



UNIVERSITAT DE BARCELONA

Dimensionality reduction in multigroup data: applications in integrative omics

Carolina Andrea Millapán Toledo

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

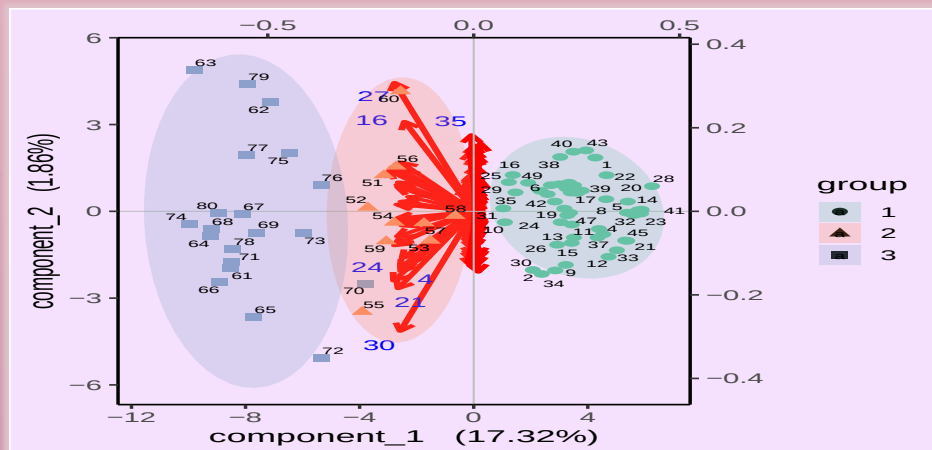
WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.



UNIVERSITAT DE
BARCELONA

Dimensionality reduction in multigroup data: applications in integrative omics.

Thesis presented by Carolina Andrea Millapán Toledo
for the Doctorate degree from the University of Barcelona



Doctorate in Biomedicine

Department of Genetics, Microbiology and Statistics

Faculty of Biology

Universitat de Barcelona

Barcelona, 27 of july 2023



UNIVERSITAT_{DE}
BARCELONA

Dimensionality reduction in multigroup data: applications in integrative omics.

Thesis presented by Carolina Andrea Millapán Toledo
for the Doctorate degree from the University of Barcelona

Doctorate in Biomedicine
Department of Genetics, Microbiology and Statistics
Faculty of Biology
Universitat de Barcelona
Barcelona, 27 of july 2023

A handwritten signature in black ink, appearing to read 'Carolina' followed by a stylized surname.

Doctoral student:
Carolina Andrea
Millapán Toledo

Director:
Ph.D. Ferran
Reverter Comes

Director:
Ph.D. Esteban
Vegas Lozano

Tutor:
Ph.D. Alex Sánchez
Pla

*To my children, Montserrat and Joel
Andrés, and my husband Joel.*

Acknowledgements

I greatly appreciate the patience and empathy of my directors, Esteban and Ferran, for guiding me, teaching me, and giving me wisdom in this research that was carried out for my doctoral thesis. I am also grateful to my directors for encouraging me in times of difficulty and helping me during the time it took me to finish my research. Without a doubt, they are great people and teachers.

I would like to thank Dr. Josep Maria Oller from the Department of Genetics, Microbiology and Statistics of the University of Barcelona for his help in formulating our multigroup PCA compactly.

I also want to thank the Department of Statistics, University of Concepcion. Through Professor Bernardo Lagos, I was able to use the server to analyze high-dimensional omic data. Moreover, my gratitude to Professor Katia Sáez, since she motivated me to continue to perfect myself in the area of statistics and is responsible for my fondness of this world of randomness.

Besides, my gratitude to my friend Jay for his help and collaboration in revising the English of my thesis. Thanks for your kindness.

Many thanks to family, especially my in-laws, Luisa and Román, who always helped me take care of my children so that I could have time to advance in my thesis research. Their great support was very fundamental to achieve this goal.

Also, many thanks to my parents, Magaly and Andrés, and grandparents, Tata Carlos and Mami Teo, even though they are not physically here with me, I know they are taking care of me wherever they are. My parents always made a great effort to help me with my undergraduate studies, despite the economic difficulties. Thank you very much, dear parents and grandparents.

