this always must be between 1 and the cardinality of the largest set of variables (i.e., `var.x` or `var.y`).

The `selection` argument specifies if the effective sample size should simply be the number of observations/rows, or whether it should be based on the sum of the weights in the survey design object. Specifying `selection = "FREQ"` will result in the weight information being used, `selection = ROWS` will result in the number of rows being used. The effective sample size is used in the calculation of *p*-values and hence included in the result tables.

4.2 Output

The output of the call to `surveycc()` in R is a list with the following structure:

- `Stats.cancor`
    - `Stats.cancor.1`
    - `Stats.cancor.2`
    - `...`
- `cc_object`

The first item in this list is another list named `Stats.cancor`, and it contains a data frame for each canonical correlate (the number being obtained from the argument `howmany`). Each data frame is the table of test results, including the results of the new Survey CC method.

The second item in this list is another list named `cc_object`, which is actually the output from the call to `cancor()` in the **candisc** package. It is beyond the scope of the present article to describe everything in the `cancor()` output as it is already done extensively in its help files, but special attention should be paid to the object `cc_object$coef`, which gives the actual canonical variate coefficient estimates. This will be explored in some of the examples below.

The **SurveyCC** package contains a plotting method for the `surveycc` object. Simply pass the saved object to `plot()`, along with two arguments, `dim1` and `dim2`, which specify which variates to use as axes in the plots. The values must both be positive integers, cannot equal each other, and must be $\leq$ `howmany`. Please see the PATH example below (PATH Study Wave 1 Adult Questionnaire Example) for an example of this.

5 Examples

In this section we present four examples to illustrate different aspects of **SurveyCC** being exalted. Although all the data sets mentioned in this section are publicly available, we have also included them as supplementary files for the convenience of the readers who would like to run the code provided in this document.

These data sets are:

- **Simple Automobile data set**: This example shows that `surveycc()` can deal with data sets without CSD factors and provide information about the statistical significance of the secondary canonical correlations, which is currently difficult to obtain using exiting CCA software. In the past it has been used as an example for other canonical correlation packages (StataCorp, 2023a). The data set is available at https://www.stata-press.com/data/r17/auto.
- **2007-2010 National Health and Nutrition Examination Survey**: This case illustrates the application of **SurveyCC** to study a survey that includes a complete set of CSD factors (survey weights, cluster and strata). The data set is available at https://wwwn.cdc.gov/nchs/nhanes/Default.aspx. The creation of this data set required merging of the 2007-2008 and 2009-2010 surveys according to the specifications provided in the *National Health and Nutrition Examination Survey:Estimation Procedures, 2007–2010 Data Evaluation and Methods Research* of the CDC's National Center for Health Statistics.
- **PATH Study Wave 1 Adult Questionnaire**: This example is based on a national survey data set with replicate weights instead of cluster and strata for standard error estimation. This data set is available at https://www.icpsr.umich.edu/web/NAHDAP/studies/36231/datadocumentation. To reduce the PATH data set to a size that would not impede the installing of **SurveyCC** by persons with little memory space in their computers or unreliable or slow internet connections, we provide a subset of PATH Wave I Adult Questionnaire data set that can be downloaded from the PATH Study website. Only the rows that have no missing values for the variables used in the example, i.e., only the rows that are in fact used during the CCA calculations, were kept.
- **2021 National Youth Tobacco Survey**: The last of our national survey examples presents a case in which the `surveycc()` handles a larger set of variables and leads to conclusions different from those obtained if only the classic statistical tests are applied. The data set is available at https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/data/index.html.

We recommend to the readers who want to replicate the results listed in this manuscript running the reproducibility scripts provided in the supplementary files as they include ancillary commands, e.g., `library`, needed for the correct execution of our examples.

5.1 Simple Automobile Example

The first of our examples serves to show how `surveycc()` can handle cases in which no CSD factors exist and the user might only be interested in examining the statistical significance of the secondary canonical correlations. For this purpose, we examine the relationships among several measures related to different models of automobiles. One group of variables express the size of the automobile (length, weight, headroom, and trunk size) while the others are related to the performance of it (displacement in cubic centimeters, miles per gallon, gear ratio, and turn ratio).

Simple Automobile Example Setup

As previously mentioned, it is necessary to create the CSD object using the **survey** package, which is then passed into the main function of the current package, `surveycc()`. This is the case even if, as in the present example, there are no CSD factors in the data set (see the design_object in the code below). Notice that the only necessary information for the creation of a survey design object is `ids`, set to `~ 1` here.

In addition to the CSD object, we define vectors for each set of variables, as well as an indicator of the desired number of canonical variates for which that statistical significance must be reported. These objects are all passed into the `surveycc()` function.

```
# Simple example
design_object <-
  survey::svydesign(
    ids = ~1,
    data = auto
  )
var.x <- c("length", "weight", "headroom", "trunk")
var.y <- c("displacement", "mpg", "gear_ratio", "turn")
howmany <- 3
out.auto <- surveycc(design_object = design_object, var.x = var.x,
  var.y = var.y, howmany = howmany, selection = "FREQ")
```

Simple Automobile Example Results

Here surveycc() is used to estimate canonical correlations p-values for the Simple example.

