Multivariate understanding of income and expenditure in United States households with statistical learning

Mingzhao Hu

Received: 10 September 2018 / Accepted: 13 June 2022 / Published online: 23 July 2022 © The Author(s) 2022

Abstract
In recent decades, data-driven approaches have been developed to analyze demographic and economic surveys on a large scale. Despite advances in multivariate techniques and learning methods, in practice the analysis and interpretations are often focused on a small portion of available data and limited to a single perspective. This paper aims to utilize a selected array of multivariate statistical learning methods in the analysis of income and expenditure patterns of households in the United States using the Public-Use Microdata from the Bureau of Labor Statistics Consumer Expenditure Survey (CE). The objective is to propose an effective data pipeline that provides visualizations and comprehensive interpretations for applications in governmental regulations and economic research, using thirty-five original survey variables covering the categories of demographics, income and expenditure. Details on feature extraction not only showcase CE as a unique publicly-shared big data resource with high potential for in-depth analysis, but also assist interested researchers with pre-processing. Challenges from missing values and categorical variables are treated in the exploratory analysis, while statistical learning methods are comprehensively employed to address multiple economic perspectives. Principal component analysis suggests that after-tax income, wage/salary income, and the quarterly expenditure in food, housing and overall as the five most important of the selected variables, while cluster analysis identifies and visualizes the implicit structure between variables. Based on this, canonical correlation analysis reveals high correlation between two selected groups of variables, one of income and the other of expenditure.

Keywords Principal component analysis · Cluster analysis · Canonical correlation analysis · Interpretive modeling · Data visualizations

Mingzhao Hu mingzhaohu@ucsb.edu

1 Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA
Introduction

Statistics has become an integral part in the modern economics research. In everyday life, the immediate economic concern for the average person is to have a balanced income and expenditure. Statistical analysis provides quantitative interpretations of the situation for making a rational decision. In particular, multivariate statistical learning methods are relied on to build data pipelines that can reveal the connections and relative significance between economic factors. On the other hand, in this digital age the datasets are increasing rapidly in complexity and scale. Methods in multivariate statistical learning offers direct and elegant solutions, with developments on existing methods in recent years addressing the new challenges in data such as missing values and computational efficiency. In this study, an array of multivariate statistical learning methods including principal component analysis (PCA), cluster analysis, and canonical correlation analysis (CCA) are selected to give rise to economic insights from different perspectives in a combined analysis on the Consumer Expenditure (CE) Survey dataset. We will consider the details on the procedure of data cleaning and feature extraction, as well as the discussion of comparative statistics and cross referencing the method performances. For the purpose of illustration, we will introduce some economic concepts and issues of interest, with quantitative and visualized support for interpretations. Furthermore, we emphasize the interactions and comparisons between the variables using multivariate analysis techniques.

The Bureau of Labor Statistics (BLS) under the United States Department of Labor has been tasked with surveying, computing and reporting how Americans spend their money through the Consumer Expenditure Survey (CE). The Consumer Expenditure Survey is unique in the sense that it is the only nationwide survey conducted by the U.S. government that investigates the income, expenditure and characteristics of the consumers. BLS has been releasing their CE datasets in the Public-Use Microdata (PUMD), which includes the Interview files based on quarterly interviews, where during an interview a Consumer Unit (CU) is asked to report the expenditure for a reference period of three months. The Interview files are divided into five categories, each with a different focus and are updated every quarter. For this analysis, data from one quarter will be chosen as the example dataset from the most recently published data from 2016 at the time of this analysis. To showcase and compare different multivariate statistical learning methods based on findings from the CE survey, the collection fmli161, the first category of Interview files covering 2016’s first quarter, is arbitrarily chosen since all four quarters in 2016 are equally suitable for consideration. The fmli161 file contains CU characteristics, income, and the summary level expenditure. The source of the original data is the website of Bureau of Labor Statistics (2016), and is open to the entire public for free download. The data also served as the core dataset for Data Challenge Expo 2017. An overview of additional related work is summarized in the editorial by Garner and Martinez (2022).

This paper develops a data pipeline that provides visualizations and interpretations for general applications on extensive social survey datasets in governmental
regulations and economics research, based on multivariate machine learning methods that address concerns over raw survey data which includes missingness issues. With a strong emphasis on details in data cleaning and feature extraction from the *fmli161* dataset analyzed in this study, we provide a general sense of its data structure as well as first-hand experience in data wrangling to help readers who are interested in research on this public data source. The dataset is of high potential for future analysis motivated by a wide range of topics. The proposed data pipeline is carefully constructed to provide comparisons and cross references between multivariate statistical learning methods in exploring the selected variables.

From the original *fmli161* survey dataset, we subjectively select thirty-five features. These are considered as the most representative and quantitative in the three categories of *Characteristics*, *Income* and *Expenditure*. PCA is adopted to understand the data structure using visual presentations. Cluster analysis is implemented to compare clusters and provide intuition on the grouping of variables, then CCA reveals the relationships between one selected group of income variables and another group of expenditure variables. For a more detailed discussion on the application of these techniques, one may refer to Johnson and Wichern (2007). R (R Core Team 2021) is used for implementations in this study, while the entire R code and data are provided for full reproducibility on the github repository, https://github.com/asa-stat-computing-and-graphics/COST-DataExpo-2017/tree/master/IncomeExpenditureU.SHouseholdsMultivariateAnalysis.

The outline of the following sections is as follows. Section 2 presents preliminary analysis based on data cleaning and feature extraction, where subsections respectively provide introduction for features of interest, discuss the important issue of missing values and examine the correlation between variables in detail. Section 3 applies PCA and compares the performance between regular and sparse PCA. Based on these findings, Sect. 4 evaluates the grouping of variables via a combined cluster analysis with three clustering criteria. Section 5 analyzes results from CCA to reveal relationships between the clusters of variables. Section 6 discusses findings from the proposed data pipeline. The detailed explanations for all selected variables in Sect. 2.1, the full correlation matrix in Sect. 2.3, and the principal components from Sect. 3 are deferred to the appendix.

## 2 Preliminary analysis

### 2.1 Feature extraction

A total of thirty-five variables were selected from the original *fmli161* dataset, which are classified into three different categories. The 807 variables in the *fmli161* files fall in three major categories. Apart from the features that are recording different aspects of income and expenditure classified as *Income* and *Expenditure* respectively, the rest are all classified as *Characteristic*, since they reflect some traits of the surveyed households and reveals information on the demographics. Not all 807 variables are necessary for answering these questions. In practice, choosing only
relevant variables is a useful strategy in terms of avoiding missing values, lowering analysis difficulties, and reducing cost. The corresponding detailed descriptions for these selected features may be found in Appendix A. Within the selected dataset, there are 6 categorical variables and 29 continuous variables over 6,426 Consumer Units, but includes a considerable proportion of missing values read in as “NA”. As a common challenge in extensive social surveys, the missing value problem in the CE survey requires good statistical reasoning to be accommodated in the analysis, as is discussed in Sect. 2.2.

The four categorical variables are urban location (BLS_URBN), housing tenure (CUTENURE), education level (EDUC_REF), race (REF_RACE), sex (SEX_REF), and occupation (OCCUCOD1). These are categorized in the survey based on natural intuitions. Moreover, some multivariate statistical learning methods, such as PCA, require continuous variables in input datasets. Therefore, these categorical features are converted in the final step of preliminary data processing.

The criterion for feature extraction from the 807 variables is based on the natural grouping by category. For each subcategory within a category, such as the subcategory of income due to rent, only one variable is selected and the preference is for the BLS derived variables. This is the common setting for most of the information included in the PUMD data files. This decision to take the BLS derived variables is supported by the fact that BLS constructed the original data with solid justifications, using multiple imputations and records from previous years in the estimation of missing components. The derivation also includes introducing anonymity to protect respondent identity. Although 773 variables were sacrificed, the extracted features offers a comprehensive coverage of the overall survey while minimizing information overlap and data complexity. To demonstrate this, based on the extracted features from the CE survey data, we may consider questions of interest as intuitive examples that are economically informative, addressing common concerns on income and expenditure.

