CDPCA: 10 years after

Adelaide Freitas
Departamento de Matemática, Universidade de Aveiro, Portugal, e Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), Aveiro, Portugal, *adelaide@ua.pt*

Keywords: Clustering; Sparse principal components analysis

Abstract: Clustering and Disjoint Principal Component Analysis (CDPCA) is a constrained principal component analysis for multivariate numerical data. The main goal is to detect clusters of objects and, simultaneously, to find a partitioning of variables such that the between cluster deviance in the reduced space of such partition is maximized. The partition formed by a disjoint set of the original variables identifies the groups of variables belonging to the CDPCA components. Recently, this methodology has been implemented in a R-function called `CDpca`. In this work, we review some theoretical issues of the CDPCA model and present two applications on real data sets using the R-function `CDpca`.

1 Introduction

In order to extract information of multivariate data, some authors apply a sequential procedure as follow: first apply Principal Component Analysis (PCA), in order to reduce the dimensionality of the data by taking the resultant score matrix associated to the first few principal components, and then proceed to the reduction of the objects in order to get homogeneous groups by applying a clustering method on that reduced score matrix. However, this reduced score matrix may mask the clustering structure of the original data [1]. To overcome this drawback, a constrained principal component analysis for multivariate numerical data, called Clustering and Disjoint Principal Component Analysis (CDPCA), was proposed ten years ago, in 2009, by Vichi and Saporta [2]. CDPCA is aimed to detect

Table 1: Distribution of citations of [2] found in SCOPUS.

	CDPCA cited in papers as			Papers with
year	Method	Application	only mention	pdf not available
2009-2011	-	-	1	2
2012	-	1	1	-
2013-2014	-	-	6	4
2015	2	-	-	2
2016	2	-	3	1
2017	2	-	1	1
2018	1	-	1	3
2019	1	-	2	-

clusters of objects and, simultaneously, to find a partitioning of variables such that the between cluster deviance in the reduced space of such partition is maximized. Concretly, the main goal of CDPCA is to providing a nonoverlapping clustering of homogeneous objects on a reduced set of sparse CDPCA components such that the set of the object centroids presents maximum variance in the reduced space defined by the components. From the point of view of practical applications, how useful has this methodology been? A quick search by Scopus (December/2019), we found 37 citations of [2]. Nevertheless, among the 24 publications made available, eight mention CDPCA in Method sections and a single paper present applications of the CDPCA methodology (Table 1) but without indication of any software used.

Recently, we have computationally developed a function in the open-source software `R` [3] to apply CDPCA on standardized data, namely the function `CDpca` available in the package biplotbootGUI [4].

The paper is organized as follows. In Section 2 a brief overview of the CDPCA model is presented. In Section 3, a description of the implemented `R`-function `CDpca` is provided. In the last section, two applications on real data sets are illustrated.

2 The CDPCA model

The CDPCA methodology is aimed at providing a description of any numerical (previously standardized) data matrix $\mathbf{X}$, with I objects and J variables, in terms of a set of P ($P < I$) object centroids, which are obtained from applying a clustering method (k-means) on the objects of the matrix $\mathbf{X}$, and a set of Q ($Q < J$) principal components resultant from PCA applied on a centroid matrix which is obtained from $\mathbf{X}$ when each original object is replaced by its cluster centroid. Concretely, given a $(I \times J)$ multivariate data matrix $\mathbf{X}$, the CDPCA model describes $\mathbf{X}$ in the following manner:

$$\begin{aligned} \mathbf{X} &= \mathbf{U}\bar{\mathbf{X}} + \mathbf{E}_1 && \text{(k-means on } \mathbf{X}) \\ &= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T + \mathbf{E}_1 + \mathbf{E}_2 && \text{(PCA on } \mathbf{U}\bar{\mathbf{X}}) \\ &= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T + \mathbf{E} && (1) \end{aligned}$$

