CATPCA in R: Application to Sensorial Data of Wine-Tasting

Andreia Gomes[1(✉)] and Ana Cristina Braga[2]

[1] School of Engineering, University of Minho, 4710-057 Braga, Portugal
andreiagomes_6@outlook.com
[2] ALGORITMI Research Centre, LASI, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
acb@dps.uminho.pt

Abstract. Analyzing large volumes of categorical data presents a recurring challenge in the social, behavioral, and biological sciences. Therefore, there is a necessity to reduce these data while maintaining minimal loss of information. CATPCA serves as an alternative to conventional Principal Component Analysis (PCA), enabling the reduction of dimensionality for variables measured on different scales. To elucidate the underlying principles of this statistical method, a review of bibliographic sources was undertaken. This search also highlighted the `Gifi` and the `Homals` packages, identified as the sole packages equipped with functions supporting the implementation of CATPCA. Both packages were explored to assess and compare their functionalities and features. The `Gifi` package is derived from the `Homals` package, aiming for enhanced user-friendliness and flexibility. To further delve into the functionalities and constraints of the `Gifi` package and to demonstrate the handling of biological data, an analysis of sensory data from wine-tasting trials was additionally conducted. In order to enable users to perform an intuitive and simplified CATPCA, a web application named ***Gify***, available at https://andreiagomes.shinyapps.io/Gify/, based on the `princals()` function from the `Gifi` package, has been developed.

Keywords: Categorical variables · PCA · CATPCA · R · Shiny

1 Introduction

Large datasets are increasingly prevalent in many areas of the sciences, such as social, behavioral, and biological fields. To interpret such datasets effectively, methods are required to significantly reduce their dimensionality in an interpretable manner while preserving a substantial amount of the information contained within them [3,11].

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14815, pp. 357–374, 2024.
https://doi.org/10.1007/978-3-031-65154-0_22

This reduction will be able to remove the noise, but retaining the signal of interest that can be crucial for understanding hidden patterns. On the other hand, the original, high-dimensional data may contain non-informative or redundant variables [17].

Many techniques have been developed for this purpose. Principal Component Analysis (PCA) is one of the oldest and most commonly used techniques of multivariate analysis [1]. Its principle is simple: to reduce the dimensionality of a dataset while simultaneously preserving the maximum variability possible [5]. This dimensionality reduction reveals the underlying structure, sometimes hidden and simplified, that often exists within it [22].

This technique allows for explaining the variance-covariance structure of a set of dependent variables, which are typically correlated, in order to transform them into a set of new, uncorrelated orthogonal variables called principal components (PCs). These PCs are linear combinations of the original set [1,4].

The PCs resulting from the decomposition of the covariance matrix are presented in descending order of their importance. That is, the first principal component (PC) explains the maximum variability in the data, the second PC explains the maximum variability remaining in the set after removing the first component, and so on, until the last PC, which expresses the least contribution to explaining variations in the initial variables. In this case, PCs with lower variability can be eliminated without leading to a significant loss of information in the initial dataset. The quality of PCA results improves as fewer linear combinations are needed to explain the total variance of the data [2,5,23].

However, this analysis is only suitable for continuous variables that are measured on a numeric level, assuming the existence of a linear relationship between them. Additionally, it requires a large sample size for accurate results. Therefore, it is not an appropriate method for dimensionality reduction when dealing with categorical variables that exhibit non-linear relationships [10,15].

Thus, the categorical principal component analysis (CATPCA), also known as nonlinear principal component analysis (NLPCA), emerges as an alternative to traditional PCA [11]. This approach aims at the same objectives as PCA, which is to reduce a dataset consisting of many variables and complicated correlation patterns to a smaller number of uncorrelated summary variables (PCs) that closely represent the information in the data. However, in this case, CATPCA is suitable for mixed-level measurement variables (nominal, ordinal, and numeric) that may not be linearly related to each other [11,15,20].

CATPCA technique quantifies categorical variables using optimal scaling, assigning numerical quantifications to the categories of each qualitative variable so that the maximum variance of the quantified variables is accounted for [24].

Therefore, the main motivation of this paper is to explore the multivariate statistical analysis tools for qualitative variables, particularly CATPCA, available in R. The utility of this methodology will be demonstrated by applying it to a real dataset of wine-tasting trials.

This paper is organized into 5 sections. Section 2 presents the theoretical and mathematical foundations of concepts related to CATPCA that support this research. Section 3 explores the `Gifi` and `Homals` packages available in the R software, which enable the execution of CATPCA. Section 4 presents a CATPCA analysis utilizing the `princals()` function from the `Gifi` package, applied to a dataset comprising wine-tasting evaluations. In Sect. 5, the ***Gify*** application, developed in R Shiny to simplify the implementation of the `princals()` function from the `Gifi` package, is introduced. Finally, Chap. 6 encapsulates the key findings described in this paper, accompanied by future outlooks for this research.

2 Methods

In CATPCA, all variables are considered categorical, however, the quantification assigned to each category takes into account the nature of these variables [11, 18]. This quantification is based on the optimal scaling process, which relies on minimizing the loss function of Alternating Least Squares (ALS) [11]. The loss to be minimized is the information loss associated with the fact that variables are now represented by a smaller number of components.