A combination of a handful of extracted features covering different categories is sufficient for providing answers to these example questions. To begin with, we may consider if people with higher education necessarily have higher income. Education level is classified as a Characteristics variable, and for total income the most appropriate variable is annual income after tax (FINCBTXM), an Income variable. The second example question is whether male respondents earn more than females and whether they spend more. To understand what the significance of gender is, apart from the Characteristics variable for sex, the earnings and spendings are sufficiently reflected by the Income variable for the annual income after tax, and the Expenditure variable for the quarterly total expenditure (TOTEXPCQ). The last example considers if people who spend more eating out necessarily have higher overall expenditure. In this case, the comparison is between the Income variable for annual income after tax and the Expenditure variable on the quarterly food expenditure (FOODCQ). One would naturally expect that the wealthier portion of the population would have higher spendings on dining, but an opposite argument can be made for the case of overworked employees who are forced to eat out regularly. To further argue for or against the opposition it would require considering the Income variable for the annual income from salary and wages (FSALARYM) and the ratio of the following
two *Expenditure* variables, the quarterly food expenditure on eating out (FDA-WAYCQ) and the quarterly food expenditure on eating at home (FDHOMECQ).

Therefore, keeping only the most representative feature in each category will be sufficient to answer a wide range of questions. This partly explains why the thirty-five selected features are suitable for the following analysis when originally 807 were given. This report will investigate the relationship between features in different categories and explore the interactions. NEWID is the row names of the dataset, and each CU is uniquely identified by NEWID. There are 14 features in the characteristics category, 10 features in the income category, and eleven features in the expenditure. The dataset produced by the thirty-five selected features offers an inclusive coverage of key characteristics, and a sufficiently large sample size that allows for further reduction and grouping procedures. This dataset is renamed as *fmli* and provides a good basis for the treatment of missing values.

### 2.2 Missing values

The next challenge is missing values, a major issue in dealing with the CE dataset, and a common obstacle in large real-life social surveys. Without further assumptions or treatment, missing values become an impediment to applying many multivariate statistical learning techniques and reaching valid interpretive results. If a sample of monthly income has missing values, then the average income cannot be directly computed based on this sample without further assumptions or addressing the missing values first.

As a special step in navigating the CE survey data, before treating the missing values in the *fmli* dataset we first consider the dummy variables in the original *fmli161* dataset that correspond to the extracted features. For example, for the count of CUs in household (HH_CU_Q), the corresponding dummy variable, known as a flag variable, is labeled similarly as HH_CU_Q_. These flag variables are for the purpose of recording the state of the data values and are labeled differently as shown in Table 1. In the *fmli161* dataset, all “.” or “NA” values in the dataset denote missing values where the respondent did not provide an answer. However, for many reported expenditure variables, the input values are “0”. Since the corresponding flag variables do not exist for these expenditure, such as the quarterly food expenditure, the “0” means the respondent gave an actual response of zero dollars.

| Table 1 Definitions of missing flag variables |
|-----------------------------------------------|
| Flag variable name | Definition |
|-------------------|------------|
| A                 | Valid non-response: a response is not anticipated |
| C                 | “don’t know”, refusal or any other type of non-response |
| D                 | Valid data value |
| T                 | Topcoding applied to value |


in expenditure. This is not the same as giving a blank answer and being recorded as “. ” or “NA”. Occasionally, CUs report “0” for several important expenditure, such as the quarterly total expenditure for the entire first quarter of 2016, which cannot happen in real life. Therefore, in these cases we consider “0” as missing. But for detailed expenditure such as the quarterly food expenditure on eating out, reporting zero spending is not considered as missing. Therefore, missing values refers to those entries in \textit{fmli161} with “. ” or “NA” values, and the special case of “0” for some expenditure.

In the first step of dealing with missing values, the focus is on the CUs reporting “0” for some expenditure. Based on the \textit{fmli} dataset, one third of all the quarterly total expenditure are zero and considered as missing. To be specific, out of the 6,426 observations, 2,171 have reported zero in the quarterly total expenditure. The dataset is first split according to whether the quarterly total expenditure has a missing value of “0”. Records of these are removed, resulting in the new \textit{fmlifull} dataset that contains 4,254 observations. No apparent pattern exists for the missing values, and no reasonable explanation holds given the current background information. Therefore, it is assumed that the values are missing completely at random. This assumption lends simplicity to the analysis and supports list-wise deletion of the rows with missing values. While no immediate underlying motive exists for reporting zero expenditure, this process increases risk of overlooking some underlying causes and concerns in exchange for the simplicity and clarity of the dataset.

The second step is to consider the number of missing values in the other detailed expenditure variables in the form of “0”, once it is ensured no missing value in the form of “0” exists in the quarterly total expenditure. As a general rule of thumb, if missing values account for less than 10% of the total information in the variable it is considered acceptable. There is no deletion in this step. The extracted features on detailed expenditure do not have corresponding flag variables, yet contain a lot of zero entries. Since the dataset has been cleaned to include only the positive quarterly total expenditure, these zeros can be seen as “valid data”. For example, for a “0” entry in the quarterly healthcare expenditure (HEALTHCQ), it does not mean the respondents intentionally ignored answering the survey question on how much was spent on healthcare during the past three months, because they have accurately reported the total expenditure. Similar conclusions hold for the quarterly transportation expenditure (TRANSCQ) and the entertainment expenditure (ENTERTCQ).

One possible explanation, especially for variables of the quarterly education expenditure (EDUCACQ), the quarterly reading expenditure (READCQ) and the quarterly apparel and services expenditure (APPARCQ), is that there is actually no spending in these aspects for some households during the quarter. These are not considered as basic needs and within a period of only three months, it is reasonable for some consumer units to have none of such expenses at all. In the following analysis, we consider them as spending zero dollars on each expense. However, this argument is harder to make for variables of unavoidable expenditure like quarterly the housing expenditure (HOUSCQ). A second potential explanation for responding “0” would be that the amount spent is negligible. Therefore, these zeros would affect conclusions and should be considered also as spending zero dollars on each expense. While trying to provide a quantitative justification for this argument, it is necessary
check the percentages of zeros and decide the measure. The percentages are listed in Table 2.

In the survey 80.22% of reading and 84.06% of educational expenditure are reported as zero. This is an interesting and somewhat concerning discovery, since it reveals that most of the interviewed CUs in the United States did not invest enough in education. Eating out, healthcare and entertainment expenditure still have considerable zeros even after discarding the 2,171 consumer units where the quarterly total expenditure is zero. This preliminary probe into the CE survey offers potential for providing quantified answers to many economic studies. Given the income levels, the effect of income status on spending emphasis can be verified. A more elaborate plan would be to investigate the cross relationship between educational level, income level, and expenditure on reading and education using this data from the CE survey to determine if more educated people earn more and invest more in education. 47.92% of households reported zero expenditure on apparel. This is in line with the lack of expenditure on other recreational expenses. All eleven expenditure features have been treated for missing values.

In the third step, missing values labeled as “NA” in Characteristics and Income categories for the remaining 4,254 observations are treated. Not only is there no “NA” recorded for any of the expenditure variables, in the characteristics variables, no missing value is observed, since it is unlikely that the respondents would withhold basic information such as gender. Missing values in the income variables are reflected in Table 3, which gives the percentage of “NA”s in all income variables in which any “NA” is found. For working hours per week (INC_HRS1) and the primary type of occupation (OCCUCOD1), both have 1,453 “NA”s, resulting in a high percentage of 34.15% being missing. Thus these two variables are dropped since they are not critical to defining the income situation. This is a regular approach in reducing the complexity of the dataset and avoiding missing values, but the robustness of results is sacrificed due to fewer dimensions and explanatory variables.

Four other income variables, including interest or dividends income (INTRD-VXM), retirement income (RETSURVM), rental income (NETRENTM) and

| Table 2  | Missing percentages for “0” in expenditure features |
|-----------------|-----------------------------------------------|
| Expenditure variable | Percentage of missing (%) |
| Quarterly food expenditure | 0.61 |
| Quarterly food expenditure on eating at home | 1.01 |
| Quarterly food expenditure on eating out | 19.42 |
| Quarterly housing expenditure | 0.50 |
| Quarterly apparel and services expenditure | 47.92 |
| Quarterly transportation expenditure | 5.92 |
| Quarterly healthcare expenditure | 20.81 |
| Quarterly entertainment expenditure | 14.34 |
| Quarterly education expenditure | 84.06 |
| Quarterly reading expenditure | 80.22 |
welfare income (WELFAREM), each have 3,519, 3,645, 4,048, and 4,209 entries that recorded “NA”. While multiple imputation is one possible approach, the original dataset has many features that are not of immediate interest, such that even after extraction the dataset retains a large number of variables. Some are not informative on the general population, such as retirement income, while others overlap in the information they provide, such as retirement income and welfare income. On the other hand, while multiple imputation is popular for treating missing values, it has its drawbacks, especially when dealing with comprehensive surveys such as the CE survey.