Finally, I would like to thank with all my heart my dear family that I have formed together with my husband Joel. We have been together for 21 years through thick and thin and he is my fundamental support along with my children, Montserrat and Joel Andrés. They are the great treasures of my life and my strength that carries me towards achieving my goals.

Abstract

The idea of this research is to propose dimensionality reduction methods considering multigroups in a dataset. In multivariate analysis, there are many multigroup methods, but they have different objectives than the goals presented in this thesis.

Classical principal component analysis (PCA) is considered in this research as it has something similar to our objective which is the exploration and visualization of the dataset. However, this unsupervised method lacks in considering the multigroup configuration.

The thesis presents two multivariate dimension reduction approaches under a multigroup configuration. Statistical simulation helps us better observe and control the parameters of interest with these new methods proposed in this research. Thus, in this way, it helps us to conclude how they contribute to the literature of multivariate techniques in the visualization and exploration of high-dimensional data analysis.

The method, multigroup principal component analysis (mgPCA), is based on maximizing the interdistances between pairs of observations when the observations belong to different groups.

The second method, multigroup dimension reduction (MDR), determines linear varieties that minimize overlap by comparing observations in one group with the rest of the observations in the other groups.

It is worth mentioning that a package was created in the R statistical programming language called `MultiGroup0` containing our two dimensionality reduction methods, and with vignettes, for better explanation and visualization of our multivariate multigroup approaches on omics datasets or any data analysis.

Resumen

La idea de esta investigación es proponer métodos de reducción de dimensionalidad considerando los multigrupos en un conjunto de datos. El análisis multivariado hay muchos métodos multigrupos, pero que tienen distintos objetivos a lo nuestro en esta tesis.

El análisis de componentes principales (PCA) clásico se considera en esta investigación, ya que tiene algo similar en nuestro objetivo que es la exploración y visualización del conjunto de datos. No obstante, este método no supervisado carece en considerar la configuración multigrupo.

La tesis presenta dos enfoques multivariantes de reducción de la dimensión bajo una configuración multigrupo. La simulación estadística nos ayuda a observar y controlar de mejor manera los parámetros de interés en estos nuevos métodos propuestos en esta investigación. Así de esta manera, nos ayuda a concluir de qué manera contribuyen en la literatura de técnicas multivariantes en la visualización y exploración del análisis de datos de alta dimensionalidad.

El método, análisis de componentes principales multigrupo (mgPCA), se basa en la maximización de las interdistancias entre pares de observaciones cuando las observaciones pertenecen a distintos grupos.

El segundo método, reducción de dimensión multigrupo (MDR), determina variedades lineales que minimizan la superposición al comparar las observaciones de un grupo con el resto de observaciones en los otros grupos.

Cabe mencionar que, se creó un paquete en el programa R llamado **MultiGroup0** que contiene nuestros dos métodos de reducción de dimensionalidad, y con viñetas, para una mejor explicación y visualización de nuestros enfoques multivariados multigrupos en conjuntos de datos ómicos o de cualquier análisis de datos.

Notation

Chapter 3

PCA	: Principal Component Analysis.
FCPCA	: Flury's Common Principal Component Analysis.
BGC	: Between Group Comparison.
DGPA	: Dual Generalized Procrustes Analysis.
STATIS	: "Structuration de Tableaux A Trois Indices de la Statistique".
ACT	: "Analyse Conjointe de Tableaux".
DSTATIS	: Dual STATIS.
MGPCA	: Multigroup Principal Component Analysis.
NIPALS	: Nonlinear Iterative Partial Least Squares.
RGCCA	: Regularized Generalized Canonical Correlation Analysis.

Chapter 4

mgPCA	: Multigroup Principal Component Analysis.
MDR	: Multigroup Dimensionality Reduction.
SDE	: Single Dimension Exploration.
E	: m -dimensional Euclidean space on real $\mathbb{R}$.
$\mathbf{b}_1, \dots, \mathbf{b}_m$	: Basis vectors of E .
$\mathbf{A}$	: Matrix associated with the scalar product of E .
$\mathcal{L}$	: Linear r -dimensional manifold, $r < m$.
$\mathbf{V}_1, \dots, \mathbf{V}_r$	: Basis vectors orthonormal of $\mathcal{L}$.
F	: SubSpace generated by the basis vectors of $\mathcal{L}$: $\langle \mathbf{v}_1, \dots, \mathbf{v}_r \rangle$.
$\mathbf{p}_1, \dots, \mathbf{p}_k$	: Elements of E grouped in q size groups $n_1, \dots, n_q$.
$d_{\mathcal{L}}^2$	: Euclidean distance induced in $\mathcal{L}$.
$\mathbf{e}_{\alpha}$	: Vector of zeros except in the position α that has the value 1.
$\mathbf{w}_1, \dots, \mathbf{w}_r$	: Eigenvector of higher r -eigenvalues.