After the quantification of the variables, the $\mathbf{Q}_{n\times m}$ matrix is obtained, replacing the original data matrix $\mathbf{H}_{n\times m}$, and it contains the transformed variables $\mathbf{q}_j = \phi_j(\mathbf{h}_j)$, where j represents the index of the variables in the matrix $\mathbf{H}$, and $\mathbf{h}_j$ is vector of size $n \times 1$, with $j = 1, ..., m$ [11]. The minimization criterion is then applied through the $\mathbf{Q}$ matrix [21]:

$$tr(\mathbf{Q} - \mathbf{YB}^T)^T(\mathbf{Q} - \mathbf{YB}^T) \tag{1}$$

Concerning the $\mathbf{Y}_{n\times p}$ matrix, where the columns represent linear functions of the columns of the $\mathbf{Q}$ matrix, this matrix is composed of the coordinates (scores) of the first p PCs, where p is the number of PCs defined for the analysis [11]. Regarding the $\mathbf{B}^T_{p\times m}$, where the (orthogonal) columns represent the eigenvectors of the variance-covariance matrix of $\mathbf{Y}$ [11,21]. In the case of the latter, the matrix is composed of the coefficients (loadings) of the first p PCs [21].

2.1 Analysis Levels

Nominal Analysis Level. In this analysis level, objects with the same category in the original dataset obtain the same optimal quantification. Orders or distances between categories are not considered [6,11].

Ordinal Analysis Level. Considers the original order of categories, in addition to taking into account objects belonging to the same category, as described in the nominal analysis. All quantifications must be less than or equal to the quantification of the category with the highest value in the original data [6,11].

Numeric Analysis Level. It is considered the most restrictive since it takes into account the objects belonging to each category while also maintaining the order and distance between categories. This distance is always equal [6].

Spline Analysis Level. Ordinal and nominal levels of analysis may lead to irregular quantifications due to the use of step functions, while spline functions offer a smoother and more stable alternative [6,10,11].

This spline transformation involves dividing the x-axis of the transformation plot into pre-defined intervals and adjusting these intervals using individual low-degree polynomial functions [10,16]. The locations where these curves intersect are called knots [7].

Various authors [11,12] recommend applying ordinal analysis when the number of categories is small and monotonic spline analysis when the number of categories exceeds the number of samples or is greater than 10. In these cases, linear transformation is too restrictive, and ordinal transformation is too specific, while monotonic spline transformation allows for maintaining the order of categories.

2.2 CATPCA Metrics

Component Loadings. Allows the quantification of the influence of the transformed variables in creating the PC [7,16]. In CATPCA, the component loading corresponds to the correlation between the transformed (quantified) variables and the retained non-linear PCs, so its value varies between -1 and 1 [11].

Variance Accounted for (VAF). It is the most important measure for quantifying the effectiveness of the method [6,18].

Nominal and ordinal level analyses are based on maximizing the p largest eigenvalues of the correlation matrix between the transformed variables $\mathbf{q}_j$. Each eigenvalue (which corresponds to the sum of the squared component loadings of each dimension) represents the amount of variance explained by each PC, and the sum of these eigenvalues divided by the number of analyzed variables is equivalent to the cumulative VAF [16]. The cumulative VAF results from the sum of the VAF of the considered components [18]. Calculating the VAF for each PC (dividing each eigenvalue by the number of analyzed variables) allows understanding the relative importance of each component in explaining the overall variance of the data, aiding in determining which PCs contribute more to its structure [10].

Crohnbach's Alpha (α_s). The most commonly used coefficients to measure internal consistency in applied psychometrics [16]. The higher its value, the closer to 1 (varies in between 0 and 1), the greater the reliability of the model and each dimension [18]. The relationship between the eigenvalue and Cronbach's alpha is used to assess the internal consistency of each dimension [16].

Principal Components Number. The number of chosen components depends on the information that can be excluded and the interpretation of these components themselves.

Its choice can be defined through the *Scree plot*, which represents the eigenvalues and the PCs. In this case, these plots show a break, or an elbow, that indicates the number of components to be considered [11,18]. Alternatively, the choice can be based on the eigenvalue, where PCs with eigenvalues greater than 1 are selected [18].

3 R Packages

The exploration of the R software, combined with the literature review, was essential to investigate which tools are currently available and enable the execution of CATPCA, thereby facilitating the handling of data with categorical variables.

The `Gifi` [14] and the `Homals` [13] packages stand out as the only ones equipped with functions (`princals()` and `homals()`, respectively) that allow the application of CATPCA. Both functions were explored using the same set of sample data, and a detailed analysis of their arguments and the values and graphs obtained was carried out to compare the functionalities and features that both packages offer.

3.1 Gifi

The `Gifi` package [14], developed in R by Patrick Mair, Jan de Leeuw, and Patrick J. F. Groenen, allows for the implementation of CATPCA (PRINCALS), multiple correspondence analysis (HOMALS), and monotone regression analysis (MORALS).

In this package, there are 12 example datasets and 9 functions, with the main function being the `princals()` command. This function adjusts the CATPCA and allows the user to define a total of 14 arguments. Simpler analysis levels such as metric, ordinal, and nominal can be specified through the `levels` argument, and the corresponding spline transformations are internally created [14].

The auxiliary function `knotsGifi()` enables the implementation of spline functions, allowing the specification of the type and number of knots for each variable. The type of knots (type) is "Q" when the knots are quantiles of the variables, "R" when the knots are equally spaced, "E" when there are no internal knots, and "D" when the internal knots correspond to data points. In the case of type = "E" and type = "D", it is not necessary to define the number of knots [9]. Similarly, it is essential to determine the degree of the spline function using the `degrees` argument [12].

The `copies` argument, generating as many copies as the number of PCs, allows for the relaxation of linear constraints on nominal variables, enabling the categories of variables to spread freely in the p space, where p is the number of PCs defined for the analysis [9,12].