More extensive and rigorous arguments and findings discussing the different approaches in managing missing values on real economic surveys datasets, in particularly the CE survey, are available (Loh et al. 2019). However, for efficiency of the proposed data pipeline, the most straightforward approach is preferred. The resulting dataset is relabeled as fmlifull, and contains fourteen characteristics variables, four income variables, eleven expenditure variables, and 4,254 observations. The third step deleted Income variables that contained a high percentage of “NA”s and kept all observations from the previous step. Using function md.pattern from the R package mice (van Buuren and Groothuis-Oudshoorn 2011) on the resulting dataset fmlifull has confirmed there are no more missing values.

### 2.3 Correlation matrix

A correlation matrix reveals the relationship between pairs of features by giving a direct illustration of the degree of association. It is also necessary for computations in Sect. 3, when finding significant linear combinations of features, and Sect. 5, which investigates the linear combination within groups of income and expenditure variables. However, as the number of features increases, the dimension of the matrix increases, making it more challenging to interpret the correlations. In dealing with a correlation matrix, it is important to remember that it only reflects association, not a causal relationship. In this case, with 29 features, the dimension is still acceptable.

However, there are four categorical variables among the 29 features, including urban location, education level, race and sex. In order to incorporate these features in multivariate statistical learning methods that requires continuous features such as PCA, it is necessary to overcome the difference between categorical

| Income variable                  | Percentage of missing (%) |
|----------------------------------|---------------------------|
| Working hours per week           | 34.15                     |
| Primary type of occupation       | 34.15                     |
| Interest or dividends income     | 82.72                     |
| Retirement income                | 85.68                     |
| Rental income                    | 95.16                     |
| Welfare income                   | 98.94                     |
variables and continuous variables. Therefore, for each categorical variable, the levels are expanded into as many new columns of binary variables correspondingly. The new columns consist solely of zeroes and ones. In total, four categorical variables has been replaced by the resulting 18 binary features in the dataset *fmlifull*. Then the correlation matrix covers all 42 variables. Given its size, it is important to focus on portions and draw conclusions in general.

The correlation matrix is given in Appendix B. Most of the correlations are below 0.5, and out of the 903 pairs of different features, only 79 pairs have correlation coefficient larger than 0.5. A selected portion of the correlation matrix including three characteristics features, two income features and three expenditure features is given in Table 4. These features are chosen not only because they are important for describing the characteristics, income and expenditure of the CU, but also because some of their correlations are notable for interpretation. For example, the correlation between the annual income from salary and wages and the annual income after tax is 0.91. This suggests that correlation between pairs of variables affiliated with each other tends to be high. A correlation of 0.67 between housing and total expenditure also reflects this trend. Furthermore, between variables of possible causal relationship, such as the before tax income and the total expenditure, there is also high correlation. The correlation matrix provides useful preliminary results, but due to the number of variables it is not an efficient representation.

| Table 4 Selected portion of correlation matrix |
|-----------------------------------------------|
| AS_COMP1 | AS_COMP2 | FAM_SIZE | FINCBTXM |
| AS_COMP1  | 1.00    | −0.02    | 0.51     | 0.28     |
| AS_COMP2  | −0.02   | 1.00     | 0.55     | 0.16     |
| FAM_SIZE  | 0.51    | 0.55     | 1.00     | 0.28     |
| FINCBTXM  | 0.28    | 0.16     | 0.28     | 1.00     |
| FSALARYM  | 0.27    | 0.16     | 0.31     | 0.91     |
| TOTEXPCQ  | 0.19    | 0.11     | 0.20     | 0.55     |
| FOODCQ    | 0.26    | 0.18     | 0.33     | 0.43     |
| HOUSCQ    | 0.14    | 0.09     | 0.19     | 0.50     |
| FSALARYM  | 0.27    | 0.19     | 0.26     | 0.14     |
| TOTEXPCQ  | 0.16    | 0.11     | 0.18     | 0.09     |
| FOODCQ    | 0.31    | 0.20     | 0.33     | 0.19     |
| HOUSCQ    | 0.91    | 0.55     | 0.43     | 0.50     |
| FSALARYM  | 1.00    | 0.50     | 0.40     | 0.48     |
| TOTEXPCQ  | 0.50    | 1.00     | 0.57     | 0.67     |
| FOODCQ    | 0.40    | 0.57     | 1.00     | 0.51     |
3 Principal component analysis

The quarterly interview of the CE survey was conducted with three sections, which include respondent characteristics, income summary and expenditure summary. This is reflected by the extracted features to compose the example `fmlifull` dataset. As the first multivariate statistical learning method in the proposed data pipeline, PCA investigates the significance of variables in explaining the total variation in the dataset. Grouping of the dataset by a single variable will also be demonstrated. This section lays the foundation for the cluster analysis in Sect. 4, and provides abundant visualizations of the dataset to assist interpretations. PCA also helps with standardizing the data, decreasing dimensions and highlighting the more important variables. Given the 42 features, selecting the more important ones and attempting dimension reduction form a main component of the proposed data pipeline. This is implemented in the following multivariate statistical learning process in addition to previous endeavors on feature extraction and dealing with missing values.

PCA originates from the foundational work done by Pearson (1901). Anderson (1963) provides a more rigorous discussion of the theoretical properties of PCA. Imposed on the data after list-wise deleting all the missing values, PCA aims to explain the variance-covariance structure of a set of variables through uncorrelated linear combinations known as principal components (PC). The first PC is the linear combination such that its variance is the largest, and the second PC has the second largest variance. As linear combinations of the variables, PCs can be used to summarize the variation of the sample, and be interpreted by considering the coefficients of each variable in this linear combination known as loadings.

There are multiple ways to apply PCA in R, among which two popular choices are the functions `princomp` and `prcomp` in the package `stats` (R Core Team 2021). Both compute the principal components but give different eigenvalues. While `princomp` uses the correlation matrix and performs spectral decomposition, `prcomp` uses the data matrix and applies singular value decomposition, which allows for better numerical accuracy. Although both functions are acceptable, the generic function `prcomp` is chosen and is performed on a scaled data matrix. Scaling prevents the magnitude of certain variables from dominating the associations between other sample variables. Using the summary of outputs, 42 principal components are produced, each is a linear combination of the 42 variables, and each explains a fraction of the total variance. The entire output is included in Appendix C. Table 5 presents the first five PCs.

Interpretation of the principal components and their correlation has been discussed in papers such as Cadima and Jolliffe (1995). However, there are new challenges when dealing with more variables. For example, the best interpretation for PC1 is as a weighted linear combination of almost all 42 features with a focus on the income, salary, total expenditure and food expenditure. With loadings of −0.06 and −0.07, details in educational expenses and reading expenses were considered less important for PC1. PC2 focused on the demographics, mainly the number of people over 64 and the age of the CU. Intuition suggests that with each
Table 5  First five principal components