See the output of the saved out.auto object below for the tables containing all statistics, degrees of freedom, test statistics, and p-values.

```
out.auto
```

```
#> $Stats.cancor
#> $Stats.cancor$Stats.cancor.1
#>                        Statistic df1 df2  Chi-Sq/F p-val
#> Wilks' Lambda            0.08973  16  NA 165.42306     0
#> Pillai's Trace           1.01956  15  NA  73.52406     0
#> Hotelling-Lawley Trace   8.93344  16  NA 581.68905     0
#> Roy's Greatest Root      0.89792   4  69 151.74255     0
#> Weighted Survey CC       0.94759  73  72  27.56469     0
#> Complex Survey CC        0.94759  73  72  27.56469     0
#>
#> $Stats.cancor$Stats.cancor.2
#>                        Statistic df1 df2  Chi-Sq/F   p-val
#> Wilks' Lambda            0.87907   9  NA   9.82963 0.36445
#> Pillai's Trace           0.12163   8  NA   9.58017 0.38553
#> Hotelling-Lawley Trace   0.13677   9  NA  10.08844 0.34337
#> Roy's Greatest Root      0.89792   4  69 151.74255 0.00000
#> Weighted Survey CC       0.34003  73  72   2.76224 0.00728
#> Complex Survey CC        0.34003  73  72   2.76224 0.00728
#>
#> $Stats.cancor$Stats.cancor.3
#>                        Statistic df1 df2  Chi-Sq/F   p-val
#> Wilks' Lambda            0.99399   4  NA   1.95328 0.74435
#> Pillai's Trace           0.00601   3  NA   1.95903 0.74329
#> Hotelling-Lawley Trace   0.00603   4  NA   1.94751 0.74541
#> Roy's Greatest Root      0.89792   4  69 151.74255 0.00000
#> Weighted Survey CC       0.06338  73  72   0.60644 0.54613
#> Complex Survey CC        0.06338  73  72   0.60644 0.54613
#>
#>
#> $cc_object
#>
#> Canonical correlation analysis of:
#>   4   X  variables:  length, weight, headroom, trunk
#>   with    4   Y  variables:  displacement, mpg, gear_ratio, turn
#>
#>      CanR  CanRSQ    Eigen  percent    cum                          scree
#> 1 0.94759 0.897924 8.796670 98.46902  98.47 ******************************
#> 2 0.34003 0.115619 0.130734  1.46343  99.93
#> 3 0.06338 0.004017 0.004033  0.04514  99.98
#> 4 0.04470 0.001998 0.002002  0.02241 100.00
#>
#> Test of H0: The canonical correlations in the
#> current row and all that follow are zero
#>
#>      CanR LR test stat approx F numDF  denDF Pr(> F)
#> 1 0.94759      0.08973  15.1900    16 202.27  <2e-16 ***
#> 2 0.34003      0.87907   0.9863     9 163.21  0.4534
#> 3 0.06338      0.99399   0.1026     4 136.00  0.9814
#> 4 0.04470      0.99800   0.1381     1  69.00  0.7113
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> attr(,"class")
#> [1] "surveycc" "list"
```

Notice that Stats.cancor does not include the canonical coefficients and hence it would not be useful to draw conclusions about the relationships among the sets of variables included in the CCA. To see the canonical correlations and coefficient estimates one would save the output of the function call, and then refer to the sub-object cc_object:

```
out.auto$cc_object$coef
```

```
#> $X
#>                 Xcan1        Xcan2         Xcan3         Xcan4
#> length   -0.009477879  0.144140174 -0.0329362330 -0.0211707964
#> weight   -0.001016175 -0.003663031  0.0009582644 -0.0006830296
#> headroom -0.035113185 -0.370142988 -1.5361072863  0.0440329417
#> trunk     0.002282333 -0.034273898  0.2135491362  0.3252799822
#>
#> $Y
#>                     Ycan1       Ycan2        Ycan3         Ycan4
#> displacement -0.005370441 -0.01254983 -0.019129726  0.0005364212
#> mpg           0.046148112 -0.04127765 -0.068339793 -0.2478094648
#> gear_ratio   -0.032958321  1.02797976 -3.659566938  1.0311435250
#> turn         -0.079392663  0.31126673 -0.003303379 -0.2240488108
```

Simple Automobile Example Conclusion

In this example the weighted Survey CC provides accurate results, because weights (equal to 1 in the case of simple random sampling) are the only survey design factor. For the first canonical correlation, all methods conclude that it is statistically significant and hence it is reasonable to assume that the conclusions drawn might be true, i.e., a positive relationship between the headroom of the cars (negative coefficient -0.0351) and its turn ratio (negative coefficient -0.0794) but a negative one with its MPG (positive coefficient 0.0461) are likely to exist. The second canonical coefficients represent the negative association that exist between the headroom (negative coefficient -0.3701) and the gear ratio (positive coefficient 1.0279) combined with the turn ratio (positive coefficient 0.3112).

Although the *p*-values for the Survey CC method and most of those provided by the classic tests led to the same conclusions for the first (statistically significant) and third canonical correlation (not statistically significant), different conclusions are reached for the second canonical correlation. Considering all these results together can help the user of the commands to reach a more informed judgment about the robustness of their conclusions.

