CIÊNCIA e NATURA

UFSM

Ci. e Nat., Santa Maria v.42, e17,
Special Edition: 40 anos, p. 1-17, 2020
DOI:10.5902/2179460X39892
ISSN 2179-460X

40 years - anniversary

Received: 27/10/19 Accepted: 28/10/19 Published 03/09/20

The usefulness of robust multivariate methods: A case study with the menu items of a fast food restaurant chain

Paulo Canas Rodrigues [I]; Rafael Almeida [II]; Kézia Mustafa [III]

ABSTRACT

Multivariate statistical methods have been playing an important role in statistics and data analysis for a very long time. Nowadays, with the increase in the amounts of data collected every day in many disciplines, and with the raise of data science, machine learning and applied statistics, that role is even more important. Two of the most widely used multivariate statistical methods are cluster analysis and principal component analysis. These, similarly to many other models and algorithms, are adequate when the data satisfies certain assumptions. However, when the distribution of the data is not normal and/or it shows heavy tails and outlying observations, the classic models and algorithms might produce erroneous conclusions. Robust statistical methods such as algorithms for robust cluster analysis and for robust principal component analysis are of great usefulness when analyzing contaminated data with outlying observations. In this paper we consider a data set containing the products available in a fast food restaurant chain together with their respective nutritional information, and discuss the usefulness of robust statistical methods for classification, clustering and data visualization.

Keywords: Multivariate statistics; Data science; Robust principal component analysis; Robust cluster analysis; Data visualization; Multivariate outlier detection

[I] Universidade Federal da Bahia, Salavador, Brazil. paulocanas@gmail.com
[II] Universidade Federal da Bahia, Salavador, Brazil. rafaell.toledo96@gmail.com
[III] Universidade Federal da Bahia, Salavador, Brazil. kzia10@hotmail.com

1 INTRODUCTION

Multivariate statistical methods play a central role in applied statistics, machine learning and data science, across many disciplines, being principal component analysis (PCA) and cluster analysis (CA) two of the most widely used. These multivariate statistical methods, similarly to many other statistical models and algorithms, are appropriated when the data satisfies certain assumptions. However, when the distribution of the data is not normal and/or the data shows heavy tails and/or outlying observations, the classic methods might induce misleading conclusions, and robust statistical methods should be used (Huber and Ronchetti, 2009). Much work has been done to develop univariate and multivariate robust statistical methodologies, which are of great usefulness when analyzing contaminated data with outlying observations.

Principal component analysis is a multivariate technique used for dimensionality reduction and visualization (Jolliffe, 2002). Its basic idea is to describe the variability of a set of initial inter-correlated variables, into a new set of non-correlated variables which are linear combinations of the original variables. These are called principal components, and are obtained sequentially, in a decreasing order of importance, in a way that the first principal component explains the most variability in the original data, the second explains the most variability in the original data that was not retained by the first principal component and is orthogonal to the first component, the third explains the most variability in the original data that was not retained by the first two principal component and is orthogonal to the first two components, and so on. The most immediate aim of PCA is to verify whether it exists a small number of the first principal components responsible to explain a higher proportion of variation of the original data. If this is the case, these few principal components can be used to represent the original data without a big loss of information. This procedure corresponds to dimensionality reduction, which can be used for data visualization and for many other analyses such as regression or cluster analysis. Another objective of PCA is the identification of latent variables (i.e. principal

components) underlying the original structure of the data that might have a physical meaning, allowing, therefore, to visualize the original structure from a different point of view (Rodrigues 2007). More details about PCA and related techniques and generalizations can be found in Jolliffe (2002), Rodrigues (2007) and Johnson and Wichern (2007).

Cluster analysis is a multivariate technique used to group similar individuals or objects based on their similarity with respect to a set of original variables. Similarity is measured by a given distance between objects and the cluster analysis can be obtained using hierarchical or non-hierarchical algorithms. Hierarchical algorithms often result in dendrograms that plot a tree showing the hierarchical relationship between individuals, being the number of clusters chosen based on the dendrograms and on the research interests. The most well-known non-hierarchical clustering algorithm is the k-means that assigns each of the individuals to one of the considered *k* (defined beforehand) clusters. More details about these methods can be found, e.g. in Johnson and Wichern (2007).

The mentioned clustering methods can be seen as unsupervised learning algorithms where similar individuals are grouped, based on their features or properties. On the other hand, classification methods are used in supervised learning where predefined labels are assigned to individuals, based on their features or properties, being the algorithm aimed at correctly predict the category where each individual belongs.

The standard algorithms for principal components analysis and for cluster analysis are sensitive to data contamination with outlying observations and, even a small percentage of outliers can make a large difference on the results. For instance, a few (multivariate) outliers can completely change the direction of the principal components or result in joining clusters artificially, or even to create non-informative clusters composed of only outlying observations (García -Escudero and Gordaliza, 1999; García-Escudero et al. 2010). Therefore, the application of robust methods in the context of PCA and CA is advisable, in order to extract the proper information from the data. Some of these methods are described in the section devoted to Materials and Methods.

In this paper we consider a data set containing the product items available in a fast food restaurant chain together with their respective nutritional information, and discuss the usefulness of robust statistical methods for classification, clustering and data visualization. Being this paper aimed to a more general audience, we decided not to include the statistical formality of the methods, while focusing on their general concept together with key references to methodological results and computational implementations.

This paper is organized as follows. Section 2 includes the description of the data and a general overview of PCA, CA, and their robust counterparts. Section 3 presents the results and discussion, including a general descriptive analysis, the multivariate outlier detection, and the comparison between the results of classic and robust PCA and CA. The paper closes with the conclusion in Section 4.