The `missing` argument allows the user to choose the treatment for missing values (NAs). Regarding the `active` argument, when a variable is considered passive, it is not optimized and therefore does not contribute to the loss function, but can be included in the plots either way [9].

The plots depicting the results can be generated using the `plot.princals()` function, also included in the `Gifi` package [14].

3.2 Homals

The `Homals` package [13], currently available in R, was developed by Patrick Mair, with co-authorship by Jan De Leeuw.

This package allows for the conduct of a homogeneity analysis, multiple correspondence analysis, and, by imposing *rank-1* constraints on the quantifications of categories (setting the `rank` argument to 1), it can be extended to perform NLPCA [8].

This package comprises a total of 4 functions and 11 example datasets, with the most important function being `homals()`.

The graphic results can be obtained using the `plot.homals()` function [13].

3.3 Comparison Between Packages

The `Gifi` package is derived from the `Homals` package, designed to be an improved version that is more user-friendly and flexible. This is primarily due to a slight difference in the formulation of the loss function and the utilization of B-splines, which provide computational and conceptual advantages to the `Gifi` package [9].

Also, the `princals()` function in the `Gifi` package is specifically tailored for CATPCA, leading to more perceptible and interpretable results in line with the initial objective.

The graphical results obtained by each package exhibit differences. It's noteworthy that while the `Homals` package offers a broader range of plots for result interpretation, the `Gifi` package encompasses the essential plots crucial for understanding the CATPCA results.

Table 1 summarizes the similarities and differences between the packages.

Table 1. Comparison table: `Gifi` vs. `Homals` package

	Gifi	Homals
Version	0.4–0	1.0–10
Last update	October 2022	June 2022
Function	princals()	homals()
Aplication	Developed for CATPCA, descends from the `Homals` package	Developed for MCA, executes CATPCA by applying a *rank-1* restriction
Methodology	• Performs the minimization of the following loss function through the ALS algorithm: $\sigma(\mathbf{X}, \mathbf{H}, \mathbf{A}) = \sum_{j=1}^{m} \mathbf{SQQ}(\mathbf{X} - \mathbf{H}_j\mathbf{A}_j)$ • Uses B-splines • Allows the presence of missing values.	• Performs the minimization of the following loss function through the ALS algorithm: $\sigma(\mathbf{X}, \mathbf{Y}) = \frac{1}{m}\sum_{j=1}^{\mathbf{m}} SQQ(\mathbf{X} - \sum_{\ell \in \mathbf{K}_j} \mathbf{G}_\ell \mathbf{Y}_\ell)$ • Uses binary matrices • Allows the presence of missing values.
Graphic results	• *Scree plot* - all components • *Transformation plot* - all graphs in a figure • *Loadings plot* • *Joint plot* - loadings and category quantifications • *Object plot* • *Biplot* - loadings and object scores	• *Scree plot* - defined number of components • *Transformation plot* - graphs in separated figures • *Loadings plot* • *Joint plot* - category quantifications and object scores • *Object plot* • *Category plot* • *Discrimination plot*
Utilization	• Computationally advantageous • More flexible • Easier to use	• Less intuitive

4 Results and Discussion

To further explore the functionalities and limitations of the `princals()` function from the `Gifi` package, chosen for its suitability in CATPCA, an analysis was conducted on a dataset comprising sensory data collected from wine-tasting trials.

In this study, a sensory dataset was utilized, collected by a research group affiliated with the Center of Biological Engineering at the University of Minho, coordinated by Professor José Oliveira. The dataset compiles information on evaluations of seven distinct wines categorized from A to G, featuring qualitative sensory descriptors determined by a panel of six wine tasters. Following the processing and elimination of non-essential variables, the dataset comprises 126 observations and 25 variables, with five of those variables having missing values. The variables are categorized into Olfactory Assessment, Gustatory Assessment, and Appearance groups, a more detailed description of the variables can be accessed on *Github*.

For the variables "Wine" and "Taster", a nominal level of analysis was chosen since they correspond to the type of wine evaluated and the respective taster, and therefore, the order is irrelevant.

The same analysis was applied to the variables "Clarity", "Color", "Intensity" and "Evolution".

For the remaining variables that have an ordinal qualitative measurement scale, an ordinal level of analysis was defined, as the number of categories does

not exceed the number of observations, and each variable does not have more than 10 categories (7 is the maximum value). Therefore, the decision was made not to apply a spline level of analysis, as this level of analysis tends to be more restrictive, leading to lower VAF values [11].

The nominal variables were automatically encoded into dummy variables by the `princals()` function. For instance, the variable "Wine" initially had 7 categories ("A" to "G") and was transformed into 6 dummy variables labeled "Wine.1', "Wine.2', "Wine.3", "Wine.4", "Wine.5" and "Wine.6". The same approach was applied to the remaining nominal variables.

For the analysis, 7 PCs were specified based on the *Scree plot*, Fig. 1.

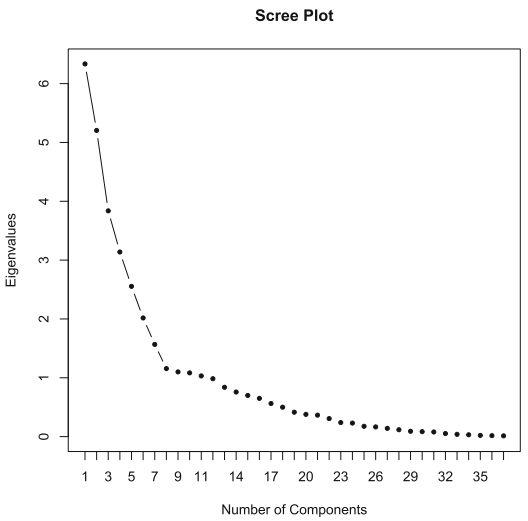

Fig. 1. *Scree plot* obtained through the `plot.princals()` function.