5.2 2007-2010 National Health and Nutrition Examination Survey (NHANES) Example

The Survey CC method was applied to the 2007-2010 NHANES data set to evaluate the existing relationship between select demographic factors (ridageyr: Age in years, at the time of the screening interview, and indhhin2: the estimated total household income) and body measurements related to obesity (bmxbmi: BMI, bmxwaist: Waist circumference, bpxpls: Resting pulse per minute, bpxsy1: Systolic blood pressure, and bpxdi1: Diastolic blood pressure) for people between 45 and 64 years old, which is the age group with the largest proportion of persons with undiagnosed diabetes, considering the appropriate CSD factors (survey weights, cluster and strata). Although many other demographic factors such are race/ethnicity and education would be considered in a serious study about the relationship between demography and body measure, the purpose of this example is to present a relatively simple case that includes CSD factors such as cluster, strata and survey weights.

NHANES Example Setup

As previously mentioned, it is necessary to create the CSD object using the survey package, which is then passed into the main function of the current package, surveycc().

```
reducedNHANESdata <- NHANESdata %>%
  dplyr::filter(ridageyr <= 64 & ridageyr >= 45)

design_object <-
  survey::svydesign(
    id = ~sdmvpsu,
    weights = ~wtmec4yr,
    strata = ~sdmvstra,
    nest = TRUE,
    data = reducedNHANESdata
  )
var.x <- c("bmxwaist", "bmxbmi", "bpxpls", "bpxdi1", "bpxsy1")
var.y <- c("ridageyr", "indhhin2")
out.NHANES <- surveycc(design_object, var.x = var.x, var.y = var.y,
                howmany = 2, selection = "ROWS")
```

NHANES Example Results

Here surveycc() is used to estimate canonical correlations p-values for the NHANES example. Notice that the specification of the CSD elements was adapted from the section *Sample Code* webpage from the *Tutorials* tab of NHANES website https://wwwn.cdc.gov/nchs/nhanes/tutorials/samplecode.aspx.

See the output of the saved out.NHANES object below for the tables containing all statistics, degrees of freedom, test statistics, and p-values.

```
out.NHANES

#> $Stats.cancor
#> $Stats.cancor$Stats.cancor.1
#>                         Statistic  df1  df2  Chi-Sq/F p-val
#> Wilks' Lambda             0.89351   10   NA 131.69647     0
#> Pillai's Trace            0.10680    9   NA 125.23226     0
#> Hotelling-Lawley Trace    0.11884   10   NA 138.63868     0
#> Roy's Greatest Root       0.10384    5 1160  26.88222     0
#> Weighted Survey CC        0.30978 1165 1164   8.80881     0
#> Complex Survey CC         0.30978 1165   31   7.11893     0
#>
#> $Stats.cancor$Stats.cancor.2
#>                         Statistic  df1  df2 Chi-Sq/F   p-val
#> Wilks' Lambda             0.99704    4   NA  4.46159 0.34713
#> Pillai's Trace            0.00296    3   NA  4.46091 0.34721
#> Hotelling-Lawley Trace    0.00297    4   NA  4.46227 0.34704
#> Roy's Greatest Root       0.10384    4 1161 33.63175 0.00000
#> Weighted Survey CC        0.05603 1165 1164  1.95003 0.05141
#> Complex Survey CC         0.05603 1165   31  2.05636 0.04825
#>
#>
#> $cc_object
#>
#> Canonical correlation analysis of:
#>   5   X  variables:  bmxwaist, bmxbmi, bpxpls, bpxdi1, bpxsy1
#>   with    2   Y  variables:  ridageyr, indhhin2
#>
#>      CanR  CanRSQ   Eigen percent    cum                          scree
#> 1 0.30978 0.09597 0.10615  97.119  97.12 ******************************
#> 2 0.05603 0.00314 0.00315   2.881 100.00 *
#>
#> Test of H0: The canonical correlations in the
#> current row and all that follow are zero
#>
#>       CanR LR test stat approx F numDF denDF Pr(> F)
#> 1 0.309784      0.90120  12.3766    10  2318  <2e-16 ***
#> 2 0.056033      0.99686   0.9134     4  1160  0.4553
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> attr(,"class")
#> [1] "surveycc" "list"
```

NHANES Example Conclusion

Similarly to the 'Simple Automobile Example' above, we can save the output of `surveycc()` to extract the canonical coefficients themselves (one could also directly apply the `candisc::cancor()` function, as is done internally in the present package).

```
out.NHANES$cc_object$coef
```

```
#> $X
#>                 Xcan1        Xcan2
#> bmxwaist -0.070475488  0.022913764
#> bmxbmi    0.129348798 -0.171143233
#> bpxpls    0.007807571 -0.034438475
#> bpxdi1    0.054467924 -0.010895040
#> bpxsy1   -0.051782536  0.009062863
#>
#> $Y
#>                Ycan1      Ycan2
#> ridageyr -0.17278878 0.05291342
#> indhhin2  0.02498235 0.07249308
```

The canonical coefficients from CCA are shown in the above output. In the first set of canonical variables, we will focus on the coefficients with the greatest magnitudes: `bmxbmi` (first set) and `ridageyr` (second set). The results show an opposite relationship between Age (negative canonical coefficient -0.1728) and BMI (positive canonical coefficient 0.1294); that is, BMI decreases as age increases for people ages 45-64 years. This correlation is considered statistically significant regardless of the method and has been observed in previous studies such as (Yang et al., 2021), (Amies-Cull et al., 2022), (Sun et al., 2022) and (Jarrett et al., 2009). The dominant second canonical coefficients express a negative association between a combined increase in age (positive canonical coefficient 0.0529) and an increase in income (positive canonical coefficient 0.0725) and a decrease in BMI (negative canonical coefficient -0.1712). In other words, increases in income and age are associated with decreases in BMI. The negative relationship between income and BMI in adults has been seen before in the literature, especially among women (Garcia Villar and Quintana-Domeque, 2009).