2 MATERIALS AND METHODS

2.1 Data Description

The data set considered to illustrate our analyses comprises the products available in a fast food restaurant chain together with their respective nutritional information. The data is available in kaggle (https://www.kaggle.com) and includes 260 products with 11 nutritional variables: calories, calories from fat, total fat, saturated fat, trans fat, cholesterol, sodium, carbohydrates, dietary fiber, sugars, and proteins. Besides these numerical variables, there is also available a categorical variable with the category of the food item, which has nine classes: (i) breakfast, (ii) beef & pork, (iii) fish & chicken, (iv) salads, (v) snacks & sides, (vi) desserts, (vii) beverages, (viii) coffee & tea, and (ix) smoothies & shakes.

2.2 Principal Component Analysis

The central idea or PCA is to reduce the dimensionality of the data while retaining as much as possible of the information in the original data, i.e. its variance-covariance structure. This dimensionality reduction is achieved by transforming the

original variables into a new set of variables, the principal components, which are uncorrelated linear combinations of the original variables (Johnson and Wichern, 2007). Although *p* principal components are needed to reproduce the full variability in the original data, a small number of the first *k* principal components can often account for much of that variability. In that case, when most of the variability in the original *p* variables can be retained by the first (few) *k* principal components, the *k* principal components can replace the original *p* variables, resulting in a dimensionality reduction of the original data set. In this case, it is possible to construct biplots (Bradu and Gabriel 1978; Gabriel 1971) and, consequently, to visualize latent variables underlying the original structure. The principal components and their graphical analysis can often help to find relationships and interpretations that were not visible initially. Two of the most widely used algorithms to obtain the principal components are the eigen decomposition and the singular value decomposition of matrices, and its computational implementation can be found, e.g., in the function "prcomp" of the package "stats", in the R software. For more details about principal component analysis, see, for example, Jolliffe (2002) and Johnson and Wichern (2007).

2.3 Cluster Analysis

Cluster analysis aims at grouping similar individuals or objects based on their similarity with respect to a set of original variables. Similarity and dissimilarity measures are used to evaluate the relation/proximity between objects and it is measured by a given distance function that must be defined by taking into account the research problem. Many similarity/distance measures are available in practice such as the standard Euclidean distance (root of sum-of-squares), the Manhattan distance (sum of absolute differences), Gower distance (if some variables are not numeric; Gower, 1971), or Mahalanobis distance (takes into consideration the correlations between variables). There are two main groups of algorithms to obtain clusters: the hierarchical methods and the non-hierarchical methods. The result of the hierarchical clustering methods is, usually, presented in the form of a dendrogram based on an agglomerative or divisive method. Without loss of generality, considering

an agglomerative method, at the beginning each individual belongs to its own cluster that are then sequentially combined in larger clusters, based on the minimum distance/maximum similarity between clusters, until all elements belong to the same cluster. Some of the most well-known methods to conduct agglomerative hierarchical clustering are: (i) single linkage, also known as nearest neighbor clustering, where the distance between two clusters is determined by the pair, one from each cluster, that is closer to each other; (ii) complete linkage where the distance between two clusters is determined by the pair, one from each cluster, that is further away from each other; (iii) average linkage where the distance between two clusters is determined by the average of the distances between all pairs of individuals, one from each cluster; and (iv) Ward minimum variance that finds clusters that maximize the homogeneity within clusters and maximize the heterogeneity between clusters. The most well-known non-hierarchical method is the k-means, where the number of clusters *k* must be decided beforehand. The k-means algorithm clusters each of the *n* individuals into the cluster with the nearest mean/centroid. More details about these methods can be found elsewhere, e.g. in Johnson and Wichern (2007). Computational implementation of hierarchical clustering can be found, e.g., under the function “hclust” of the “stats” package in R, and the k-means algorithm can be found under the function “kmeans” of the same package.

2.4 Robust Principal Component Analysis

Many extensions of PCA, including robust SVD algorithms, have been developed to deal with contaminated data with outlying observations. Examples of robust extensions of the SVD can be found in the literature. Hawkins et al. (2001) used the L1 norm instead of the more usual least squares L2 norm, to compute a robust approximation to the SVD of a rectangular matrix. This method is not robust in the presence of leverage points. Hubert et al. (2005) proposed a robust PCA algorithm that combines projection-pursuit (PP) and robust covariance estimation (minimum covariance determinant; MCD) techniques to compute the robust loadings. Croux et al. (2007) proposed a robust grid algorithm that uses PP to compute PCA estimators,

being the optimization done via the grid search algorithm in the plane instead of the p-dimensional space. Locantore et. al. (1999) proposed robust spherical PCA that uses a spherical principal components procedure. Croux and Ruiz-Gazen (2005) proposed a robust projection pursuit algorithm that uses PP to compute the robust eigenvalues and eigenvectors without going through robust covariance estimation. More details on robust methods for principal component analysis and applications can be found in Maronna (2005), Todorov and Filzmoser (2009), Filzmoser and Todorov (2013), and Rodrigues et al. (2016).

In this paper we consider a robust version of PCA that uses the S-estimator and that is available in the package “mdqc” (Cohen Freue et al., 2007), under the function “prcomp.robust”.

2.5 Robust Cluster Analysis

Several developments have been made to generalize the algorithms for hierarchical and non-hierarchical clustering in order to account for data contaminated with outlying observations. One of the strategies that can be used to remove the influence of outliers when obtaining clusters is to take into account the concept of trimming. Trimmed k-means clustering is an algorithm for robust clustering analysis proposed by Cuesta-Albertos et al. (1997) that outperforms the standard methods when the data includes outlying observations. The trimmed k-means clustering is implemented under the function “tkmeans” of the R package “tclust”, that includes other related robust methods for cluster analysis. More details on robust clustering methods can be found in García-Escudero et al. (2010) and references therein.