The argument `copies` was set equal to the number of components for nominal variables to ease linear restrictions on this type of variable [12]. Regarding missing values, option `a` was chosen within the `missing` argument of the `princals()` function.

The initial 7 components collectively explained 66.64% of the total variance in the original variables, with the top two components contributing the highest percentage of VAF. The first PC has a percentage of VAF of 17.12%, an eigenvalue of 6.335, and a Cronbach's alpha of 0.877, while the second PC has a VAF of 14.07%, with an eigenvalue of 5.205 and a Cronbach's alpha of 0.842.

It is relevant to emphasize that, since the first two PCs, together, explain the majority of the variance, with a cumulative VAF value of 31.19%, the subsequent plots will only illustrate the first and second dimensions. Therefore, the results will be based on the first two PCs.

To illustrate the impact of choosing a specific quantification level for variables, *Transformation plots* were applied. In this paper, for example purposes, only the plots for the variables "Wine" and "Taster" are represented in Fig. 2.

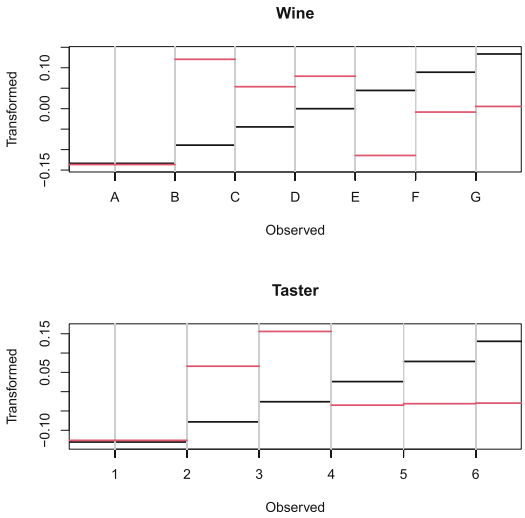

Fig. 2. *Transformation plots* for the "Wine" and "Taster" variables obtained through the `plot.princals()` function.

This graph represents the optimal transformations assigned to the categories of each variable based on the optimal scaling technique and provides insights into the relationships between variables. On the x-axis, the real values of the categories are displayed, and on the y-axis, the corresponding optimal quantifications [10].

The black lines depict the transformation in the first dimension and the red lines in the second dimension. Due to the *rank-1* restriction [12], only the transformations in the first dimension are significant.

The obtained component loadings, along with the VAF(%), eigenvalues, and Cronbach's alpha for each PC, are summarized in Table 2. Component loadings greater than 0.5 are considered significant for the analysis [24] and are highlighted in bold in the table.

Considering the values of the component loadings together with the *Joint plot*, in Fig. 3, which represents in a single plot the component loadings of ordinal variables and the quantifications of the categories of nominal variables, it is possible to establish the most important relationships between variables for the first two components.

The variables "Bitterness", "Carbon_dioxide" and "Dry_fruit" are overlapped, indicating a positive correlation among them. This suggests that wines tend to exhibit these characteristics together. However, they are in opposite

Table 2. Principal components extracted from CATPCA, with their respective component loadings, Cronbach's alphas, eigenvalues, and VAF(%).

	Principal Components						
Variables	1	2	3	4	5	6	7
"Wine.1"	–0.044	–0.020	–0.090	–0.044	0.200	–0.279	0.439
"Wine.2"	0.015	0.098	0.061	0.140	–0.099	0.170	0.356
"Wine.3"	0.032	0.047	0.021	0.021	–0.129	0.186	–0.258
"Wine.4"	–0.025	–0.155	–0.143	–0.102	0.144	–0.425	–0.197
"Wine.5"	0.039	0.047	0.063	0.055	–0.146	0.306	–0.250
"Wine.6"	0.018	0.125	0.120	0.071	-0.101	0.303	0.384
"Taster.1"	0.053	-0.517	0.268	**0.635**	–0.226	–0.255	–0.070
"Taster.2"	**0.831**	0.108	0.360	–0.102	0.233	–0.113	0.003
"Taster.3"	0.093	0.342	0.034	–0.243	**-0.689**	–0.436	–0.039
"Taster.4"	–0.414	**0.529**	0.511	0.264	0.223	–0.210	–0.126
"Taster.5"	–0.231	–0.447	**0.604**	–0.500	–0.055	0.072	0.098
"Citrus_fruit"	**–0.532**	-0.369	–0.398	–0.412	0.125	0.027	0.154
"Tree_fruit"	0.285	–0.131	**0.520**	–0.207	0.112	–0.494	–0.146
"Tropical_fruit"	0.003	0.094	–0.448	**0.523**	0.439	0.109	0.020
"Dry_fruit"	0.173	**0.583**	0.325	0.115	0.376	0.017	–0.141
"Floral"	–0.422	0.096	–0.566	**–0.593**	–0.199	0.120	–0.001
"Fresh_vegetable"	–0.249	**–0.791**	0.394	–0.006	0.142	0.254	0.046
"Dry_vegetable"	0.159	**–0.530**	0.355	0.084	–0.374	0.209	–0.171
"Honeyed"	0.047	–0.399	0.263	0.198	–0.278	0.091	–0.257
"Mineral"	0.185	–0.011	0.014	0.132	–0.352	0.074	0.354
"Carbon_dioxide"	0.185	**0.633**	0.541	–0.034	0.060	0.292	0.014
"Sweetness"	–0.523	**0.647**	0.000	–0.329	0.196	0.077	–0.023
"Acidity"	**–0.768**	0.025	0.467	–0.200	–0.274	–0.100	0.009
"Heat"	–0.361	–0.038	0.418	-0.441	–0.114	–0.182	0.017
"Bitterness"	0.244	**0.805**	0.154	–0.305	–0.189	0.123	0.022
"Astringency"	–0.217	0.199	–0.331	–0.211	–0.255	0.205	–0.056
"Structure"	**–0.891**	0.154	0.102	0.249	–0.019	–0.066	–0.018
"Balance"	**–0.941**	0.062	0.126	0.238	0.040	-0.057	–0.038
"Persistence"	**–0.863**	0.078	0.145	0.362	0.040	–0.049	–0.031
"Clarity.1"	0.291	**0.693**	–0.410	–0.073	–0.052	–0.255	–0.207
"Clarity.2"	–0.006	–0.289	0.059	–0.276	**0.527**	0.058	–0.440
"Color"	**0.569**	0.474	0.443	–0.136	–0.181	–0.173	0.089
"Intensity.1"	**–0.755**	0.108	0.211	–0.162	0.004	0.034	–0.099
"Intensity.2"	–0.225	0.141	–0.157	**0.564**	–0.433	–0.361	0.166
"Color_quality"	0.067	**–0.585**	–0.095	–0.431	–0.123	–0.109	0.093
"Evolution.1"	–0.095	–0.255	–0.466	–0.065	-0.091	-0.435	–0.302
"Evolution.2"	0.075	0.172	–0.150	0.161	**-0.531**	0.408	–0.474
Cronbach's alpha	0.877	0.842	0.770	0.710	0.634	0.525	0.377
Eigenvalue	6.335	5.205	3.838	3.138	2.555	2.017	1.568
VAF(%)	17.12	14.07	10.37	8.48	6.90	5.45	4.24