where $\mathbf{U}$ is a $(I \times P)$ binary and row stochastic matrix storing the assignment of the I objects into the P clusters, $\mathbf{A}$ is a $(J \times Q)$ columnwise orthonormal matrix (*i.e.*, $\mathbf{A}^T\mathbf{A} = \mathbf{I}$) that represents the component (unit-)loading (i.e., the coefficients of the linear combinations of the original variables) matrix having only one nonzero element by row which assigns the component (column) for each variable identified by row, $\bar{\mathbf{X}}$ is the $(P \times J)$ object centroid matrix in the original space, $\bar{\mathbf{Y}} := \bar{\mathbf{X}}\mathbf{A}$ is the $(P \times Q)$ object centroid matrix in the reduced space of the components and $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$ with $\mathbf{E}_1$, $\mathbf{E}_2$ the $(I \times J)$ error matrices arising from k-means and PCA, respectively. In this sense, the structure of the matrix $\mathbf{X}$ defined by model (1) is visually described by "blocks" as illustrated in a toy example depicted in Figure 1.

The parameters $\mathbf{U}$, $\bar{\mathbf{Y}}$ and $\mathbf{A}$ of the CDPCA model can be estimated by minimizing the error associated to the model. Nevertheless, such optimization problem is quite difficult to solve [2]. Hence, an alternating least-squares (ALS) algorithm have been proposed in [2] for the parameter estimation. Later, the ALS algorithm was described in [5] in terms of two basic steps: the assignment of objects via

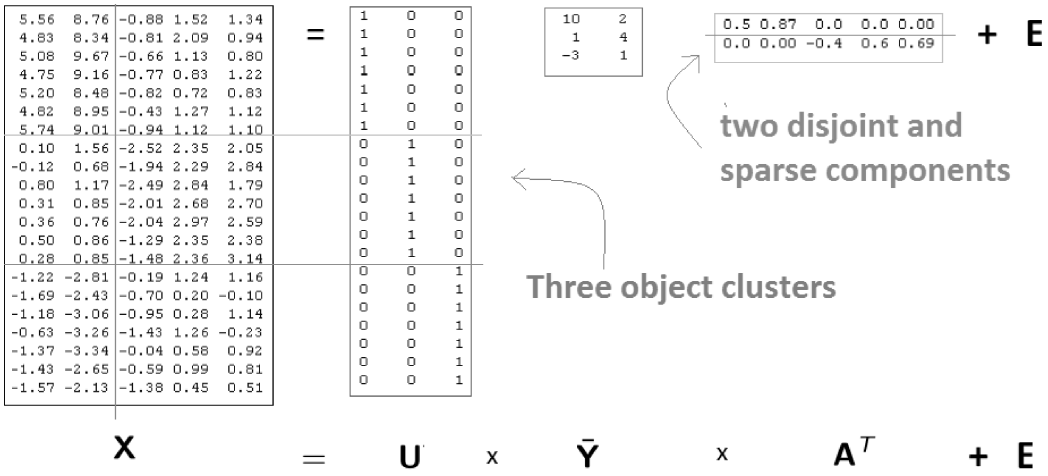

Figure 1: An illustration of the CDPCA model.

k-means, and the reduction of the attribute space via application of PCA to the resulting centroids. This is the algorithm implemented in the R-function `CDpca`. This algorithm is a heuristic procedure which starts from an initialization step (random generation of the matrices **U** and **V**) and uses iterative schemes in the estimation of the matrices **U**, **V** and **A**. In order to increase the chance of finding the global optimal solution, and reducing the sensitivity of the ALS algorithm on the initial matrices **U** and **V**, CDPCA's authors have been recommending to run the algorithm at least 30 times, for different initial assignment matrices **U** and **V** randomly chosen at the beginning of each run. Each run executes the two steps iteratively until a tolerance condition has been achieved [5].