3 RESULTS AND DISCUSSION

3.1 Preliminary Descriptive Analysis and Data Visualization

Figure 1 shows the star plot for all 260 food items, considering all 11 variables, divided by category: (i) breakfast (1-42), (ii) beef & pork (43-57), (iii) fish & chicken (58-84), (iv) salads (85-90), (v) snacks & sides (91-102), (vi) desserts (104-110), (vii)

Figure 1 – Star plot for all 260 food items, considering all 11 nutritional variables, each represented in one direction and with one color. The food items are divided by category: (i) breakfast (1-42), (ii) beef & pork (43-57), (iii) fish & chicken (58-84), (iv) salads (85-90), (v) snacks & sides (91-102), (vi) desserts (104-110), (vii) beverages (111-137), (viii) coffee & tea (138-232), and (ix) smoothies & shakes (233-260)

beverages (111-137), (viii) coffee & tea (138-232), and (ix) smoothies & shakes (233-260). In a preliminary analysis, some similarities can be seen within each category. There is some heterogeneity for the category “coffee & tea” as it includes drinks from plain iced tea to large frappé chocolate chip. Similarity can also be seen between the categories beef & pork and fish & chicken for obvious reasons. It is also visible a possible outlier for the food item 83, which represents a large portion of 40 chicken nuggets. The boxplots for each category within each nutritional variable can be found in the Figures S1-S11 in the Supplementary Material.

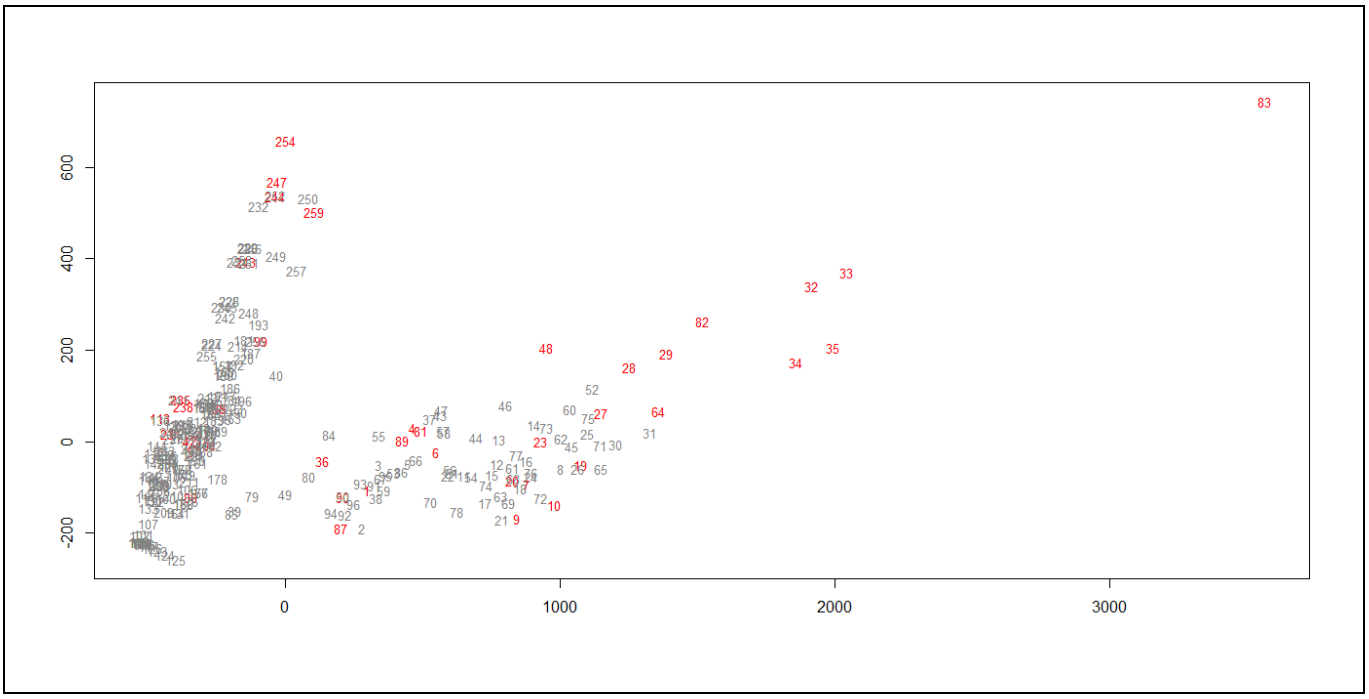
This preliminary analysis motivates the hypothesis that the data might contain some outlying observations and that robust statistical methods might be more appropriated for the analysis than their classical counterparts.

3.2 Multivariate Outlier Detection

In this subsection we present the study of the hypothesis that the data might contain some outlying observations. We consider the methods proposed by Filzmoser et al. (2005) and their implementation in the R package “mvoutlier” for multivariate outlier detection. In particular, we used the function “aq.plot” to obtain the plot depicted in Figure 2. This plot shows the adjusted quantile plot showing the outliers detected by the 97.5% quantile of the chi-square distribution 10 degrees of freedom (the number of variables after removing “trans fat” because of its high proportion of zeros), using 99% of the observations for the minimum covariance determinant (MCD) estimations, and a maximum thresholding proportion of 0.05. The multivariate outliers detected, shown in red in Figure 2, and with the same codes as in Figure 1, are: 1, 4, 6, 9, 10, 19, 20, 27, 28, 29, 32, 33, 34, 35, 36, 41, 42, 48, 81, 82, 83, 87, 88, 89, 90, 97, 98, 99, 104, 113, 235, 238, 244, 254, 259. The formal multivariate outlier detection shown in Figure 2 endorses the hypothesis that the data includes outlying observations and confirms the specific outliers that show a visual discrepancy in Figure 1.

In the following sections we make a comparison between classic and robust methods to access the influence of these multivariate outliers.