Studying the results of this section, the conclusion about the statistical significance of the second canonical correlation would be different depending on whether a classic test or the Survey CC method is used for this purpose. For example, Wilks' lambda p-value is 0.3471 and Pillai's trace p-value is 0.3472, both leading to the conclusion that it is not statistically significant. In contrast, Survey CC, which accounts for the complex survey design elements, leads to the opposite conclusion that the correlation is statistically significant (p= 0.048).

5.3 PATH Study Wave 1 Adult Questionnaire Example

This example includes material about how the variables and units graphs can be used to examine the associations between certain smoking- and drinking-related variables. The evaluation of the statistical significance of the canonical correlations takes in to consideration the replicate weights needed for their accurate calculation. Once again, a serious examination between smoking and drinking alcohol would require a larger number of variables but given the purpose of this manuscript, we believe that it is acceptable to present this preliminary analysis to further illustrate how `surveycc()` can help researchers extract information from national survey data sets with replicate weights. Moreover, it is important to notice that on occasion the classic statistical tests and the proposed Survey CC method lead to the same conclusions as is the case presented in this subsection.

The smoking variables are:

- `R01_AC1022`: In past 30 days, number of days smoked cigarettes.
- `R01_AE1022`: In past 30 days, number of days used an e-cigarette.
- `R01_AG1022CG`: Number of days smoked cigarillos in past 30 days.

The variables related to alcohol drinking are:

- R01_AX0075: Number of days drank one or more drinks of an alcoholic beverage in past 30 days.
- R01_AX0076: Number of alcoholic drinks usually consumed each day on days drank in past 30 days.

PATH Study Example Setup

As previously mentioned, it is necessary to create the CSD object using the **survey** package, which is then passed into the main function of the current package, surveycc().

In addition to the CSD object, we also define here vectors for each set of variables, as well as an indicator for how many canonical variates we want. These objects are all passed into the surveycc() function.

```
design_object <-
 survey::svrepdesign(
 id = ~PERSONID,
 weights = ~R01_A_PWGT,
 repweights = "R01_A_PWGT[1-9]+",
 type = "Fay",
 rho = 0.3,
 data=reducedPATHdata,
 mse = TRUE
 )
var.x <- c("R01_AC1022", "R01_AE1022", "R01_AG1022CG")
var.y <- c("R01_AX0075", "R01_AX0076")
howmany <- 2
out.PATH <- surveycc(design_object, var.x, var.y, howmany = howmany,
  selection = "ROWS")
```

PATH Study Example Results

Here surveycc() is used to estimate canonical correlations *p*-values for the PATH Study example. Notice that the specification of the CSD elements in the function survey::svrepdesign() comes directly from the *Population Assessment of Tobacco and Health (PATH) Study [United States] Public-Use Files ICPSR Public-Use Files User Guide*. These files can be found at the PATH Study website https://www.icpsr.umich.edu/web/NAHDAP/studies/36231. See the output of the saved out.PATH object below for the output table containing all statistics, degrees of freedom, test statistics, and *p*-values.

```
out.PATH
```

```
#> $Stats.cancor
#> $Stats.cancor$Stats.cancor.1
#>                          Statistic df1 df2 Chi-Sq/F   p-val
#> Wilks' Lambda              0.93636   6  NA  9.47145 0.14875
#> Pillai's Trace             0.06395   5  NA  9.33935 0.15537
#> Hotelling-Lawley Trace     0.06763   6  NA  9.60680 0.14222
#> Roy's Greatest Root        0.05866   3 128  2.65880 0.05106
#> Weighted Survey CC         0.24187 131 130  1.15077 0.25194
#> Complex Survey CC          0.24187 131  98  1.13501 0.25914
#>
#> $Stats.cancor$Stats.cancor.2
#>                          Statistic df1 df2 Chi-Sq/F   p-val
#> Wilks' Lambda              0.99471   2  NA  1.76100 0.41458
#> Pillai's Trace             0.00529   1  NA  1.76162 0.41445
#> Hotelling-Lawley Trace     0.00532   2  NA  1.76036 0.41471
#> Roy's Greatest Root        0.05866   2 129  4.01935 0.02026
#> Weighted Survey CC         0.07375 131 130  0.79444 0.42839
#> Complex Survey CC          0.07375 131  98  0.89128 0.37496
#>
#>
#> $cc_object
#>
```