|                          | PC1  | PC2  | PC3  | PC4  | PC5  |
|--------------------------|------|------|------|------|------|
| HH_CU_Q                  | 0.04 | −0.07| 0.03 | −0.09| 0.16 |
| AS_COMP1                 | −0.18| 0.03 | −0.31| 0.17 | −0.25|
| AS_COMP2                 | −0.11| −0.13| 0.30 | 0.30 | −0.15|
| CUTENURE                 | 0.15 | −0.20| 0.00 | −0.14| 0.07 |
| FAM_SIZE                 | −0.22| −0.21| 0.00 | 0.38 | −0.24|
| NO_EARNR                 | −0.25| −0.23| −0.12| 0.13 | −0.06|
| PERSLT18                 | −0.15| −0.26| 0.02 | 0.28 | −0.12|
| PERSOT64                 | 0.12 | 0.39 | 0.17 | 0.13 | −0.24|
| VEHQ                     | −0.17| 0.11 | −0.09| 0.18 | −0.05|
| AGE_REF                  | 0.11 | 0.40 | 0.15 | 0.08 | −0.19|
| FINCBTXM                 | −0.32| 0.06 | −0.01| −0.07| 0.06 |
| FSALARYM                 | −0.31| −0.05| −0.06| −0.07| 0.08 |
| FRRETIRM                 | 0.12 | 0.39 | 0.16 | 0.13 | −0.22|
| FSMFPFRXM                | −0.11| 0.04 | 0.02 | −0.04| 0.08 |
| TOTEXPCQ                 | −0.31| 0.16 | 0.11 | −0.10| 0.12 |
| FOODCQ                   | −0.31| 0.10 | 0.11 | −0.01| −0.04|
| FDHOMEQCQ                | −0.25| 0.05 | 0.10 | 0.10 | −0.10|
| FDAWAYCQ                 | −0.25| 0.11 | 0.07 | −0.12| 0.05 |
| HOUSCQ                   | −0.27| 0.09 | 0.10 | −0.11| 0.08 |
| APPARCQ                  | −0.12| 0.02 | 0.08 | −0.08| 0.10 |
| TRANSCQ                  | −0.12| 0.10 | 0.07 | −0.01| 0.06 |
| HEALTHCQ                 | −0.13| 0.21 | 0.08 | 0.00 | −0.01|
| ENTERTCQ                 | −0.15| 0.12 | 0.08 | −0.07| 0.08 |
| EDUCACQ                  | −0.06| 0.00 | 0.01 | −0.09| 0.12 |
| READCQ                   | −0.07| 0.12 | 0.06 | −0.06| 0.08 |
| BLS_URBN_URBAN           | −0.06| −0.08| 0.11 | −0.36| −0.40|
| BLS_URBN_RURAL           | 0.06 | 0.08 | −0.11| 0.36 | 0.40 |
| EDUC_REF_NEVER           | 0.01 | −0.01| 0.01 | 0.01 | −0.03|
| EDUC_REF_FIRST           | 0.02 | −0.03| 0.00 | 0.12 | −0.05|
| EDUC_REF_SECOND          | 0.07 | −0.04| 0.02 | 0.04 | −0.11|
| EDUC_REF_HIGH            | 0.07 | 0.03 | −0.05| 0.16 | −0.06|
| EDUC_REF_ASSC            | 0.00 | −0.04| 0.02 | 0.05 | 0.00 |
| EDUC_REF_UDGRD           | −0.09| 0.00 | −0.01| −0.15| 0.10 |
| EDUC_REF_GRD             | −0.10| 0.09 | 0.05 | −0.12| 0.07 |
| REF_RACE_WH              | −0.01| 0.20 | −0.15| 0.22 | 0.29 |
| REF_RACE_BLK             | 0.06 | −0.17| 0.12 | −0.17| −0.22|
| REF_RACE_NA              | −0.01| −0.02| 0.00 | 0.00 | −0.04|
| REF_RACE_AS              | −0.06| −0.08| 0.07 | −0.12| −0.13|
| REF_RACE_PI              | −0.01| −0.02| 0.00 | −0.02| −0.04|
| REF_RACE_MULTI           | −0.02| −0.04| 0.03 | −0.04| −0.07|
| SEX_REF_MALE             | −0.05| 0.17 | −0.53| −0.12| −0.18|
| SEX_REF_FEMALE           | 0.05 | −0.17| 0.53 | 0.12 | 0.18 |
PC as a linear combination of all 42 features, the coefficients can be interpreted such that one PC might be representing the contrast between income and expenditure as two groups, another representing contrast between income after tax and retirement income, etc. However, since PC1 has nonzero loadings for 42 features and PC2 for 40 features, direct interpretations of PCs cannot offer clear insights into the grouping or relative significance of the features. Furthermore, as shown in Table 6, the first nine PCs together explains about 51% of the total variation in the dataset. In fact, it takes 19 PCs to explain 76% of the total variation, and 34 PCs to explain 99%.

Table 6 Importance of first nine principal components

|       | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SD    | 2.46| 1.76| 1.57| 1.53| 1.37| 1.34| 1.21| 1.16| 1.10|
| Proportion of variance | 0.14 | 0.07 | 0.06 | 0.06 | 0.04 | 0.04 | 0.03 | 0.03 | 0.03 |
| Cumulative proportion | 0.14 | 0.22 | 0.28 | 0.33 | 0.38 | 0.42 | 0.45 | 0.49 | 0.51 |

Figure 1 presents a scree plot for the first ten principal components. The scree plot is a direct visualization of the importance of PCs by the amount of variation each PC explains. In general, one should look for the sharpest convex turn, or “elbow”, to determine the best amount of principal components that could be considered as important. Since the variation explained by PCs is always decreasing, such a turn indicates that the following PCs explain much less variation compared to PCs before the turn. In this case, the most prominent “elbow” can be observed after PC3, while

![Scree plot for first ten principal components](image_url)
Fig. 2  Variables in first and second principal components
the second is after PC5. With three PCs, 28% of the total variation is explained, which is not sufficient for showing the majority of the variance. For 5 PCs, similarly only 38% of the total variation is explained. Furthermore, the percentage of variance explained by each PC is about the same from PC2 to PC6, which shows that the traditional PCA is not ideal in showing relative significance among variables for this particular CE dataset. The results have demonstrated the advantages of PCA, which include the robustness, identification of the important features via visualizations and no sparsity constraint requirement on the features. These are strong reasons for adopting PCA as the first step in the data pipeline of multivariate statistical learning methods and conducting a thorough investigation on its results.

It would naturally follow that the variables be plotted according to their importance in the PCs, since the PCs are considered as linear combinations of original variables that are always orthogonal to each other. As is shown in Fig. 2, this offers a new perspective to the CE data, overcoming the difficulty for having a direct graphical presentation in two dimensions for all 42 features. Figure 2 visualizes the dispersion of 42 features in the plot of their coefficients in the two most important PCs.

For this particular example dataset, Fig. 2 doesn’t identify any prominent outlying features due to that PC1 and PC2 only explains 22% of the total variance, and that the rest of the PCs each explains a small portion of variation. This reveals that the full dataset represents a sphere in 42 dimensions, since there is no
sharp variation along any feature. Figure 2 shows that clustering may exist around a subset of the characteristic variables. It is worth pointing out that since the PC1 vs. PC2 plot covers only two dimensions, having a small distance in Fig. 2 does not guarantee proximity when the 40 remaining dimensions have been taken into account.

Therefore, the proposed data pipelines further considers whether there is a strong grouping effect in the PC1 - PC2 plane by certain categorical social factors. In particular, we investigate the stratification based only on the race of the reference person in the CU. To avoid the overlap of colored points, the plots are displayed separately in Fig. 3 according to categorization of the ethnicity feature. The immediate perception is that white respondents forms the majority while there are very few Pacific Islanders or Native Americans. Out of 4,254 observations, 3,489 are labeled as “White”, and only 13 are labeled as “Pacific Islanders or Native Americans”. Also, for each subplot the general dispersion of observation points is similar to the subplot of “White”. For “Pacific Islanders or Native Americans" and “Multiple Races", there are too few observation points to make a conclusive argument, but for “Asians" the relative magnitude of dispersion along PC1 against dispersion along PC2 is larger compared to that in the subplot of “White". This means that in the PC1 - PC2 plane the “Asian" community’s projections have differences from the “White" community, when all features have been taken into account. It is also worth noting that after dividing into groups the restriction of the small sample size made it difficult to reach robust conclusions.

*Fig. 4* Biplot of first PC vs second PC
### Table 7
First five PCs from PCA based on selected CUs