Figure 2 – Adjusted quantile plot showing the outliers detected by the 97.5% quantile of the chi-square distribution 10 degrees of freedom, using 99% of the observations for the MCD estimations, and a maximum thresholding proportion of 0.05. The grey numbers represent the non-outlier observations and the red numbers represent the outlier observations



3.3 Principal Component Analysis: Classic vs. Robust Methods

Figure 3 shows the biplot of the first two principal components for the 260 food items and the 11 nutritional variables, considering the correlation matrix, representing a total of 80.04% of the variability in the original data. We can see that categories “Coffee & Tea” and “Smoothies & Shakes” show a similar behavior, being “Beverages” quite close. Also the categories “Chicken & Fish” and “Beef & Pork” seem to have some similarities, with one individual in “Chicken & Fish” (the same as identified in Figure 1) showing extreme behavior (right-bottom corner of Figure 3). The biplot of the first two principal components for the 260 food items and the 11 nutritional variables, considering the covariance matrix, accounts for a total of 98.57% of the original information (Figure S12), being the variables “Calories” and “Sodium” the most important for that graphical representation because of their variability and magnitude.

Figure 3 – Biplof of the first two principal components for the 260 food items (colored dots accordingly to the restaurant chain category) and the 11 nutritional variables, considering the correlation matrix, representing a total of 80.04% of the original information

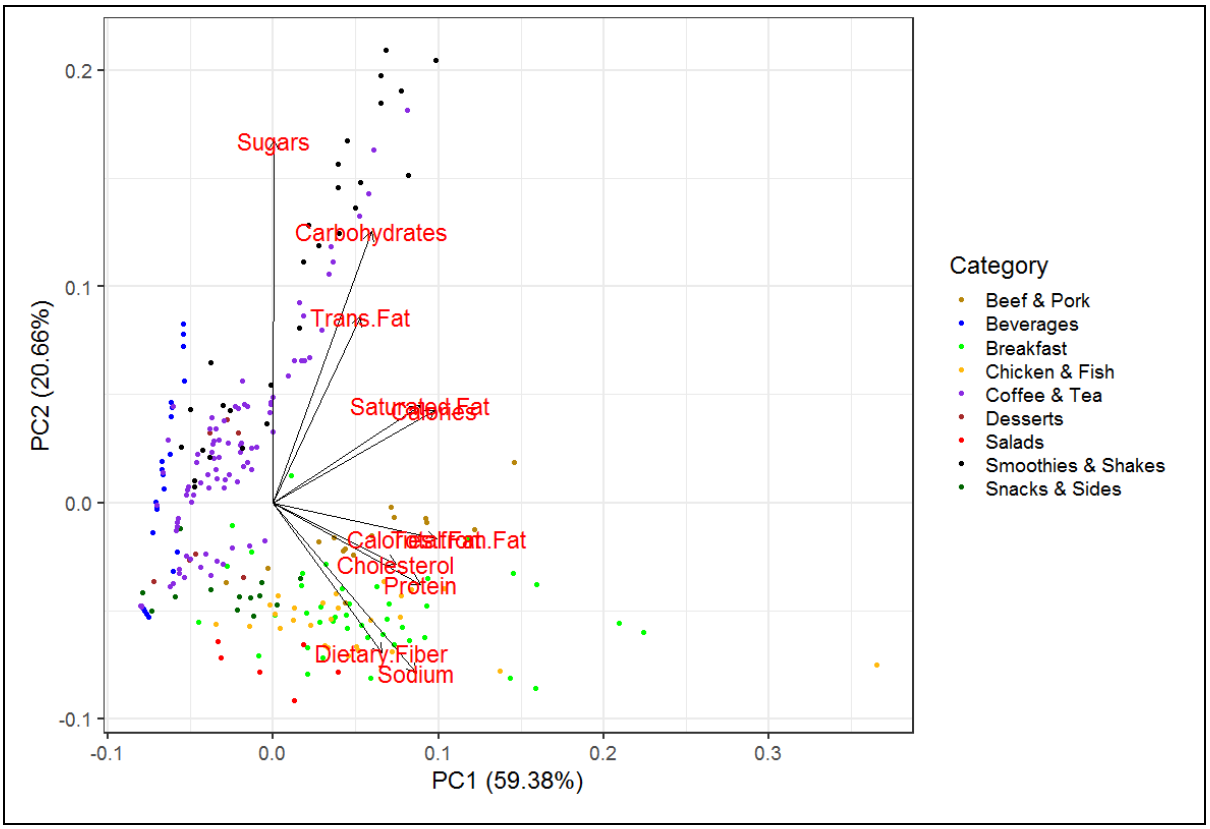
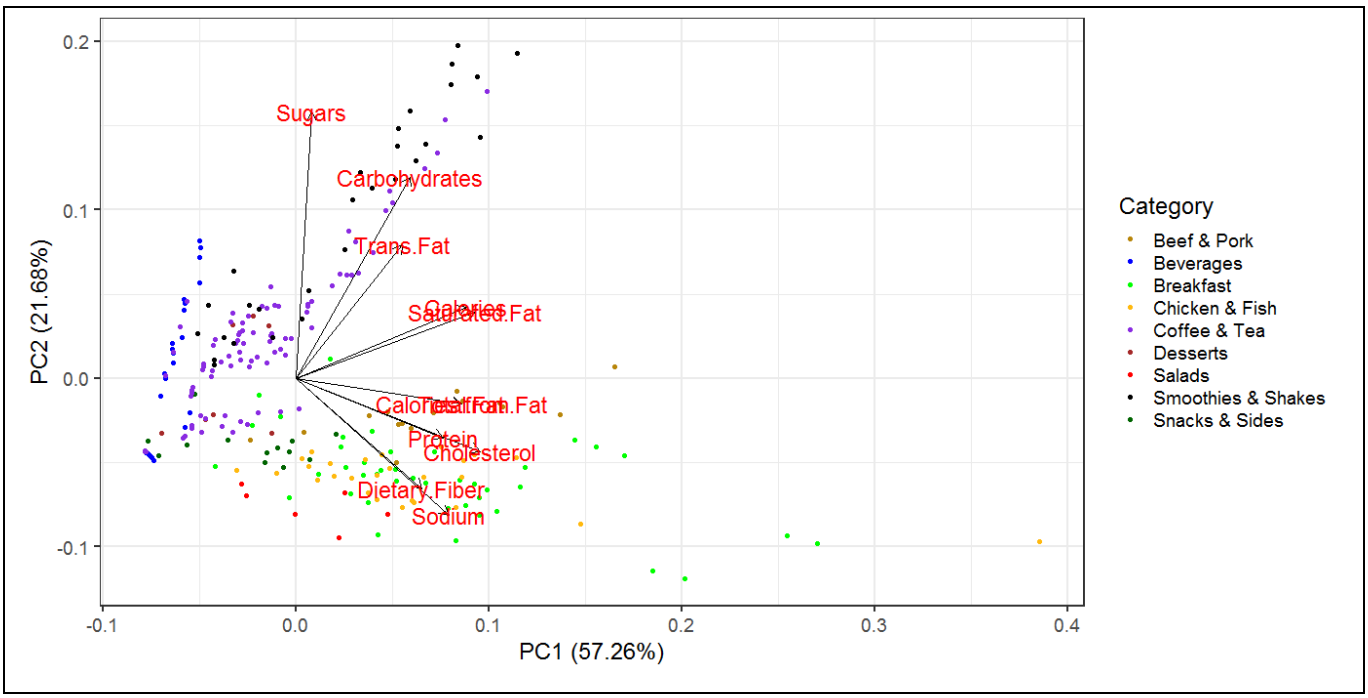