3 CDPCA in R

CDPCA, with the ALS algorithm, is available in a R-package called `biplotbootGUI` [4] given by the `CDpca` function. Recently, the R-function `CDpca` was evaluated on high-dimensional data [6]. To execute `CDpca`, some parameters must be introduced as input. For

instance, for `lymphoma` data set (library "`spls`") with $P = 3$ clusters of objects and $Q = 2$ clusters of variables, the following instruction can be considered:

```
> data(lymphoma)
> data <- lymphoma$x
> class <- lymphoma$y
> Q = 2 # desired number of subsets of variables
> P = 3 # desired number of clusters of objects
> tol = 10^(-5) # accuracy; tolerance criterium
> maxit = 1000 # maximum of iterations of each run
> r = 20 # number of runs of the ALS algorithm
```

and, finally,

```
> CDpca (data, class, P, Q, tol= 10^(-5), maxit,
        + r, cdpcaplot=TRUE)
```

The input `cdpcaplot=TRUE` is useful where the true object clusters are known and the purpose is to compare them with the object clustering provided by the CDPCA methodology. It is worth to mention that the `CDpca` function starts by standardizing the input `data`.

4 Applications

Basically, CDPCA is aimed at reducing both objects and variables, which is very important to make easier the interpretation of data. For illustration of applications of the CDPCA methodology, two real multivariate data sets, where the object clustering is known, are analyzed: *MoraviaWine* (where $I > J$; the data is available in [7]) and *Lymphoma* (where $I << J$; the data is available in the R-package `spls` [8]).

MoraviaWine data set (30 objects × 8 variables)

For this data set, the purpose is to explore the presence of grouping structures in 30 commercially available wines from South Moravia taking into account eight phenolic acids (vanillic, gentisic, protocatechuic, syringic, gallic, coumaric, ferulic and caffeic) measured by gas chromatography mass spectrometry. In these 30 wines we have: 9 white wine, 1 rosé and 20 red wine. In order to visualize the data in a 2-dimensional graph, we propose the partitioning of the variables in 2 groups (i.e., we define 2 disjoint components). Thus, we applied CDPCA with $P = 3$ and $Q = 2$, taking 500 runs of the ALS algorithm, each run with a maximum of 1000 iterations, and such that the difference of the objetive function yielded between two consecutive iterations is less than 10^5, as follows:

```
> res=CDpca(data=wine, class=c(rep(1,9),2,rep(3,20)),
   + P=3,Q=2,tol= 10^(-5),maxit=1000,r=500,cdpcaplot=T)
```

The R-object `res$A` provides the component loading matrix reported in Table 2. The first (second) component explains 32.7% (29.5%, resp.) of the total variance and it is more characterized by the variables: vanillic, syringic and gallic acids (coumaric and ferulic acids, resp.). Applying PCA, the two first (no correlated) components are almost the same of the disjoint components obtained from CDPCA methodology both in terms of variance explained and of variables most contributing to specify the components; however, the reduced space defined by the two first disjoint components which are correlated (Pearson's coefficient = 0.153, observed from the correlation matrix obtained from the R-object `var(res$Y)`) exhibits more homogeneous clusters since the correspondent between cluster deviance (bcd) is 73.5% of total deviance (obtained from R-object `res$bcdev`), while the bcd obtained when the k-means is applied on the reduced space of the first two principal components of PCA is less than 50% of the total deviance. This result on the bcd shows the benefits of applying CDPCA instead of applying k-means on the score matrix resultant of PCA. The plot of the 30 wines in the first

Table 2: Component loadings for PCA and CDPCA

	PCA		**CDPCA**	
Phenolic acid	PC1	PC2	PC1	PC2
vanillic	0.55	-0.04	0.53	0
gentisic	-0.14	0.52	0	0.35
protocatechuic	0.30	0.06	0.33	0
syringic	0.51	0.10	0.52	0
gallic	0.47	-0.07	0.57	0
coumaric	0.20	0.43	0	0.56
ferulic	-0.04	0.60	0	0.59
caffeic	0.29	0.41	0	0.47
% Var	35.3	29.3	32.7	29.5

two disjoint components plan is shown in Figure 2 (on the right). Based on the interpretations of that components, we can conclude that vanillic, syringic and gallic acids can be identified as presumed better markers in red wines since they occur in relatively higher concentrations. While coumaric and ferulic acids do not exhibit distinguishable features among white, rosé and red wines, they suggest that these features can reflect presumable particular wines (containing higher concentrations of coumaric and ferulic acids).