| Variable                  | PC1  | PC2  | PC3  | PC4  | PC5  |
|---------------------------|------|------|------|------|------|
| HH_CU_Q                   | 0.01 | 0.02 | −0.10| 0.09 | −0.16|
| AS_COMP1                  | −0.17| −0.41| −0.03| −0.09| 0.17 |
| AS_COMP2                  | 0.08 | 0.45 | −0.01| −0.03| −0.09|
| CUTENURE                  | 0.08 | −0.02| −0.25| 0.03 | 0.03 |
| FAM_SIZE                  | −0.10| 0.25 | −0.17| −0.32| 0.17 |
| NO_EARNR                  | −0.25| 0.06 | −0.24| −0.08| −0.13|
| PERSLT18                  | −0.07| 0.23 | −0.19| −0.31| 0.17 |
| PERSOT64                  | 0.18 | 0.04 | 0.39 | 0.02 | 0.19 |
| VEHQ                      | −0.16| −0.05| 0.13 | −0.08| −0.14|
| AGE_REF                   | 0.16 | 0.01 | 0.42 | 0.02 | 0.15 |
| FINCBTXM                  | −0.33| 0.04 | 0.12 | 0.07 | −0.12|
| FSALARYM                  | −0.33| 0.03 | −0.02| 0.03 | −0.15|
| FRRETIRM                  | 0.16 | 0.03 | 0.39 | 0.00 | 0.18 |
| FSMFRXRM                  | −0.10| 0.02 | 0.04 | 0.05 | −0.06|
| TOTEXPCQ                  | −0.34| 0.10 | 0.19 | 0.06 | −0.03|
| FOODCQ                    | −0.29| 0.13 | 0.10 | −0.12| 0.27 |
| FDHOMEQC                  | −0.20| 0.16 | 0.08 | −0.18| 0.26 |
| FDAWAYCQ                  | −0.27| 0.04 | 0.07 | −0.01| 0.17 |
| HOUSCQ                    | −0.25| 0.09 | 0.14 | 0.03 | 0.07 |
| APPARCQ                   | −0.12| 0.09 | −0.01| 0.04 | 0.07 |
| TRANSCQ                   | −0.17| 0.05 | 0.12 | 0.05 | −0.14|
| HEALTHCQ                  | −0.08| 0.03 | 0.26 | 0.03 | 0.02 |
| ENTERTCQ                  | −0.11| 0.04 | 0.11 | 0.04 | −0.01|
| EDUCACQ                   | −0.06| 0.01 | −0.04| 0.06 | −0.04|
| READCQ                    | −0.11| 0.03 | 0.13 | 0.07 | −0.14|
| BLS_URBN_URBAN            | −0.06| 0.04 | −0.07| 0.53 | 0.12 |
| BLS_URBN_RURAL            | 0.06 | −0.05| 0.08 | −0.53| −0.12|
| EDUC_REF_NEVER            | 0.02 | 0.00 | 0.00 | 0.02 | −0.02|
| EDUC_REF_FIRST            | 0.04 | 0.00 | 0.00 | −0.08| 0.01 |
| EDUC_REF_SECOND           | 0.09 | 0.01 | −0.01| 0.04 | 0.18 |
| EDUC_REF_HIGH             | 0.07 | −0.05| 0.05 | −0.19| 0.04 |
| EDUC_REF_ASSC             | 0.00 | 0.03 | −0.02| −0.07| −0.05|
| EDUC_REF_UDGRD            | −0.09| 0.00 | −0.04| 0.15 | −0.14|
| EDUC_REF_GRD              | −0.12| 0.04 | 0.08 | 0.09 | −0.02|
| REF_RACE_WH               | 0.00 | −0.09| 0.20 | −0.16| −0.43|
| REF_RACE_BLK              | 0.03 | 0.08 | −0.17| 0.11 | 0.39 |
| REF_RACE_NA               | 0.00 | −0.02| −0.03| 0.03 | 0.05 |
| REF_RACE_AS               | −0.04| 0.04 | −0.07| 0.09 | 0.12 |
| REF_RACE_PI               | 0.00 | −0.01| −0.03| 0.01 | 0.02 |
| REF_RACE_MULTI            | −0.02| 0.00 | −0.04| 0.06 | 0.02 |
| SEX_REF_MALE              | −0.15| −0.45| −0.02| −0.06| 0.15 |
| SEX_REF_FEMALE            | 0.15 | 0.44 | 0.02 | 0.07 | −0.16|
Using the R package *factoextra* (Kassambara and Mundt 2020), the proposed data pipeline implements a very useful graphical presentation of PCA results, the biplot. Important information for quantitative interpretation is included in this visualization where the extracted features are represented as vectors. The length of a vector is the standard deviation, the dot product between two vectors is the covariance, the cosine of the angle between two vectors is the correlation, the inter-point distance between variables is the Mahalanobis distance, and the length of difference of two vectors is the standard deviation of the difference of the two variables. The transparency and the color reflects the contribution of each feature to PC1 and PC2 respectively. Based on Fig. 4, we can conclude that in the PC1 - PC2 plane, retirement income and age are close, and the number of earners and the family size are also close. Furthermore, expenditure variables including total expenditure, expenditure on eating out, expenditure on eating at home and housing expenditure are in the same direction as income before tax. The pattern of stretching from top-right to lower-left observed before in subplots of Fig. 3 can now be explained as variations due to income and expenditure variables.

While Fig. 3 investigates the influence of a categorical feature in the current dataset from the CE survey, the proposed data pipeline can also explore gender equality. We select households with only males among members over the age of 16, and households with only females among members over the age of 16, based on the variables for the number of males age 16 and over in the CU (AS_COMP1) and the number of females age 16 and over in the CU (AS_COMP2). In the CE survey, the variable for sex only records the gender of the reference person in the CU, but using these two features instead allows for learning the gender of all members in the household over 16. Then we apply PCA to the selected CUs with the same 42 features in the data pipeline.

Based on the subsample of the selected 1,696 consumer units, Tables 7 and 8 shows that the first three PCs explains a total of 30% of the variation. Compared with the PCA on the full sample, there is no apparent improvement in the interpretability of the loadings of PCs, nor is there a notable change in the proportion of variance explained by each PC. This potentially suggests that for those households with all members over the age of 16 belonging to the same gender, the income and expenditure traits are not vastly different from the other types of U.S. households. In reality, these selected households most likely include same sex couples, single parent, and people living on their own, among other possible scenarios. Therefore, intuition suggests that some difference in household income and consumer expenditure should exist when this subsample is compared to the national average. Given that subsampling under the current selection standard before performing PCA failed

|                | PC1 | PC2 | PC3 | PC4 | PC5 |
|----------------|-----|-----|-----|-----|-----|
| Standard deviation | 2.30 | 1.98 | 1.80 | 1.46 | 1.41 |
| Proportion of Variance | 0.13 | 0.09 | 0.08 | 0.05 | 0.05 |
| Cumulative Proportion | 0.13 | 0.22 | 0.30 | 0.35 | 0.39 |
to improve Tables 7 and 8 in terms of interpretability, the proposed data pipeline proceeds to analyze the graphical portrayal on this selected subsample in Fig. 5 for investigating the gender effect.

Using the package `ggplot2` (Wickham 2016) and the package `rgl` (Murdoch and Adler 2021) in R, we include the novel rotational three dimensional representation of PCA shown in Fig. 5 to provide a flexible and in-depth visualization for the data pipeline. This offers the researcher an innovative presentation of the dataset projections from a graphical perspective. The gray observation points represent the households with only females above the age of 16, while the black represent the households with only males above the age of 16. The upper-left subplot represents the PC1 - PC2 plane, the upper-right subplot shows the PC1 - PC3 plane, and the bottom subplot is the PC2 - PC3 plane. In these subplots, the three “axes” are the three PCs. The first observation is that clear segregation can be found due to gender. This is in line with the intuition that gender is a special factor in this selected subsample.

Fig. 5 3D representation of grouping in the first three PCs from PCA based on selected CUs
### Table 9  First three PCs of sparse PCA

| Variable                  | PC1  | PC2  | PC3  |
|---------------------------|------|------|------|
| HH_CU_Q                   | 0.00 | 0.00 | 0.00 |
| AS_COMP1                  | 0.00 | 0.00 | -0.37|
| AS_COMP2                  | 0.00 | 0.00 | 0.30 |
| CUTENURE                  | 0.00 | 0.00 | 0.00 |
| FAM_SIZE                  | 0.00 | 0.00 | 0.00 |
| NO_EARNR                  | 0.00 | -0.41| 0.00 |
| PERSLT18                  | 0.00 | 0.00 | 0.00 |
| PERSOT64                  | 0.00 | 0.51 | 0.00 |
| VEHQ                      | 0.00 | 0.00 | -0.13|
| AGE_REF                   | 0.00 | 0.47 | 0.00 |
| FINCBTXM                  | 0.00 | 0.00 | 0.00 |
| FSALARYM                  | 0.00 | -0.33| 0.00 |
| FRRETIM                   | 0.00 | 0.50 | 0.00 |
| FSMPFRXM                  | 0.00 | 0.00 | 0.00 |
| TOTEXPCQ                  | -0.44| 0.00 | 0.00 |
| FOODCQ                    | -0.52| 0.00 | 0.00 |
| FDHOMEQC                  | -0.43| 0.00 | 0.00 |
| FDAWAYCQ                  | -0.43| 0.00 | 0.00 |
| HOUSCQ                    | -0.42| 0.00 | 0.00 |
| APPARCCQ                  | 0.00 | 0.00 | 0.00 |
| TRANSCQ                   | 0.00 | 0.00 | 0.00 |
| HEALTHCQ                  | 0.00 | 0.00 | 0.00 |
| ENTERTCQ                  | 0.00 | 0.00 | 0.00 |
| EDUCACQ                   | 0.00 | 0.00 | 0.00 |
| READCQ                    | 0.00 | 0.00 | 0.00 |
| BLS_URBN_URBAN            | 0.00 | 0.00 | 0.00 |
| BLS_URBN_RURAL            | 0.00 | 0.00 | 0.00 |
| EDUC_REF_NEVER            | 0.00 | 0.00 | 0.00 |
| EDUC_REF_FIRST            | 0.00 | 0.00 | 0.00 |
| EDUC_REF_SECOND           | 0.00 | 0.00 | 0.00 |
| EDUC_REF_HIGH             | 0.00 | 0.00 | 0.00 |
| EDUC_REF_ASSC             | 0.00 | 0.00 | 0.00 |
| EDUC_REF_UDGRD            | 0.00 | 0.00 | 0.00 |
| EDUC_REF_GRD              | 0.00 | 0.00 | 0.00 |
| REF_RACE_WH               | 0.00 | 0.00 | 0.00 |
| REF_RACE_BLK              | 0.00 | 0.00 | 0.00 |
| REF_RACE_NA               | 0.00 | 0.00 | 0.00 |
| REF_RACE_AS               | 0.00 | 0.00 | 0.00 |
| REF_RACE_PI               | 0.00 | 0.00 | 0.00 |
| REF_RACE_MULTI            | 0.00 | 0.00 | 0.00 |
| SEX_REF_MALE              | 0.00 | 0.00 | -0.61|
| SEX_REF_FEMALE            | 0.00 | 0.00 | 0.61 |
The second finding is that most data points are dispersed in a three-dimensional region enclosed by three “axes”, as can be seen most clearly in the upper-right subplot. This visual presentation helps the construction of an intrinsic idea of how the data is organized in reference to the PCs.