Figure 4 shows the biplot for the first two robust principal components for the 260 food items and the 11 nutritional variables, considering the correlation matrix, representing a total of 78.94% of the original information in the data. The differences for the graphical representation in Figure 3 are minor, giving the idea that, although several outliers were detected in Section 3.2, they do not have a strong role in the model with two principal components. We should also keep in mind that the principal components are ordered by decreasing importance (the first is the most important and the last is the least important), and that the first principal components are usually associated to the signal in the data and the last principal components to the noise in the data. However, since we have identified a number of multivariate outliers, the graphical representation of Figure 4 is more reliable that the one in Figure 3, as it is associated to a robust method.

Figure 4 – Biplof of the first two robust principal components for the 260 food items (colored dots accordingly to the restaurant chain category) and the 11 nutritional variables, considering the correlation matrix, representing a total of 78.94% of the original information



3.4 Clustering analysis: Classic vs. Robust Methods

3.4.1 Hierarchical clustering

As a preliminary overall cluster analysis of the data, Figure 6 shows the heat map for the bi-clustering of food items and nutritional variables with a dendrogram for the individuals and another for the variables, considering the Euclidean distance. The nine colors below the dendrogram for the food items are associated to the categories attributed by the fast food restaurant chain. This gives us a general overview of the similarity between individuals and between variables, which has (obvious) nutritional interpretation specially for the variables.

Figure 5 – Heat map for the bi-clustering of food items and nutritional variables with a dendrogram for the individuals and another for the variables, considering the Euclidean distance. The nine colors below the dendrogram for the food items are associated to the categories attributed by the fast food restaurant chain