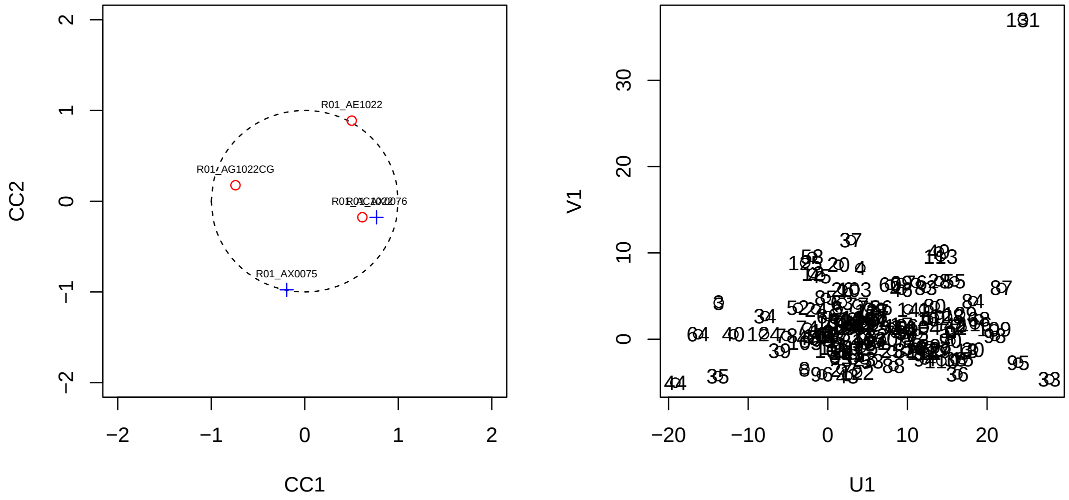

Figure 1: SurveyCC plot for the PATH example

```
#> Canonical correlation analysis of:
#>   3   X  variables:  R01_AC1022, R01_AE1022, R01_AG1022CG
#>   with    2   Y  variables:  R01_AX0075, R01_AX0076
#>
#>      CanR   CanRSQ    Eigen percent    cum                          scree
#> 1 0.24187 0.058499 0.062134  91.911  91.91 ******************************
#> 2 0.07375 0.005439 0.005469   8.089 100.00 ***
#>
#> Test of H0: The canonical correlations in the
#> current row and all that follow are zero
#>
#>       CanR LR test stat approx F numDF denDF Pr(> F)
#> 1 0.241867      0.93638   1.4145     6   254  0.2093
#> 2 0.073748      0.99456      NaN     2   NaN     NaN
#>
#> attr(,"class")
#> [1] "surveycc" "list"
```

PATH Study Example Conclusion

Notice how the p-values for all tests are statistically insignificant, hence, the same conclusions are reached regardless of the method used. This might be especially true for national survey data sets since they have sample sizes so large that p-values for all tests might be considered statistically significant.

In this example the plotting method is demonstrated, using the following code:

```
par(mfrow = c(1,2))
dim1 <- 1
dim2 <- 2
plot(out.PATH, dim1 = dim1, dim2 = dim2)
```

The variables graph in Figure 1 shows that the number of cigarillo-smoking days in past 30 days (R01_AG1022CG) and the number of alcoholic drinks usually consumed each drinking day in past 30 days (R01_AX0076) lie on opposite sites of the unit circle and hence it should be expected that when one increases the other one should be expected to decrease. On the other hand, both variables are extremely close inside a unit circle, so this association is not strong and more work would be needed to explore it. Also, Figure 1 shows that, in this example, observation 131 is a unit that has an extraordinary value in the new two-dimensional plane described by the canonical variates, and may be an outlier.

5.4 2021 National Youth Tobacco Survey (NYTS) Example

This example also shows how yet another national survey with CSD elements can be handled by the surveycc() function further demonstrating its versatility. For illustrative purposes, the proposed

Survey CC method was applied to the 2021 NYTS data set to evaluate the relationships of multiple tobacco and/or e-cigarette product use with measures of exposure to e-cigarette Internet marketing and perceived addiction to various tobacco products relative to cigarettes among Asian Americans (n=1150). In this case **SurveyCC** handles a larger set of variables, providing evidence of its potential usefulness in a real-life research scenario. The definitions of the variables in this example is provided in a supplementary file. The detailed examination of the canonical structure produced by these commands is beyond the scope of this manuscript and more appropriate for a paper focused on tobacco and nicotine research.

NYTS Example Setup

As previously mentioned, it is necessary to create the CSD object using the **survey** package, which is then passed into the main function of the current package, `surveycc()`.

In addition to the CSD object, here we define vectors for each set of variables, as well as an indicator for the desired number of canonical variates we want. These objects are all passed into the `surveycc()` function.

```
design_object <-
  survey::svydesign(
    ids = ~psu2,
    nest = TRUE,
    strata = ~v_stratum2,
    weights = ~finwgt,
    data = reducedNYTS2021data
  )
var.x <- c("qn9", "qn38", "qn40", "qn53", "qn54", "qn64", "qn69",
           "qn74","qn76", "qn78", "qn80", "qn82", "qn85", "qn88",
           "qn89")
var.y <- c("qn128", "qn129", "qn130", "qn131", "qn132", "qn134")
howmany <- 3
out.NYTS <- surveycc(design_object = design_object, var.x = var.x,
  var.y = var.y, howmany = howmany, selection = "ROWS")
```

NYTS Example Results

Here `surveycc()` is used to estimate canonical correlations p-values for the NYTS example. Notice that the specification of the CSD elements in the function `survey::svrepdesign()` were adapted from the document *Guide to weighting harmonized NYTS data for the 1999 and 2000 survey years available at Integrated Public Use Microdata Series (IPUMS) NYTS* [`https://www.ipums.org/projects/ipums-nyts`].