Adopting PCA in the data pipeline, we have revealed details based on a categorical feature in Fig. 3 and found support for an intuitive grouping of the features through the biplot in Fig. 4. PCA has provided useful insights into variable interactions, as well as data stratification by race and gender. Furthermore, on a selected subset of households, the PCs and the three dimensional projections show that better interpretations are attainable with a smaller dataset for consideration. While PCA is a statistical learning method that is easy to implement and explain, given the 42 variables in the dataset, one deficiency of the usual PCA is the difficulty in interpreting the PCs and ineffectiveness in dimension reduction. The usual PCA has laid the foundation for pursuing the more appropriate sparse PCA approach in the following section of the data pipeline.

Sparse PCA is suitable when not every feature in the dataset has a nonzero coefficient in the PCs. To achieve this, a sparsity constraint needs to be enforced, as was proposed by Zou et al. (2006). The R package *nsprcomp*(Sigg and Buhmann 2008)
contains the function \texttt{nsprcomp} for conducting sparse PCA. Sparse PCA is based on finding PCs under the constraint that the number of nonzero components in the PC is less than or equal to a chosen constant. Function \texttt{nsprcomp} achieves this through Expectation-Maximization (EM) iterations, where soft-thresholding produces the desired sparsity, producing PCs and the proportion of variance explained. However, it is important to set the starting number for random number generation in advance to ensure the results are reproducible. Centering and scaling have been imposed on \texttt{fmlifull} and the loadings of the PCs are also allowed to be negative. We consider the first three PCs for simplicity of interpretation and presentation. The number of nonzero factor loadings for each PC is set to five. The resulting PCs and their importance are shown in Tables 9 and 10.

There are notable differences in the amount of variance respectively explained by the three PCs. PC1 explains 38%, PC2 explains 34.0% and PC3 accounts for the remaining 28%. Computations in the sparse PCA involves forcing the constraint on the dataset, therefore forcing most coefficients to be zero, leaving the specified number of nonzero coefficients for each PC. No apparent “elbow” can be found in Fig. 6, the scree plot. Based on the loadings in Table. 9, PC1 represents the five most important factors of income and expenditure, including the annual income after tax, the annual income from salary and wages, the quarterly total expenditure, the quarterly food expenditure and the quarterly housing expenditure. PC2 is more concerned with the age structure, a contrast between the pair formed by the number of earners (NO_EARNR), the number of children less than 18 in the CU (PERSLT18), and the trio consisting of the number of adults over 64 in the CU (PERSLT64), age and FRRETIRM, which is the amount of Social Security and Railroad Retirement income prior to deductions for medical insurance and Medicare. PC3 is about the
gender, male contrasting against female, comparing the trio of the number of males age 16 and over in the CU, the number of owned vehicles (VEHQ), and the indicator for having a male reference person (SEX_REF_MALE), against the pair formed by the number of females age 16 and over in the CU and the indicator for having a male reference person (SEX_REF_FEMALE). It is interesting to find that number of vehicles owned is counted as an index of masculinity. Sparse PCA has improved the clarity of interpretations on results from the data pipeline.

However, in terms of visualizations sparse PCA does not provide a informative biplot. The more instructive plot is the PC1 vs. PC2 plot, since not only that PC1 and PC2 account for 72% of total variance for this dataset, but PC1 also reflects influence of income and expenditure while PC2 contrasts age differences. The results from the sparse PCA using nsprcomp are based on that of the regular PCA using prcomp. Therefore, in Fig. 7, the new PC1 vs. PC2 plot for the sparse PCA, we are interested in the dispersion of observations instead of the dispersion of features in Fig. 2. Each observation is marked as a dot in the PC1 - PC2 plane in this plot. Figure 7 shows a different dispersion pattern compared with results from the regular PCA, with dispersion along PC1 being approximately twice the dispersion along PC2.

In order to compare the effectiveness of the sparse PCA for interpretations with the regular PCA based on the selected subsample of 1,696 observations used for Tables 7 and 8, the same subsample selected according to gender is analyzed. Tables 11 and 12 give the results of this sparse PCA. Among the three PCs computed in the sparse PCA, PC1 explains 38% of the total variation, and the other two each accounts for about 30%. PC1 can be interpreted as a contrast between male and female respondents. PC2 is a weighted average of the important income and expenditure variables, including the annual income after tax, the annual income from salary and wages, the quarterly total expenditure and the quarterly food expenditure. PC3 is a contrast between the number of people in the CU, reflected by features for the number of members in the CU (FAM_SIZE), the number of earners, the age of the reference person (AGE_REF), and the presence of senior citizens in the CU, represented by PERSOT64 and the Social Security and Railroad Retirement income. The first three PCs from the sparse PCA each explains much more variation than the first three PCs from the regular PCA shown in Table 8, but this is partly due to subjectively choosing to use exactly three PCs to explain the entire variation in the construction of sparse PCA.

We present in Fig. 8 the three dimensional PC1 – PC2 – PC3 projections of the dataset with black points indicating CUs with male reference persons and grey points indicating female. A clear segregation along PC1 is because this PC represents the contrast of gender. Sparse PCA has successfully captured the most important characteristic of this selected subsample by producing PC1 as the contrast of genders. Sparse PCA has also made interpretations less demanding and revealed details the standard PCA did not uncover. Also, there appears to be some stratification of the sample points along PC1. The three PCs produced by sparse PCA identify related variables that are important to consider further in the data pipeline, providing us with the first criterion for further selection and grouping in Sect.4.
Table 11  First three three PCs from sparse PCA based on selected CUs

|                   | PC1 | PC2 | PC3 |
|-------------------|-----|-----|-----|
| HH_CU_Q           | 0.00| 0.00| 0.00|
| AS_COMP1          | -0.48| 0.00| 0.00|
| AS_COMP2          | 0.48 | 0.00| 0.00|
| CUTENURE          | 0.00 | 0.00| 0.00|
| FAM_SIZE          | 0.15 | 0.00| 0.23|
| NO_EARNR          | 0.00 | 0.00| 0.43|
| PERSLT18          | 0.00 | 0.00| 0.00|
| PERSOT64          | 0.00 | 0.00| -0.51|
| VEHQ              | 0.00 | 0.00| 0.00|
| AGE_REF           | 0.00 | 0.00| -0.50|
| FINCBTXM          | 0.00 | 0.48| 0.00|
| FSALARYM          | 0.00 | 0.46| 0.00|
| FRTCRETIRM        | 0.00 | 0.00| -0.49|
| FSPMPFRXM         | 0.00 | 0.00| 0.00|
| TODTEXPCQ         | 0.00 | 0.49| 0.00|
| FOODCQ            | 0.00 | 0.38| 0.00|
| FDHOMEFCQ         | 0.00 | 0.00| 0.00|
| FDWAYACYC         | 0.00 | 0.00| 0.00|
| HOUSCQ            | 0.00 | 0.42| 0.00|
| APARCCQ           | 0.00 | 0.00| 0.00|
| TRANSCQ           | 0.00 | 0.00| 0.00|
| HEALTHCQ          | 0.00 | 0.00| 0.00|
| ENTERTCQ          | 0.00 | 0.00| 0.00|
| EDUCACQ           | 0.00 | 0.00| 0.00|
| READCQ            | 0.00 | 0.00| 0.00|
| BLS_URBN_URBAN    | 0.00 | 0.00| 0.00|
| BLS_URBN_RURAL    | 0.00 | 0.00| 0.00|
| EDUC_REF_NEVER    | 0.00 | 0.00| 0.00|
| EDUC_REF_FIRST    | 0.00 | 0.00| 0.00|
| EDUC_REF_SECOND   | 0.00 | 0.00| 0.00|
| EDUC_REF_HIGH     | 0.00 | 0.00| 0.00|
| EDUC_REF_ASSC     | 0.00 | 0.00| 0.00|
| EDUC_REF_UDGRD    | 0.00 | 0.00| 0.00|
| EDUC_REF_GRD      | 0.00 | 0.00| 0.00|
| REF_RACE_WH       | 0.00 | 0.00| 0.00|
| REF_RACE_BLK      | 0.00 | 0.00| 0.00|
| REF_RACE_NA       | 0.00 | 0.00| 0.00|
| REF_RACE_AS       | 0.00 | 0.00| 0.00|
| REF_RACE_PI       | 0.00 | 0.00| 0.00|
| REF_RACE_MULTI    | 0.00 | 0.00| 0.00|
| SEX_REF_MALE      | -0.51| 0.00| 0.00|
| SEX_REF_FEMALE    | 0.51 | 0.00| 0.00|
Based on the results from Sect. 3, the intuitive grouping of features into categories of Characteristics, Income and Expenditure is reasonable. Furthermore, sparse PCA has revealed features that coincide with this intuition, which is supported by findings from the usual PCA. Further details need to be explored in the data pipeline by