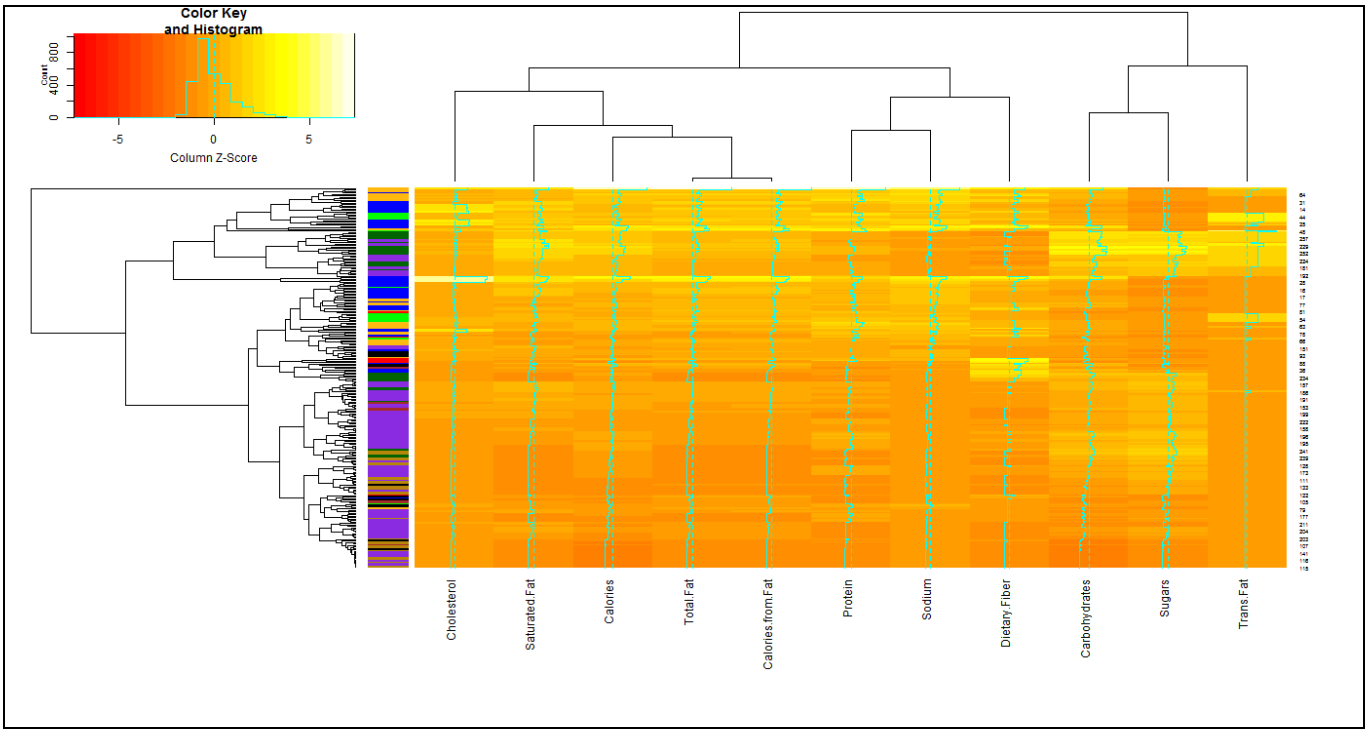
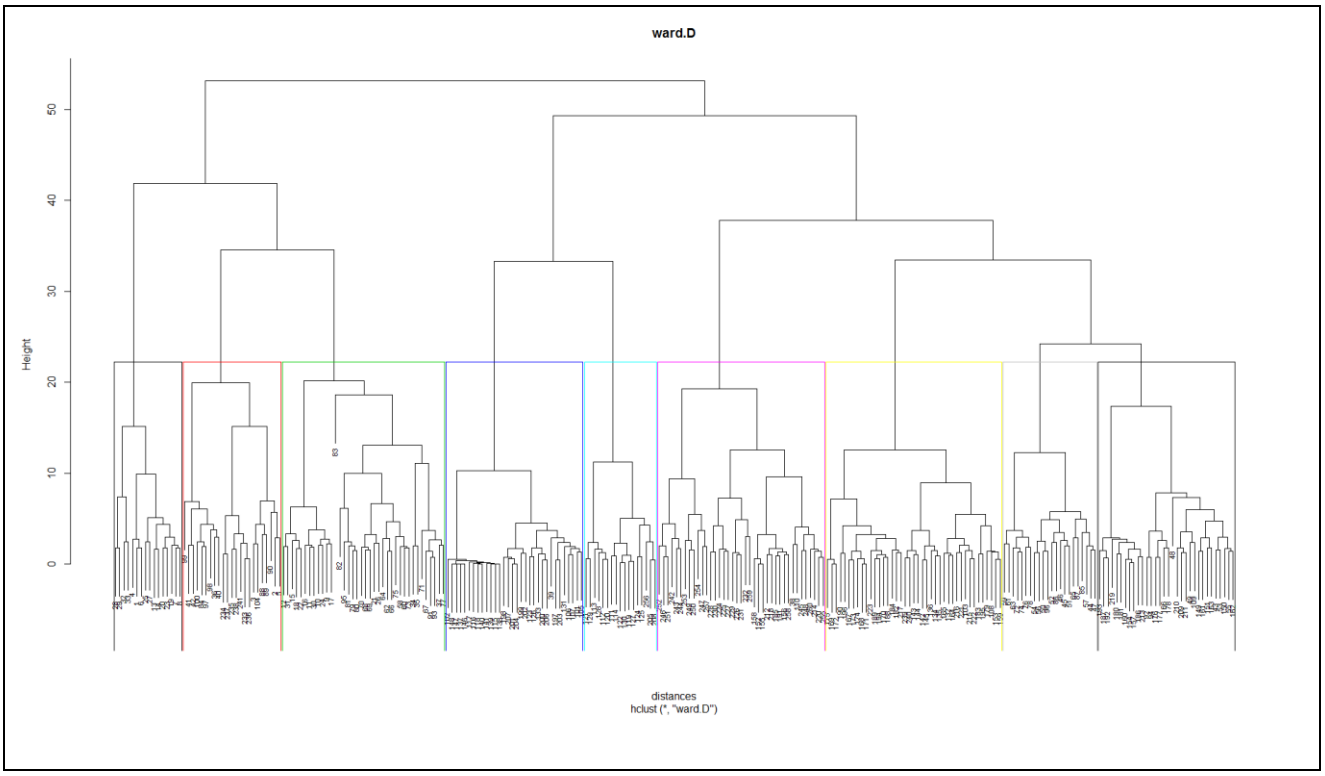


Table 1 shows the classification error rate when comparing the food category given by the restaurant chain and the results obtained for hierarchical clustering. Here we consider the three distance measures: Euclidean, Gower and Mahalanobis; and five linked methods: single, complete, average and two versions of the Ward. The best classification error rate was obtained when the Mahalanobis distance and the Ward.D method were used. Its dendrogram can be found in Figure 6.

Table 1 – Classification error rate between the food category given by the restaurant chain and the results obtained for hierarchical clustering obtained for the methods listed in the first column and for the distance measures listed in the first row

	Euclidean	**Gower**	**Mahalanobis**
single	69.103	73.297	73.876
complete	27.048	27.080	35.522
average	28.164	33.935	67.241
ward.D	24.036	22.278	19.943
ward.D2	23.228	22.521	24.963

Figure 6 – Dendrogram for the individual food items considering the Mahalanobis distance and the Ward.D method