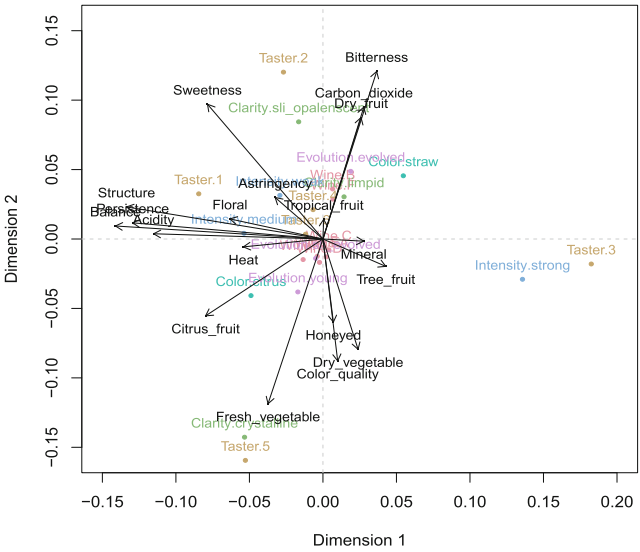

Fig. 3. *Joint plot* for the first two components obtained through the `plot.princals()` function.

positions compared to the variable "Fresh_vegetable" implying a strong negative correlation. In other words, wines with a fresh vegetable olfactory characteristic did not exhibit a bitter taste, nor did they have an aroma of dry fruits and carbon dioxide.

Similarly, the variables "Structure", "Persistence", "Balance" and "Acidity" are positively correlated, as is the group comprising the variables "Honeyed", "Dry_vegetable" and "Color_quality".

Geometrically analyzing the plot it is possible to see that the variables "Structure", "Persistence", "Balance", "Acidity" and "Citrus_fruit" contribute more to the first PC/dimension, as they have an angle $\beta < 45°$ with the axis representing it (x-axis). Also, the variables "Floral", "Heat", "Mineral" and "Tree_fruit" meet this condition; however, their contribution is much smaller as their vectors are shorter and closer to the origin of the plot.

On the other hand, the second PC is more influenced by the variables "Bitterness", "Carbon_dioxide" and "Dry_fruit", with positive values on the second PC/dimension, and by the variables "Honeyed", "Dry_vegetable", "Color_quality" and "Fresh_vegetable", with negative values. The variable "Tropical_fruit" also has an angle $\beta < 45°$ with the y-axis, however, it does not have a significant influence on its formation.

The variable "Sweetness" exhibits an angle $\beta \approx 45°$ with the axes and therefore has a similar influence on both components. The variable "Astringency" has the same characteristic, however, its influence is residual.

Regarding the points of the categories of nominal variables, and since the position of these points reflects their relationship with ordinal variables [11], it was possible to conclude that the "Intensity" and "Color" variables are discriminants in the first PC, the "Clarity" variable discriminates in the second PC, and the "Evolution" variable does not have a significant influence on either of the components. As for the categories "A", "C", "D", "E" and "G" of the "Wine" variable, they are mainly accumulated at the origin and are not associated with any of the components. Wines "B" and "F" are similarly related to the variables "Bitterness", "Dry_fruit", "Carbon_dioxide" and "Sweetness". The tasters can be grouped according to the characteristics they evaluated, as the variables they are close to in the plot are highly influential in their assessments. Thus, tasters "1" and "3" are associated with the first PC, tasters "2" and "5" with the second PC, and tasters "4" and "6" are close, indicating similar evaluations, however, they are located near the origin of the plot, which does not allow for significant conclusions.