**Table 12** Importance of first three PCs from sparse PCA based on selected CUs

|                | PC1  | PC2  | PC3  |
|----------------|------|------|------|
| Standard deviation | 1.93 | 1.75 | 1.71 |
| Proportion of Variance | 0.38 | 0.32 | 0.30 |
| Cumulative Proportion    | 0.38 | 0.70 | 1.00 |

**Fig. 8** 3D representation of grouping in the first three PC from sparse PCA based on selected CUs

### 4 Cluster analysis

Based on the results from Sect. 3, the intuitive grouping of features into categories of Characteristics, Income and Expenditure is reasonable. Furthermore, sparse PCA has revealed features that coincide with this intuition, which is supported by findings from the usual PCA. Further details need to be explored in the data pipeline by
choosing the more important and informative features and grouping features of the sample points accordingly to reveal implicit interactions. The proposed data pipeline utilizes another multivariate statistical learning method, cluster analysis, and aims to discover clusters of features based on distances such that it either reinforces or contradicts the intuitive classification of features. Moreover, the process may help indicate the relative significance of some features. Since cluster analysis is one of the most developed multivariate statistical learning methods, we conduct a comparison of several different approaches for clustering on the survey dataset. Clustering of features is an essential step before advancing to investigate the dependencies between features.

The idea of cluster analysis is to use the values of the features to cluster the observations into groups such that each cluster only contains similar observations. Here, the interest lies in the clustering of features instead of the observations. One may refer to Anderberg (1973) or Everitt et al. (2001) for a more in-depth discussion on the topic. A more theoretical discussion can be found in Fraley and Raftery (2002). Clustering of observations leads to either a single cluster, or clusters with no interpretable traits. In R, there are several packages that aim at providing a robust clustering of features. One popular R package is ClustOfVar (Chavent et al. 2017b), which includes a function hclustvar that provides ascendant hierarchical clustering of a set of features. This method centers around the first PC produced by PCAmix (Chavent et al. 2017a), which was developed to incorporate a mixture of categorical and continuous features. Since it has been verified in Sect. 3 that PCA produces reasonable results, we implement this procedure to reach the cluster dendrogram shown in Fig. 9.

![Cluster dendrogram](image-url)
Figure 9 brings straightforward insights that match with intuition. Age, the number of seniors, and retirement income are in a cluster, which is in line with the second PC of the sparse PCA in Sect. 3. Food related expenditure form a small cluster, as do education, reading, transportation and the total expenditure. Not surprisingly,
references to having an undergraduate or graduate degree is close to income, housing expenditure and salary. As expected, the features for the gender of the reference person are in the same cluster that was identified in the start. Similar explanation holds for the urban references. On the more general level, there are three main clusters. From left to right, clusters respectively represent age and family composition, education as well as income and expenditure, and demographic categories of gender, race and geography. This is different from the intuitive grouping into categories of Characteristics, Income and Expenditure, because the clustering of features by \textit{hclustvar} is based on homogeneity for the clusters when merging in an hierarchical fashion. Therefore, similarities between the features become the deciding factor, and the pattern of response is compared between the features. The criterion is no longer the intuitive sociological classifications. The clustering performed is useful for future investigations of the dependencies between groups of features.

Although \textit{ClustOfVar} is a powerful multivariate statistical learning tool, we implement alternative R packages for clustering and compare the results from the data pipeline. A comprehensive R package for multivariate statistical learning is \textit{Hmisc} (Harrell Jr 2021), which includes the function \textit{varclus} that conducts hierarchical cluster analysis on features using the Hoeffding D statistic, squared Pearson or Spearman rho correlations. It is important to note that the function conducts pair-wise deletion of “NA” values implicitly, and the default aggregation criterion is the squared Spearman correlation coefficient, which is used for detecting monotonic but nonlinear relationships, while Hoeffding’s D statistic is sensitive to many types of dependence, including highly non-monotonic relationships. On the other hand, the difference between Spearman’s correlation and Pearson’s
correlation is that one is more suitable for ordinal scales and the other for interval scales. Compared to the Spearman’s correlation, Pearson’s correlation is better for linear relationships. However, the actual mathematical justification is beyond these explanations. The best approach is to compare all three methods, separately given in Figs. 10, 11, and 12.

The limitations imposed by mathematical implications of each criterion is a challenge for the interpretation and comparison of the results. We focus on verifying the grouping of features, revealing implicit connections, and selecting suitable clustering for further analysis. These objectives requires a direct comparison of the clustering results in the dendrograms. In general, although the clusters are still hierarchical, there are no longer clear levels of aggregation, as was in Fig. 9. All three have the tendency of grouping all features into one cluster in total. In particular, this is most prominent in Fig. 11. However, at the more detailed level, the three criteria have been successful in selecting the important features and clustering on a smaller scale. The clustering of FOODCQ and FDHOMEQCQ, the clustering of AGE_REF, PERSOT64 and FRRETIR, the clustering of FAM_SIZE and PERSLT18, the clustering of FINCBTXM and FSALARYM, and the clustering of TOTEXPCQ, HOUSCQ and TRANSCQ are found under all three criteria. This reaffirms the results from ClustOfVar, and provides a strong basis for solving the challenge in statistical learning due to high correlation between similar features.

5 Canonical correlation analysis

To investigate how groups of features influence each other as groups instead of individual features, we incorporate findings in previous sections from PCA and cluster analysis on important features and reasonable grouping to obtain meaningful grouping of features. CCA is implemented in the proposed data pipeline to find the linear combination within each group that achieves the maximum possible correlation between two groups. CCA was pioneered by the works of Hotelling (1935), while the fundamental idea behind this multivariate statistical learning method was discussed in detail in Hotelling (1936). Further discussion on the topic can be found in Knapp (1978). Rencher (1992) provides a comprehensive discussion on the interpretation of the CCA results.

Canonical Correlation Analysis (CCA) requires the user to provide two groups of features and computes the linear combination within each group that will give the largest correlation between the two linear combinations. Once the two groups of features are selected, CCA proceeds to find pairs of canonical features to take a linear combination of features in first group with unit variance and a linear combination of features in the second group with unit variance. CCA forms a pair of canonical features from the linear combinations. The first pair, \((U_1, V_1)\), maximizes the correlation. Then define \(E_1 = \Sigma_1^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\), \(E_2 = \Sigma_2^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\). \(\Sigma_{11}\), \(\Sigma_{22}\) are the covariance matrices of the first and second groups of features, and \(\Sigma_{12}\) is the covariance between the two groups. The coefficients in \(U_1\) are given by the eigenvector of \(E_1\) based on the covariance structure of the two groups, whereas
the coefficients in $V_1$ is given by the eigenvector of $E_2$ based on the covariance structure of the two groups. The maximum correlation based on the first pair of canonical features is the square root of the first nonzero eigenvalue of $E_1$, which is equal to the first nonzero eigenvalue of $E_2$.