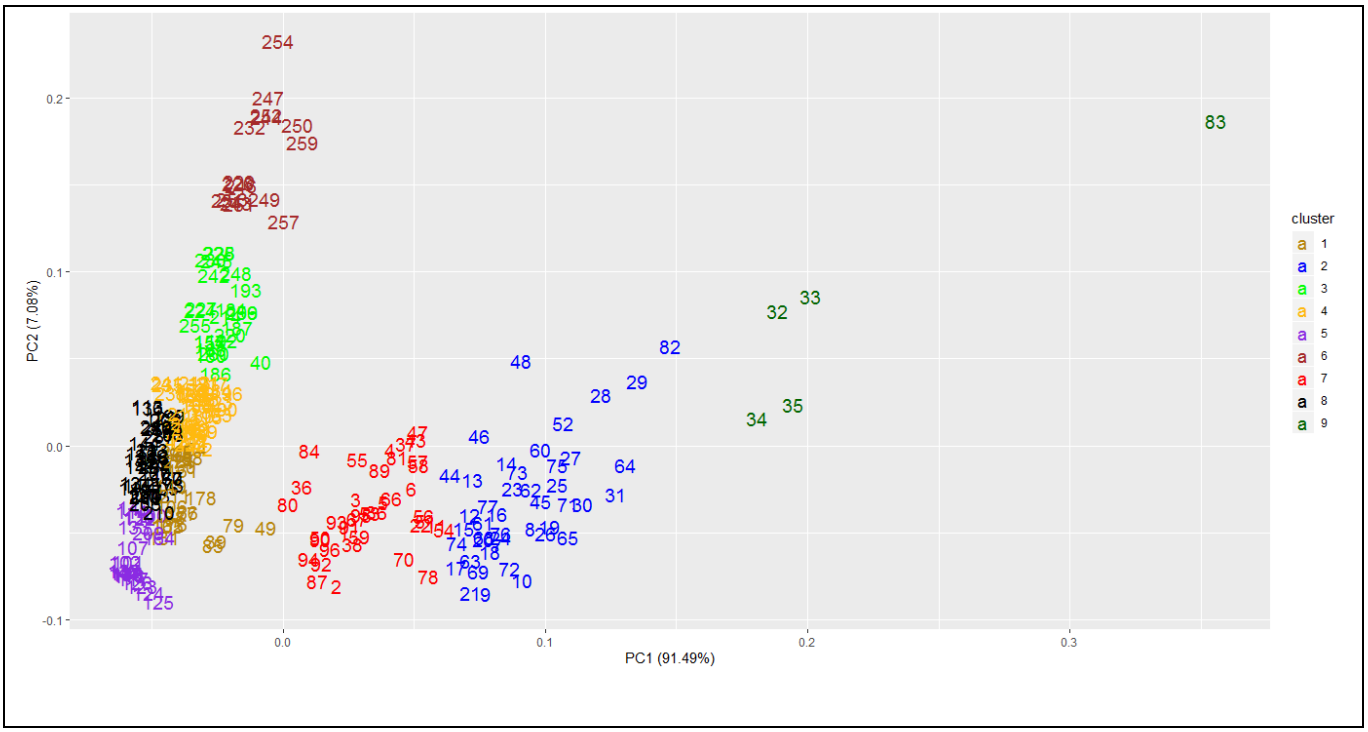


3.4.2 Non-hierarchical clustering

Figure 7 shows the output of the k-means cluster analysis with two components, representing about 98.60% of the original variability, for the 260 food items considering nine clusters. The clusters show some differences from the original

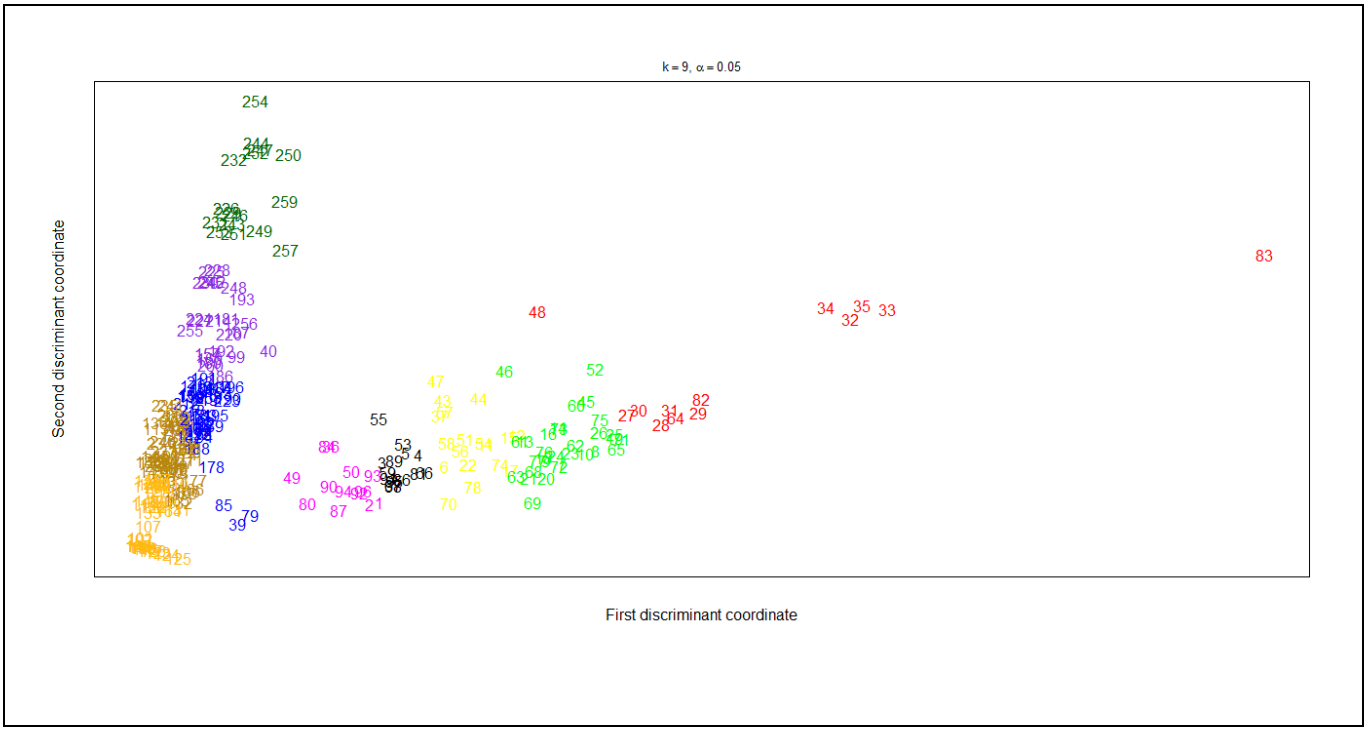
categories given by the fast food restaurant chain as the clusters are obtained exclusively by the nutritional information about the food items. The classification error rate between the nine clusters obtained by the classical k-means and the fast food restaurant chain categories is 22.9%.