Lymphoma data set (62 objects × 4026 variables)

For this gene expression data set which contains $J = 4026$ genes and $I = 62$ samples extracted from $P = 3$ types of cancer (dimension of each type-group: 42/9/11), the purpose is to fit the standardized data to a CDPCA model. We considered 2 disjoint components. The model parameters are then estimated using the following instruction:

```
> CDpca(data=lymphoma\$x, class=lymphoma\$y,
  + P=3,Q=2,tol= 10^(-5),maxit=1000,r=20,cdpcaplot=F)
```

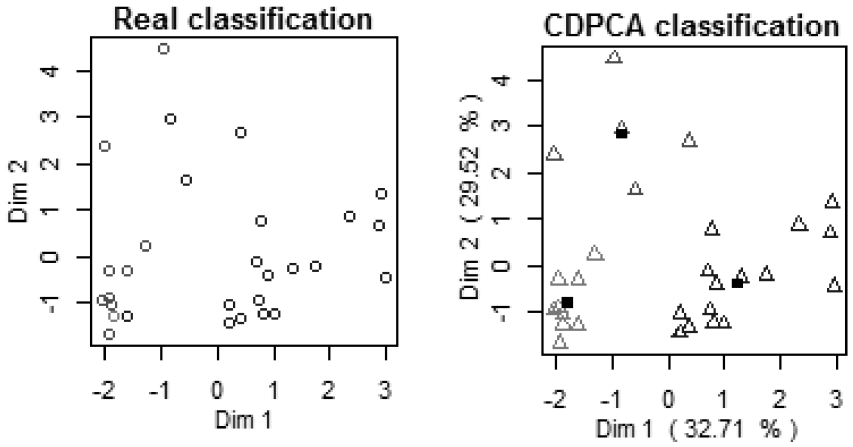

Figure 2: On the left, CDPCA with three color showing the real classification (red: red wine; green: rosé; blue: "red wine"), and on the right CDPCA with the object clustering described by three different colors (centroids represented in black) on the MoraviaWine data set.

Due to high computational effort involved in the execution of the ALS algorithm using the CDpca function when the data is described by a high number of variables, only 20 runs of that algorithm were considered in the heuristic procedure. The bdc related to the final estimated CDPCA model was about 88%. The first two disjoint components explain almost the same (about 12%-13% each one).

Since this data set contains a high number of variables, we empirically examined the variability of the results related to the variables when the `CDpca` function is executed several times on *Lymphoma* data. We observed small variability in terms of the proportion of the variance explained by the components and their correlation. However, the position of nonzero elements into the component loading matrix was exhibiting many changes among executions.

A study to analyze how much the variables that belong to each disjoint component can be differently assigned, we applied CDPCA on *Lymphoma* data by executing $R = 30$ times the instruction above for

$Q = 2, 3, 4$. The summarized results are reported in Figure 3. The proportions of variance explained by the two first disjoint components and the correlation values between them show low variability.