Therefore, an important task is to find the appropriate groups from features in the data pipeline. This should be based on the research question. For example, for the CE survey we would like to explore if there is dependency between the income features and expenditure features in the *fmlifull* dataset. Ideally, we would include as many features as possible to avoid loss of information and negligence of implicit influence between and within each group. After standardizing the features, we proceed to form the first group with the four income features and the second group with all eleven expenditure features. The income features and the expenditure features measure different periods, since they are based on income in the last twelve months and the expenditure in the last three months, respectively. While this should not cause any computational hardship, it increases the difficulty for making a reasonable interpretation. The resulting eigenvalues in the CCA, as the square of the canonical correlations, are complex. The correlation matrix in Sect. 2.3 has revealed that high correlations exist within the group of income features and the group of expenditure features. For example, there is a 0.85 correlation between the quarterly food expenditure and the quarterly food expenditure on eating at home. This will heavily influence the performance of CCA. Therefore, improvements on the CCA are necessary in the data pipeline. Here we revisit previous grouping results from Sect. 4. Instead of eleven expenditure features, the four most representative ones are chosen since they have less correlation. Then CCA is implemented on the four income features. The selection criterion is based primarily on Fig. 9, where four clusters are highlighted to support the decision of selecting the four representative features since within each cluster, the features have higher similarity. To determine which variable is most likely to be the most representative within each cluster, we revisit the contrast between groups of features reported by PCs in Table 9 from sparse PCA. The first three PCs point to the quarterly food expenditure, the quarterly total expenditure and the quarterly housing expenditure. The last cluster contains only one expenditure variable, the

| Table 13 | Eigenvectors for new $E_1$, $E_2$ |
|----------|----------------------------------|
|          | Eigenvector 1 | Eigenvector 2 | Eigenvector 3 | Eigenvector 4 |
| $E_1$    | 0.99          | 0.57          | -0.57         | -0.68         |
|          | 0.13          | -0.58         | 0.73          | 0.66          |
|          | 0.05          | 0.57          | 0.33          | 0.14          |
|          | -0.03         | 0.11          | -0.19         | 0.28          |
| $E_2$    | 0.79          | 0.02          | 0.61          | 0.60          |
|          | 0.31          | -0.16         | 0.18          | -0.79         |
|          | 0.53          | -0.31         | -0.75         | 0.06          |
|          | 0.05          | 0.94          | -0.18         | -0.09         |
quarterly healthcare expenditure. This forms the group of four expenditure features needed for CCA.

Square root of eigenvalues of the CCA using four income features and four expenditure features are (0.59, 0.19, 0.04, 0.00) for $E_1$ and for $E_2$, whereas eigenvectors are given in Table 13. Complex eigenvalues are avoided. The first square root of eigenvalues is 0.59 for both $E_1$ and $E_2$, and represents the correlation between the linear combination of the first group of income features, $U_1$, and the linear combination of the second group of selected expenditure features, $V_1$.

$$U_1 = 0.99 \text{FINCBTXM} + 0.57 \text{FSALARYM} - 0.57 \text{FRRETIRM} - 0.68 \text{FSMPFRXM},$$

$$V_1 = 0.79 \text{TOTEXPCQ} + 0.02 \text{FOODCQ} + 0.61 \text{HOUSCQ} + 0.60 \text{HEALTHCQ}. \tag{1}$$

This correlation of 0.59 is quite high, and the result is consistent with the intuitive understanding that expenditure features should have dependence on the income features. However, it is important to avoid mistakenly taking the linear combination coefficients of $U_1$ and $V_1$ to interpret them as the coefficients for quantifying the relationship with the original features. As shown in the results from the proposed data pipeline, this is not valid. Instead, using correlation coefficients, $U_1$ can be seen as a measure of high before tax income and salary with low retirement and self-employment income. $V_1$ is a combination of the total expenditure and the expenditure on housing and healthcare. The food expenditure is less notable. Furthermore, CCA verified the previous conclusions from cluster analysis.

However, CCA does have several shortcomings. Implicit influences not included in either of the two groups are not accounted for. The basic requirement of having appropriate, meaningful groups prior to conducting CCA may present difficulty for larger datasets where many factors do not have distinct and interpretable meanings for grouping using intuition. More importantly, CCA uses linear combinations of original features to show the between-group correlation, but the relationship between the linear combinations and the original features is best reflected in the correlations instead of the linear coefficients. This adds complexity to the interpretation. In conclusion, CCA in the proposed data pipeline utilizes the results from previous PCA and cluster analysis and discovers the 0.59 correlation between $U_1$ which represents higher income stable employment and $V_1$ which represents bulk expenditure. This relationship is consistent with intuition and reflects the dependency of expenditure on income for households in the United States.

6 Conclusion

This study proposes a data pipeline based on multivariate statistical learning methods on an economic survey dataset. The data pipeline is available for research on general social science surveys, and the discussion has provided details that promotes the significance of the CE survey data as a well-designed and comprehensive public data source. The missing value problem that is common in large survey datasets has
been considered carefully in this study. Moreover, the CE dataset has provided challenges for multivariate statistical learning methods such as PCA and CCA. The data pipeline offers a flexible framework, allowing for adjustments or extensions to be adopted to overcome real data challenges and provide comparisons for results from different approaches. Informative interpretation based on direct visualizations is the main theme of the data pipeline. Reasonable grouping of features and relationships between groups have been analyzed for economic understanding. Further research on the CE dataset may extend to applications of deep learning methods to explore in greater details the relative importance of the variables. We are currently combining the CE dataset with Alzheimer’s Disease survey dataset from Alzheimer’s Disease Neuroimaging Initiative (ADNI), so links between consumption/income and the onset of the disease can be analyzed with the data pipeline using multivariate statistical learning, developing on the prediction framework of Hu et al. (2017). Furthermore, in future research alternative predictive methods, including the grey forecasting models (Wang et al. 2017), may be explored and incorporated for fresh perspectives based on the CE dataset with a focus on uncertainty considerations.

Supplementary information  The online version contains supplementary material available at (https://doi.org/10.1007/s00180-022-01251-2).

Acknowledgements  The author would like to express his gratitude to the Bureau of Labor Statistics for providing the dataset on which this research was based. Furthermore, guidance from Prof. Weiyin Loh of the department of statistics, University of Wisconsin Madison is greatly appreciated. More importantly, the author would like to thank the reviewers for their expert opinions and detailed comments which were crucial to the revision of the manuscript.

Open Access  This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Anderberg MR (1973) Cluster analysis for applications. Academic Press, New York
Anderson TW (1963) Asymptotic theory for principal component analysis. Ann Math Stat 34(1):122–148
Cadima J, Jolliffe I (1995) Loading and correlations in the interpretation of principle components. J Appl Stat 22(2):203–214
Chavent M, Kuentz V, Labenne A, Liquet B, Saracco J (2017a) PCAmixdata: Multivariate Analysis of Mixed Data. https://CRAN.R-project.org/package=PCAmixdata, R package version 3.1
Chavent M, Kuentz V, Liquet B, Saracco J (2017b) ClustOfVar: Clustering of Variables. https://CRAN.R-project.org/package=ClustOfVar, R package version 1.1
Everitt BS, Landau S, Leese M (2001) Cluster analysis. Hodder Arnold, London
Fraley C, Raftery AE (2002) Model based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97(458):611–631
Garner T, Martinez W (2022) The 2017 data challenge of the american statistical association. Computational Statistic Forthcoming
Harrell Jr FE (2021) Hmisc: Harrell Miscellaneous. https://CRAN.R-project.org/package=Hmisc, R package version 4.6-0

Hotelling H (1935) The most predictable criterion. J Educ Psychol 26(2):139–142

Hotelling H (1936) Relations between two sets of variables. Biometrika 26(3/4):321–377

Hu M, Zhang Y, Dowling NM (2017) An analysis of factors predicting memory loss in alzheimer’s disease prevention. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp 1281–1288, https://doi.org/10.1109/IJCNN.2017.7966000

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River, NJ

Kassambara A, Mundt F (2020) factoextra: Extract and Visualize the Results of Multivariate Data Analyses. https://CRAN.R-project.org/package=factoextra, R package version 1.0.7

Knapp T (1978) Canonical correlation analysis: a general parametric significance-testing system. Psychol Bull 85(2):410–416

Loh WY, Eltinge JL, Cho MJ, Li Y (2019) Classification and regression trees and forests for incomplete data from sample surveys. Stat Sin 29(1):431–453

Murdoch D, Adler D (2021) rgl: 3D Visualization Using OpenGL. https://CRAN.R-project.org/package=rgl, R package version 0.108.3

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Phil Mag 2(11):559–572

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

Rencher AC (1992) Interpretation of canonical discriminant functions, canonical variates and principal components. Am Stat 46:217–225

Sigg CD, Buhmann JM (2008) Expectation-maximization for sparse and non-negative pca. In: Proc. 25th International Conference on Machine Learning, https://doi.org/10.1145/1390156.1390277

van Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. J Stat Softw 45(3):1–67. https://doi.org/10.18637/jss.v045.i03

Wang H, Zhao L, Hu M (2017) The morbidity of multivariable grey model MGM(1, m). Int J Differ Eq 2017:1–5. https://doi.org/10.1155/2017/2495686

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York

Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15(2):265–286

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.