Figure 7 – Classical k-means cluster analysis with two components for the 260 food items considering nine clusters, being the code number the same as defined in Figure 1



The output from the trimmed k-means clustering (Cuesta-Albertos et al. 1997) can be found in Figure 8. The red codes on the right hand side of Figure 8 are 5% food items to be trimmed, being each of the other colors associated to each of the nine clusters. The classification error rate between the nine clusters obtained by the trimmed k-means and the fast food restaurant chain categories was 21.6%, a slight improvement from the classical k-means.

Figure 8 – Trimmed k-means cluster analysis with two components for the 260 food items considering nine clusters and 5% of the observations to be trimmed, being the code number the same as defined in Figure 1



4 CONCLUSION

This paper discussed the usefulness of using robust statistical methods such as robust principal component analysis and robust cluster analysis when the data is contaminated with outlying observations. The methods under consideration were applied to a data set containing the products available in a fast food restaurant chain together with their respective nutritional information.

Visualization tools were used to have a general overview of the multivariate data, and multivariate outlier detection techniques were used before conducting the comparisons between the classic and robust versions of principal component analysis and cluster analysis. Slightly different patterns were visible for classic and robust methods when doing clustering. When the comparisons were made in terms of classification, considering the food category in the fast food restaurant chain as the benchmark, there was a slight improvement in terms of classification error rate for

the robust non-hierarchical clustering. However, we should bear in mind that we are clustering and classifying food items based on nutritional information, which might have a great overlap between food categories (e.g. breakfast vs. smoothies and shakes).

Overall, it is of great importance to make a proper preliminary analysis of the data and it is recommended to consider robust statistical methods when the data is contaminated with outlying observations.

REFERENCES

CUESTA-ALBERTOS, J.A.; GORDALIZA, A.; MATRÁN, C. "Trimmed k-means: an attempt to robustify quantizers". Annals of Statistics. 1997;25:553-576.

COHEN FREUE, G.V.; HOLLANDER, Z.; SHEN, E.; ZAMAR, R.H.; BALSHAW, R.; SCHERER, A.; MCMANUS, B.; KEOWN, P.; MCMASTER, W.R.; NG, R.T. MDQC: A New Quality Assessment Method for Microarrays Based on Quality Control Reports. Bioinformatics. 2007;23:3162 – 3169.

CROUX, C.; RUIZ-GAZEN, A. High breakdown estimators for principal components: The projection-pursuit approach revisited. Journal of Multivariate Analysis. 2005;95:206–226.

CROUX, C.; FILZMOSER, P.; OLIVEIRA, M. Algorithms for Projection-Pursuit Robust Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems. 2007;87:218–225.

FILZMOSER, P.; GARRETT, R.G.; REIMANN, C. Multivariate outlier detection in exploration geochemistry. Computers and Geosciences. 2005;31:579-587.

FILZMOSER, P.; TODOROV, V. Robust tools for the imperfect world. Information Sciences. 2013;245:4–20.

GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. Biometrika. 1971;58:453–467

GARCÍA-ESCUDERO, L.A.; GORDALIZA, A. Robustness Properties of k-Means and Trimmed k-Means. Journal of the American Statistical Association. 1999;94:956–969

GARCÍA-ESCUDERO, L.A.; GORDALIZA, A.; MATRÁN, C.; MAYO-ISCAR, A. A Review of Robust Clustering Methods. Advances in Data Analysis and Classification. 2010;4:89–109.

GOWER, J.C. A general coefficient of similarity and some of its properties. Biometrics. 1971;27:857–874.

HAWKINS, D.M.; LIU, L.; YOUNG, S. Robust Singular Value Decomposition, National Institute of Statistical Sciences. Technical Report Number. 2001;122.

HUBER P.J.; RONCHETTI E.M. Robust Statistics. 2nd ed. USA: Wiley; 2009.

HUBERT, M.; ROUSSEEUW, P.J.; BRANDEN, K.V. Robpca: a new approach to robust principal component analysis. Technometrics. 2005;47:64–79.

JOHNSON R.A. and WICHERN D.W. Applied Multivariate Statistical Analysis. 6th ed. USA: Pearson; 2007.

JOLLIFFE, I.T. Principal component analysis. New York: Springer; 2002.

LOCANTORE, N.; MARRON, J.S.; SIMPSON, D.G.; TRIPOLI, N.; ZHANG, J.T.; COHEN, K.L. Robust principal components for functional data. Test. 1999;8:1–28

MARONNA. R. Principal components and orthogonal regression based on robust scales. Technometrics. 2005;47:264–273.

RODRIGUES, P.C.; MONTEIRO, A.; LOURENÇO, V.M. A Robust additive main effects and multiplicative interaction model for the analysis of genotype-by-environment data. Bioinformatics. 2016;32:58–66.

RODRIGUES, P.C. Componentes Principais: o método e suas generalizações (Principal Components: the method and its generalizations). In Lisbon, Portugal [dissertation]. Lisbon: Instituto Superior Técnico, Technical University of Lisbon; 2007.

TODOROV, V.; FILZMOSER, P. An Object Oriented Framework for Robust Multivariate Analysis. Journal of Statistical Software. 2009;32:1–47.