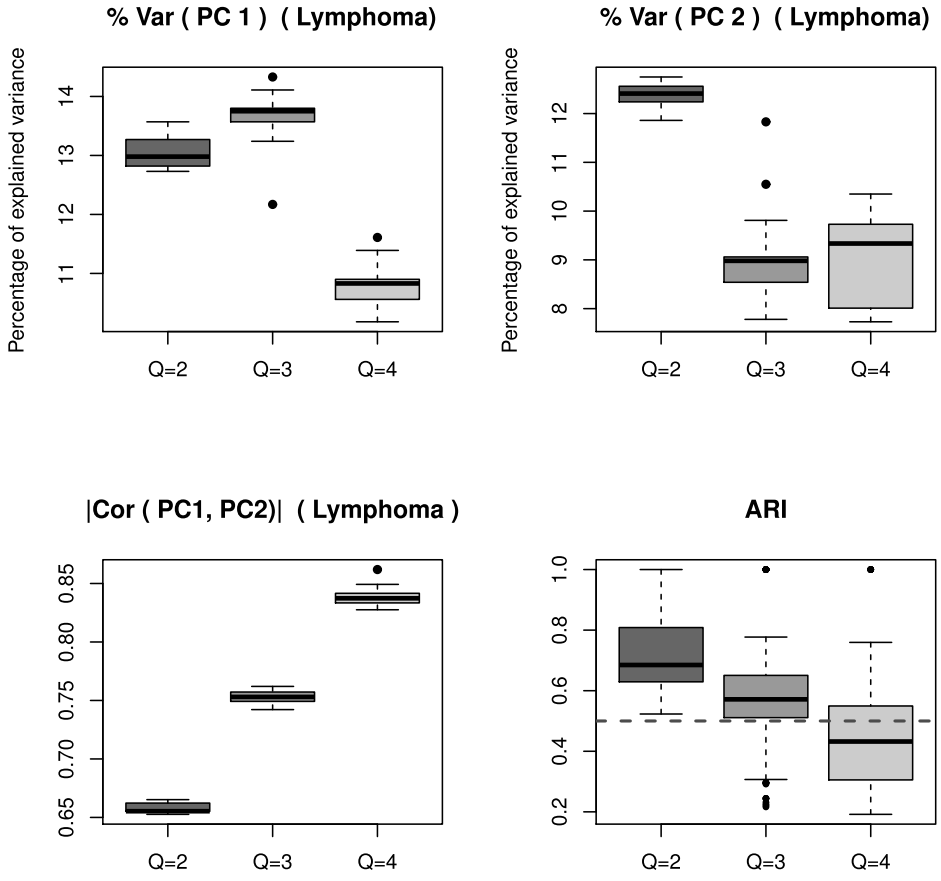

Figure 3: Boxplots of several features of CDPCA results.

Using the Adjusted Rand Index (ARI), we analyze the similarity between any two estimated partitions of variables among all the combination pairs of two applications that we can get from the $R = 30$ executions of CDPCA. The values of ARI are displayed in Figure 3. In many cases, the ARI index exhibited better results for $Q = 2$.

This behaviour of the ARI for comparing variable clusterings among multiple executions of CDPCA, may be consider as a criterium to select the number of clusters of variables for *Lymphoma* data, and it suggests to group the $J = 4026$ genes in two big clusters. This is an idea to explore in future work.

Acknowledgements

This work was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA, University of Aveiro) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), reference UIDB/04106/2020.

References

[1] DeSarbo, W.S., Jedidi, K., Cool, K., Schendel, D. (1990). Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters* 2, 129–146.

[2] Vichi, M., Saporta, G. (2009). Clustering and Disjoint Principal Component Analysis. *Computational Statistics & Data Analysis*, 53, 3194–3208.

[3] R Development Core Team. (2019). R: A Language and Environment for Statistical Computing (http://www.R-project.org/).

[4] Nieto-Librero, A.B., Galindo-Villardon, M.P., Freitas, A. (2019). biplotbootGUI: Bootstrap on Classical Biplots and Clustering Disjoint Biplot. R package version 1.2. `https://CRAN.R-project.org/package=biplotbootGUI`.

[5] Macedo, E., Freitas, A. (2015). The Alternating Least-Squares Algorithm for CDPCA. In Plakhov, A. et al (eds.): Optimiza-

tion in the Natural Sciences, *Communications in Computer and Information Science*, Springer Verlag, 499, 173–191.

[6] Freitas, A., Macedo, E., Vichi, M. An empirical comparison of two approaches for CDPCA in high-dimensional data. *Statistical Methods & Applications* (24 pages, to be published).

[7] Hron, K., Jelínková, M., Filzmoser, P., Kreuzigerd, R., Bednář, R., Barták, P. (2012). Statistical Analysis of Wines Using a Robust Compositional Biplot. *Talanta*, 90, 46–50.

[8] Chung, D., Chun, H., Keles, S. (2019). spls: Sparse Partial Least Squares (SPLS) Regression and Classification. R package version 2.2-3. `https://CRAN.R-project.org/package=spls`.