See the output of the saved `out.NYTS` object below for the output table containing all statistics, degrees of freedom, test statistics, and p-values:

```
out.NYTS
```

```
#> $Stats.cancor
#> $Stats.cancor$Stats.cancor.1
#>                        Statistic  df1  df2  Chi-Sq/F   p-val
#> Wilks' Lambda            0.88398   90   NA 143.58266 0.00029
#> Pillai's Trace           0.12173   89   NA 142.94303 0.00032
#> Hotelling-Lawley Trace   0.12495   90   NA 144.22768 0.00025
#> Roy's Greatest Root      0.03663   15 1134   2.87474 0.00018
#> Weighted Survey CC       0.19387 1149 1148   1.56828 0.11709
#> Complex Survey CC        0.19387 1149   75   1.51094 0.13501
#>
#> $Stats.cancor$Stats.cancor.2
#>                        Statistic  df1  df2  Chi-Sq/F   p-val
#> Wilks' Lambda            0.91759   70   NA 102.88481 0.00641
#> Pillai's Trace           0.08509   69   NA 102.56555 0.00679
#> Hotelling-Lawley Trace   0.08692   70   NA 103.20553 0.00604
#> Roy's Greatest Root      0.03663   14 1135   3.08280 0.00010
#> Weighted Survey CC       0.17412 1149 1148   1.91871 0.05527
#> Complex Survey CC        0.17412 1149   75   1.86544 0.06603
```

```
#>
#> $Stats.cancor$Stats.cancor.3
#>                        Statistic  df1  df2 Chi-Sq/F   p-val
#> Wilks' Lambda            0.94624   52   NA 68.54747 0.06171
#> Pillai's Trace           0.05482   51   NA 68.44115 0.06277
#> Hotelling-Lawley Trace   0.05570   52   NA 68.65328 0.06066
#> Roy's Greatest Root      0.03663   13 1136  3.32286 0.00005
#> Weighted Survey CC       0.14872 1149 1148  2.02639 0.04296
#> Complex Survey CC        0.14872 1149   75  2.02187 0.04676
#>
#>
#> $cc_object
#>
#> Canonical correlation analysis of:
#>  15  X  variables:  qn9, qn38, qn40, qn53, qn54, qn64, qn69, qn74, qn76, qn78, qn80, qn82, qn85, qn88, qn89
#>   with    6   Y  variables:  qn128, qn129, qn130, qn131, qn132, qn134
#>
#>      CanR   CanRSQ    Eigen percent    cum                          scree
#> 1 0.19387 0.037585 0.039053  30.904  30.90 *******************************
#> 2 0.17412 0.030317 0.031265  24.741  55.64 *************************
#> 3 0.14872 0.022118 0.022618  17.898  73.54 ******************
#> 4 0.11860 0.014066 0.014267  11.290  84.83 ***********
#> 5 0.10631 0.011301 0.011430   9.045  93.88 *********
#> 6 0.08762 0.007678 0.007737   6.123 100.00 ******
#>
#> Test of H0: The canonical correlations in the
#> current row and all that follow are zero
#>
#>       CanR LR test stat approx F numDF  denDF   Pr(> F)
#> 1 0.193869      0.88276  1.58342    90 6355.7 0.0003999 ***
#> 2 0.174117      0.91724  1.40840    70 5384.0 0.0144608 *
#> 3 0.148720      0.94591  1.21872    52 4382.5 0.1358675
#> 4 0.118602      0.96731  1.05132    36 3345.3 0.3859128
#> 5 0.106305      0.98111  0.98695    22 2266.0 0.4776264
#> 6 0.087625      0.99232  0.87743    10 1134.0 0.5539335
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> attr(,"class")
#> [1] "surveycc" "list"
```

NYTS Example Conclusion

The importance of our proposed Survey CC method is shown in all three tables produced by `surveycc`, which compare the p-values obtained using the classic tests available in the existing statistical software that consider only survey weights (weighted CCA) against those obtained using our newly developed method, Survey CC, which incorporates the remaining complex survey design elements. For the Asian-Americans in this data set, the Survey CC method results in different conclusions regarding the significance of the first canonical correlation and the relationships that can be inferred from the examination of canonical coefficients. For example, in the tables `Stats.cancor\$Stats.cancor.1` and `Stats.cancor\$Stats.cancor.2` the classic tests lead to p-values smaller then 0.05, e.g., Wilks' Lambda has p-value=0.00029, Pillai's Trace has p-value=0.00032, Hotelling-Lawley Trace has p-value=0.00025 and Roy's Greatest Root has p-value=0.00018 for the first canonical correlation, while the Survey CC method does the opposite (p-value=0.11709 taking into consideration only the survey weights and p-value=0.13501 considering all CSD factors).

6 Conclusions

The inclusion of all complex survey design elements, i.e., survey weights, cluster, strata, etc., is essential for the accurate calculation of any statistical significance, including that of canonical correlations. The incorporation of design elements is important as CCA is a multivariate method that it is especially suitable to study problems that involve sets of variables, such as those presented as examples in this manuscript. Our proposed function surveycc() allows the user not only to address this issue but to test the statistical significance of each canonical correlation separately instead of a canonical correlation and all those that come after it. Consideration of these results with those from classic statistical tests allows the user of surveycc() to obtain a complete and accurate assessment of the statistical significance of each of the canonical correlations and hence of the conclusions obtained from the examination of the canonical structure.