Based on the interpretation, the first PC can be defined as "Wine Quality", as many variables contributing to its formation describe fresh characteristics and a persistent and balanced flavor, where color intensity has a significant influence. The second PC is moderately influenced by wines with olfactory characteristics of fruits, dried vegetables, and carbon dioxide, and a sweet taste, and strongly influenced by the aroma of fresh vegetables and a bitter taste. Thus, this PC can be defined as "Wine Aroma". The *Object plot* in Fig. 4, represents the scores of the objects in this analysis, which reveals the existence of three different groups highlighted with circles. Objects from the same group, meaning those closer geometrically, share identical scores on the PCs [18].

The cluster of points in the graph complicates its interpretation, but this division into 3 groups seems to be related to the tasters' rating patterns. The group on the right consists mostly of evaluations made by taster "3" where the choice of the "strong" category predominates for the "Intensity" variable. Meanwhile, the group located in the third quadrant is mostly evaluated by taster "5" and includes evaluations where only the "crystalline" category of the "Clarity" variable was chosen. The larger group brings together the remaining tasters "1", "2", "4" and "6" being abstract regarding the selection of categories and the influence of the evaluated variables, as it encompasses numerous observations.

Additionally, it is possible to relate the object scores to the component loadings through the *Biplot*, shown in Fig. 5.

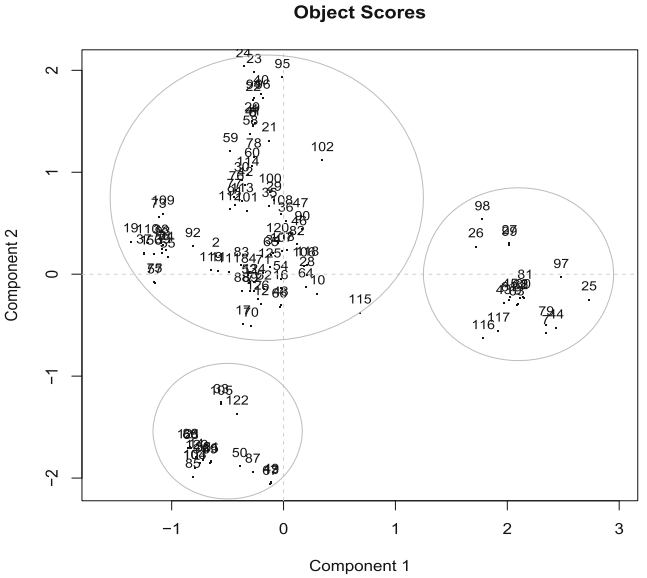

Fig. 4. *Object Plot* for the first two components obtained through the `plot.princals()` function.

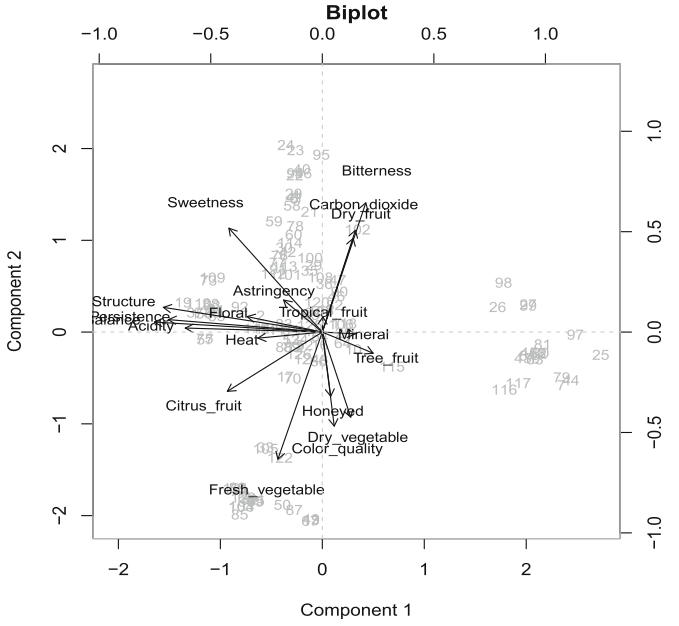

Fig. 5. *Biplot* for the first two components obtained through the `plot.princals()` function.

This plot only allows representing the loadings of ordinal variables, following the vector model. Therefore, it is not possible to relate the object scores to the quantifications of categories for nominal variables, which follow the centroid model.

Nevertheless, this plot reinforces the previously explained observations. The larger group located at the origin and in the second quadrant is overlapped by most loading vectors, explaining the heterogeneous characteristics of this group. The group located in the third quadrant is more influenced by the variable "Fresh_vegetable" and the group of observations on the right does not experience a considered significant influence from any ordinal variable. As observed, the vectors do not reach the area where the scores of the objects in this group are located.

4.1 Shiny: Gify App

To enable users to perform CATPCA without writing a single line of code, a web application in R Shiny was developed [19]. Named ***Gify***, this application provides a more intuitive and simplified way for users to utilize the `princals()` function from the `Gifi` package.

This application can be accessed at https://andreiagomes.shinyapps.io/Gify/ from any internet-enabled device.

It was created to allow users to upload their dataset to the web page, define analysis parameters, execute the CATPCA, and review the results, all processed through selection buttons and input spaces.

A user guide is provided in the application, containing information about the operation of the entire application and links to CRAN support documents.