Moreover, an example with simple random sample data, i.e., devoid of complex survey design elements, showed that surveycc() can be useful and provide additional information compared to the existing CCA R, Stata and SAS commands, even if the user is not examining national survey data, as it provides necessary information for assessing the statistical significance of secondary canonical correlations using different methods. As seen in the examples included in this manuscript, the examination of the secondary canonical coefficients leads to insights about the relationships that exist among the variables included in the problem not provided by the first canonical coefficients. Therefore, the accurate evaluation of the statistical significance of the corresponding secondary canonical correlations is essential to determine the validity of the conclusions drawn from these analyses. Also, the ability to provide the variables and units graphs makes **SurveyCC** a complement to the existing CCA packages in R.

The results of our examples show the limited usefulness of the upper limit of the Largest Root statistic as proposed in (Johnstone, 2009), which is a limitation of the current implementation of our package. For the secondary canonical correlations, it consistently provided *p*-values significantly smaller than those of all the other methods, making its interpretation exceedingly difficult. Another limitation of our package is that, in all national survey examples, the option FREQ led to sample sizes so large that all methods provided a *p*-value of <0.0001 for all canonical correlations. These trivially significant results are due to the same issues present in (Nelson et al., 2020). Although national survey data sets include a large number of subjects, when the option FREQ is omitted, the sample sizes are not always large enough to make all statistical tests lead to a small *p*-values, especially when examining sub-populations of interest, as shown in the last example presented in Section 5 of this paper. The user of our command should keep these issues in mind and refrain from declaring a canonical correlation statistically significant based solely on these criteria. Also, this paper focuses on introducing the Survey CC algorithm and the SurveyCC package, but it does not cover any preprocessing steps that may be required for proper analysis of specific datasets, such as adjustments to survey weights when analyzing a population subdomain. Finally, to our knowledge our method Survey CC is the first to incorporate CSD factors into the calculation of the statistical significance of all canonical correlations using the exact adjustments provided in the methodological documentation of the surveys being studied, but there might be other ways to combine or modify existing statistical procedures to reach this same objective.

We presented several examples coming from a diverse set of national surveys that show the versatility of the command surveycc() and its ability to work appropriately for a diverse group of national data sets. Given the wealth of information contained in these surveys, an examination of the survey data that appropriately accounts for their complex design is essential. Our package can become an important tool in this process through the accurate examination of the canonical correlations statistical significance. In addition to the national surveys listed in this manuscript, we expect our package to be able to handle data sets such as the Behavioral Risk Factor Surveillance System, the Medical Expenditures Panel Survey, the Health Information National Trends Survey, the Tobacco Use Supplement to the Current Population Survey, and other national and international surveys produced using complex survey design.

In conclusion, our package **SurveyCC** can help to extract accurate information about the relationships between sets of variables and the statistical significance of those relationships in data sets created using complex survey design, hence having the potential to help researchers interested in using this type of data sets to address scientific questions.

References

B. Amies-Cull, J. Wolstenholme, L. Cobiac, and P. Scarborough. Estimating bmi distributions by age and sex for local authorities in england: a small area estimation study. *BMJ Open*, 2022. URL https://doi.org/10.1136/bmjopen-2022-060892. [p10]

P. Boedeker and R. Henson. *Canonical Correlation Analysis*. SAGE Publications Ltd, 2020. URL https://doi.org/10.4135/9781526421036883301. [p2]

R. Borland, N. Owen, D. Hill, and P. Schofield. Predicting attempts and sustained cessation of smoking after the introduction of workplace smoking bans. *Health Psychology*, 1991. URL https://doi.org/10.1037//0278-6133.10.5.336. [p1]

T. Caliński, M. Krzyśko, and W. WOłyński. A comparison of some tests for determining the number of nonzero canonical correlations. *Communications in Statistics - Simulation and Computation*, 2006. URL https://doi.org/10.1080/03610910600716290. [p2, 4]

D. Clark. Understanding canonical correlation analysis. *Norwich: Geo Abstracts Ltd*, 1975. URL https://www.qmrg.org.uk/catmog/index.html. [p1]

N. Correa, Y. Li, T. Adali, and V. Calhoun. Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. *IEEE Journal of Selected Topics in Signal Processing*, 2008. URL https://doi.org/10.1109/JSTSP.2008.2008265. [p1]

R. Cruz-Cano and M.-L. Lee. Fast regularized canonical correlation analysis. *Computational Statistics and Data Analysis*, 2014. URL https://doi.org/10.1016/j.csda.2013.09.020. [p5]

J. Garcia Villar and C. Quintana-Domeque. Income and body mass index in europe. *Economics and Human Biology*, 2009. URL https://doi.org/10.1016/j.ehb.2009.01.006. [p10]

R. Gittins. Canonical analysis: a review with applications in ecology. 2012. URL https://link.springer.com/book/10.1007/978-3-642-69878-1. [p1, 4, 5]

D. Hahs-Vaughn, C. McWayne, R. Bulotsky-Shearer, X. Wen, and A. Faria. Methodological considerations in using complex survey data: an applied example with the head start family and child experiences survey. *Evaluation Review*, 2011. doi: https://doi.org/10.1177/0193841X11412071. [p2]

J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson. *Multivariate data analysis*. Cengage Learning EMEA, 2019. URL https://www.cengage.com/c/multivariate-data-analysis-8e-hair-babin-anderson-black/9781473756540/. [p1]

D. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 2011. URL https://doi.org/10.1007/s10994-010-5222-7. [p1]

T. Hill and P. Lewicki. *Electronic Statistics Textbook*. StatSoft Inc., 1997. URL https://statsoft.com. [p2]

D. Holt, T. Smith, and P. Winter. Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):474–487, 1980. URL https://doi.org/10.2307/2982065. [p2]

H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936. URL https://doi.org/10.1093/biomet/28.3-4.321. [p1]

B. Jarrett, G. Bloch, D. Bennett, B. Bleazard, and D. Hedges. The influence of body mass index, age and gender on current illness: a cross-sectional study. *International Journal of Obesity*, 2009. URL https://doi.org/10.1038/ijo.2009.258. [p10]

I. M. Johnstone. Approximate null distribution of the largest root in multivariate analysis. *The annals of applied statistics*, 3(4):1616, 2009. URL https://doi.org/10.1214/08-AOAS220. [p2, 4, 15]

S. Killip, Z. Mahfoud, and K. Pearce. What is an intracluster correlation coefficient? crucial concepts for primary care researchers. *The Annals of Family Medicine*, 2(3):204–208, 2004. URL https://doi.org/10.1370/afm.141. [p2]

J. D. Leo and E. Wulfert. Problematic internet use and other risky behaviors in college students: An application of problem-behavior theory. *Psychol Addict Behav.*, 2013. URL https://doi.org/10.1037/a0030823. [p1]

T. H. Lewis. *Complex survey data analysis with SAS*. Chapman and Hall/CRC, 2016. URL https://doi.org/10.1201/9781315366906. [p2]

T. Lumley. Analysis of complex survey samples. *Journal of statistical software*, 9:1–19, 2004. URL https://doi.org/10.18637/jss.v009.i08. [p2]

W. Meredith. Canonical correlations with fallible data. *Psychometrika*, 1964. URL https://doi.org/10.1007/BF02289567. [p2]

D. Morris, A. Davis, K. Lauritsen, C. Rieth, M. Silvestri, J. Winters, and S. CHermack. Substance use consequences, mental health problems, and readiness to change among veterans seeking substance use treatment. *Journal of Substance Abuse Treatment*, 2018. URL https://doi.org/10.1016/j.jsat.2018.08.005. [p1]

D. R. Nelson, S. H. Wong-Jacobson, E. Lilly, and Company. %surveycorrcov macro: Complex survey data correlations for multivariate analysis and model building. 2020. URL https://api.semanticscholar.org/CorpusID:215747405. [p3, 15]

SAS. *Survey Weights: A Step-by-Step Guide to Calculation*. SAS Institute Inc., 2012. URL https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_cancorr_toc.htm. [p2]

StataCorp. *Stata Statistical Software: Release 18*. Stata Press, 2023a. URL https://www.stata-press.com/manuals/documentation-set. [p6]

StataCorp. *Stata 18 Multivariate Statistics Reference Manual*. Stata Press, 2023b. URL https://www.stata.com/manuals/mvcanon.pdf. [p2]

J. Stowe, C. Watson, and T. Robertson. Relationships between the two sides of the balance sheet: A canonical correlation analysis. *Journal of Finance*, 1980. URL https://doi.org/10.2307/2327214. [p2]

J.-Y. Sun, W. Huang, Y. Hua, Q. Qu, C. Cheng, H. Liu, X. Kong, Y. Ma, and W. Sun. Trends in general and abdominal obesity in us adults: Evidence from the national health and nutrition examination survey (2001–2018). *Frontiers in Public Health*, 2022. URL https://doi.org/10.3389/fpubh.2022.925293. [p10]

F. Travis and Y. Lagrosen. Creativity and brain-functioning in product development engineers: A canonical correlation analysis. *Creativity Research Journal*, 2014. URL https://doi.org/10.1080/10400419.2014.901096. [p2]

R. Valliant and J. Dever. *Survey Weights: A Step-by-Step Guide to Calculation*. Stata Press, 2018. URL https://www.stata-press.com/books/survey-weights. [p4]

C. Winship and L. Radbill. Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2):230–257, 1994. URL https://doi.org/10.1177/0049124194023002004. [p2]

Y. Yang, C. Walsh, M. Johnson, D. Belsky, M. Reason, P. Curran, A. Aiello, M. Chanti-Ketterl, and K. Harris. Life-course trajectories of body mass index from adolescence to old age: Racial and educational disparities. *Proc Natl Acad Sci*, 2021. doi: 10.1073/pnas.2020167118. [p10]

X. Zhuang and Z. Yang. A technical review of canonical correlation analysis for neuroscience applications. *Hum Brain Mapp*, 2020. URL https://doi.org/10.1002/hbm.25090. [p1]

Raul Cruz-Cano
Department of Epidemiology and Biostatistics, IU School of Public Health
Indiana University School of Public Health
Bloomington, IN
USA
raulcruz@iu.edu

Aaron Cohen
Department of Epidemiology and Biostatistics, IU School of Public Health
Biostatistics Consulting Center
Department of Epidemiology and Biostatistics
Indiana University School of Public Health
Bloomington, IN
USA
cohenaa@iu.edu

Erin Mead-Morse
Department of Public Health Sciences, UConn Health, School of Medicine
UConn Health, School of Medicine
Farmington, CT
USA
mead@uchc.edu