In the Upload file section, represented in Fig. 6, the user is prompted to upload the data file (.csv or .txt) that he/she plans to analyze. The user is also requested to specify necessary parameters for reading the file, such as indicating whether the file contains a header, if the first column represents row names, and selecting the separator used in the original file ('.', ';', or ',').

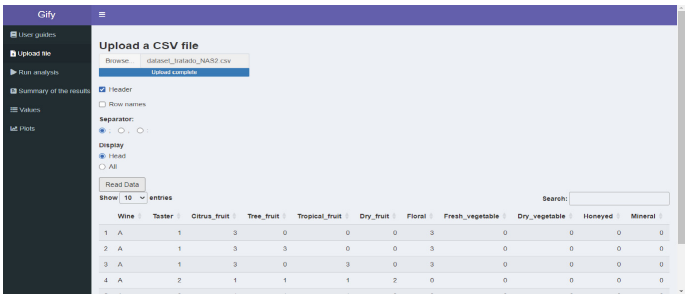

Fig. 6. Uploading the dataset file in the ***Gify*** app.

The Run analysis tab, Fig. 7, allows the user to set the analysis parameters. The parameters in the left column are fixed and refer to the number of dimensions to be computed, handling ties and missing values, whether object scores are z-scores, the maximum number of iterations, and the convergence criteria.

Meanwhile, the parameters in the right column only appear after uploading the data file. In this case, the first question allows the user to choose what type of analysis they want to perform, similar to what happens in the original function.

If the user chooses the "predefined levels" option, Fig. 7, they should, for each variable in the dataset, choose the analysis level ("ordinal", "nominal" or "metric"), spline degree (this argument is maintained in this type of analysis as it does in the original function), the number of copies, and whether the variable is considered active for analysis.

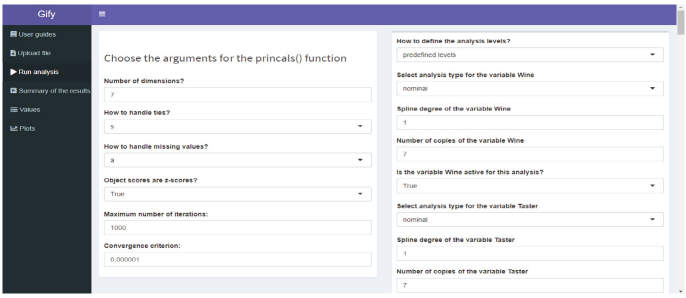

Fig. 7. Specifying the analysis parameters within the ***Gify*** app.

There is also the possibility to configure a spline analysis manually, choosing the "manual splines" option, for that option only the question regarding the selection of the analysis type is replaced by questions specifying whether the variable in question follows an order, the type of internal knots for the variable ("D", "E", "Q", and "R"), and the number of knots for that variable [14]. The remaining questions regarding the spline degree, number of copies, and whether the variable is active persist unchanged.

All initially presented values correspond to the default values of the function.

After defining all parameters, the user will reach the end of the page where the "Run Analysis" button is located, which will automatically redirect the user to the next tab containing a summary of the analysis results.

In the Summary tab, shown in Fig. 8, summaries of the obtained results can be accessed. This includes the function, loss function value, number of iterations, eigenvalues, component loadings (with a cutoff of 0.5), VAF, and cumulative VAF.

At the Values tab, Fig. 9, it is possible to check all the values obtained through the `princals()` function. Among the options are commands such as `transform`, `rhat`, `evals`, `objectscores`, `scoremat`, `quantifications`, `dmeasures`, `lambda`,

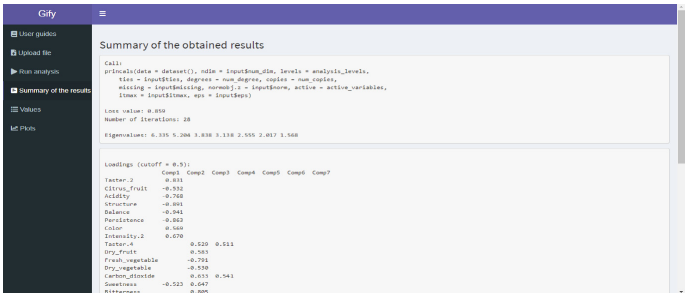

Fig. 8. Summary of the results obtained, displayed in the Summary tab in the ***Gify*** app.

`weights`, `loadings`, `ntel`, `f`, `data`, `datanum`, `ndim`, and `call`, all of these commands and their respective descriptions can be found at [14].

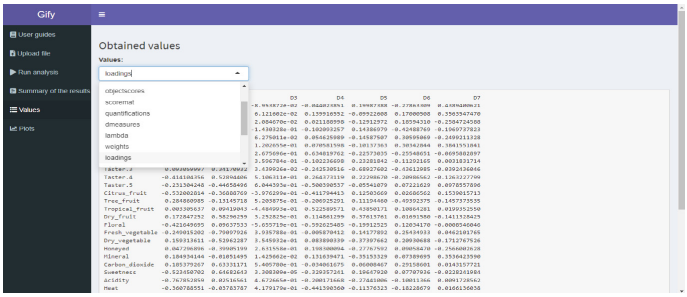

Fig. 9. Overview of the values obtained, displayed in the Values tab in the ***Gify*** app.

Additionally, users can utilize the application for accessing and downloading graphical results such as the *Scree plot, Transformation plot, Loadings plot, Joint plot, Object plot* and *Biplot* generated from the analysis.

5 Conclusion and Future Work

This study aimed to explore the currently available tools in R for applying CATPCA. A comprehensive search identified the `Gifi` and `Homals` packages, with their respective functions `princals()` and `homals()`, as the only ones capable of achieving this objective.

By studying these packages, the `Gifi` package was selected as the most flexible and comprehensive. To investigate the capabilities of the `Gifi` package, an analysis of a wine-tasting dataset was also conducted. This analysis provided valuable insights into the dependency structure of the wine data, which holds promise for future research within the wine industry.

Subsequently, based on the `princals()` function of this package, a web application in R Shiny, named ***Gify***, was developed. ***Gify*** streamlines the execution of CATPCA through a dynamic and simplified interface, making it accessible to users without programming or R software expertise.

In the future, it would be highly interesting to explore the inclusion of other software tools for performing CATPCA, as well as testing additional functionalities of this package. Improving the dynamics of the application and keeping it up-to-date would also be of great interest, given its utility across various fields.

References

1. Abdi, H., Williams, L.: Principal component analysis. Wiley Interdisc. Rev. Comput. Stat. **2**, 433–459 (2010). https://doi.org/10.1002/wics.101
2. Barth, J., Katumullage, D., Yang, C., Cao, J.: Classification of wines using principal component analysis. J. Wine Econ. **16**(1), 56–67 (2021). https://doi.org/10.1017/jwe.2020.35
3. Carreño, E.R., et al.: Study of risk factors for healthcare-associated infections in acute cardiac patients using categorical principal component analysis (catpca). Sci. Rep. **12**(1), 28 (2022). https://doi.org/10.1038/s41598-021-03970-w
4. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer, New York (2002). https://doi.org/10.1007/b98835
5. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Phil. Trans. Roy. Soc. A: Math. Phys. Eng. Sci. **374**(2065), 20150202 (2016). https://doi.org/10.1098/rsta.2015.0202
6. Lavado, N.: Análise em Componentes Principais Não - Linear. Master's thesis, Universidade Nova de Lisboa (2004)
7. Lavado, N.: Extensões via splines da Análise em Componentes Principais. Master's thesis, Universidade Nova de Lisboa (2012)
8. de Leeuw, J., Mair, P.: Gifi methods for optimal scaling in r: the package homals. J. Stat. Softw. **31**(4), 1–21 (2009). https://doi.org/10.18637/jss.v031.i04
9. de Leeuw, J., Mair, P., Groenen, P.: Multivariate Analysis with Optimal Scaling (2016)
10. Linting, M., Kooij, A.V.D.: Nonlinear principal components analysis with catpca: a tutorial. J. Pers. Assess. **94**(1), 12–25 (2012). https://doi.org/10.1080/00223891.2011.627965
11. Linting, M., Meulman, J.J., Groenen, P.J., van der Kooij, A.J.: Nonlinear principal components analysis: introduction and application. Psychol. Methods **12**(3), 336–358 (2007). https://doi.org/10.1037/1082-989X.12.3.336
12. Mair, P.: Modern Psychometrics with R, vol. 62 (2018)
13. Mair, P., de Leeuw, J.: homals: gifi methods for optimal scaling (2022). https://CRAN.R-project.org/package=homals. Accessed 28 Jan 2023
14. Mair, P., de Leeuw, J., Groenen, P.J.F.: Gifi: multivariate analysis with optimal scaling (2022). https://CRAN.R-project.org/package=Gifi. Accessed 28 Jan 2023
15. Manisera, M., van der Kooij, A.J., Dusseldorp, E.: Identifying the component structure of satisfaction scales by nonlinear principal components analysis. Qual. Technol. Quant. Manag. **7**(6), 97–115 (2010). https://doi.org/10.1080/16843703.2010.11673222

16. Meulman, J., Kooij, A., Heiser, W.: Principal Component Analysis with Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data, pp. 49–70. Sage Publications, Thousands Oaks (2004). https://doi.org/10.4135/9781412986311
17. Nguyen, L.H., Holmes, S.: Ten quick tips for effective dimensionality reduction. PLoS Comput. Biol. **15**(6), e1006907 (2019). https://doi.org/10.1371/journal.pcbi.1006907
18. Ribeiro, C.: Estudo de estratégias de valorização económica do Alto Douro Vinhateiro: Aplicação de Técnicas Multivariadas Qualitativas. Master's thesis, Universidade de Trás os Montes e Alto Douro (2015)
19. Satyahadewi, N., Perdana, H.: Web application development for inferential statistics using r shiny. In: Proceedings of the 1st International Conference on Mathematics and Mathematics Education (ICMMEd 2020), pp. 425–429. Atlantis Press (2021). https://doi.org/10.2991/assehr.k.210508.099
20. Saukani, N., Ismail, N.A.: Identifying the components of social capital by categorical principal component analysis (catpca). Soc. Indic. Res. **141**(2), 631–655 (2019). https://doi.org/10.1007/s11205-018-1842-2
21. Sebastião, F.J.: Técnicas multivariadas de redução de dados: comparação, aplicabilidade e convergência de resultados. Ph.D. thesis, UL, Lisboa (2016)
22. Shlens, J.: A tutorial on principal component analysis. Educational **51** (2014). https://doi.org/10.48550/arXiv.1404.1100
23. Todorov, H., Fournier, D., Gerber, S.: Principal components analysis: theory and application to gene expression data analysis. Genom. Comput. Biol. **4**(2), 100041 (2018). https://doi.org/10.18547/gcb.2018.vol4.iss2.e100041
24. Vilela, A., Monteiro, B., Correia, E.: Sensory profile of port wines: categorical principal component analysis, an approach for sensory data treatment. Ciência e Técnica Vitivinícola **30**(1), 1–8 (2015). https://doi.org/10.1051/ctv/20153001001