Contents

Acknowledgements	i
Abstract	i
Resumen	iii
Notation	v
List of Figures	ix
List of Tables	xiii
Chapter 1. Introduction	1
1.1. Background and Motivation	1
1.2. Simulation	3
1.3. Omics Real Datasets	3
1.4. Organization of the Present Study	4
Chapter 2. Goals	7
Chapter 3. Existing Methods of Analysis of Multigroup Data	9
3.1. Flury's Common Principal Component Analysis	11
3.2. Multigroup Principal Components Analysis	13
3.3. DUAL STATIS	14
3.4. DUAL Generalized Procrustes Analysis	17
3.5. Between Groups Comparison	18
3.6. Multi-group PCA retrieved	20
3.7. Comparison of Methods	22
3.8. Regularized Generalized Canonical Correlation Analysis	25
Chapter 4. Main Contribution to Multivariate Methods	31
4.1. Multigroup Principal Component Analysis	31
4.2. Multigroup Dimensionality Reduction	39
Chapter 5. MultiGroupO Package	47

5.1. R Package	47
Chapter 6. Simulation Study and Omic Real Data	57
6.1. The Data Simulation Function	58
6.2. Real Dataset	62
Chapter 7. Results	65
7.1. Simulation Scenarios	65
7.2. Real Dataset	83
Chapter 8. Discussion and Conclusion	89
8.1. Discussion	89
8.2. Conclusion	91
References	93

List of Figures

1	Graphical display of the common vector of loadings ($\mathbf{a}^{(1)}$), the group vectors of loadings ($\mathbf{a}_1^{(1)}, \dots, \mathbf{a}_M^{(1)}$) and the group components ($\mathbf{t}_1^{(1)}, \dots, \mathbf{t}_M^{(1)}$) by Eslami et al.	9
2	Flury Common Principal Component Analysis on the iris dataset	12
3	Multigroup Principal Components Analysis on the iris dataset	14
4	DUAL STATIS on the iris dataset	17
5	DUAL Generalized Procrustes Analysis on the iris dataset	19
6	Between Groups Comparison on the iris dataset	21
1	Structure of the MultiGroup0 package: the implemented functions of the methods proposed in this thesis.	48
2	Output by method of the function <code>pca()</code> using SRBCT dataset in plsgenomics package.	49
3	Output by method of the function <code>mgPCA()</code> using SRBCT dataset in plsgenomics package.	50
4	Output by method of the function <code>MDR()</code> using SRBCT dataset in plsgenomics package.	51
1	PCA with 30 observations and correlation (0.5,0.2,0.1) for each group, respectively; in the upper and lower triangular. Different variances for each group (3,3,5). The variables with signal are 100 and the variable noise are 10. Graph Biplot PCA.	60
2	Multigroup PCA with 30 observations and correlation (0.5,0.2,0.1) for each group, respectively; in the upper and lower triangular. Different variances for each group (3,3,5). The variables with signal are 100 and the variable noise are 10. Graph Biplot Multigroup PCA.	60
3	MDR with 30 observations and correlation (0.5,0.2,0.1) for each group, respectively; in the upper and lower triangular. Different variances for each group (3,3,5). The variables with signal are 100 and the variable noise are 10. Graph display MDR.	62
4	Overview of the progressive advance in the methods to study genes, transcripts and proteins in the computer science [1].	63

1	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	66
2	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	67
3	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	67
4	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	68
5	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	68
6	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	69
7	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	70
8	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	70
9	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	70
10	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	71
11	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	72
12	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	72
13	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	73
14	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	73
15	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	74
16	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	75
17	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	75
18	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	75
19	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	76
20	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	77
21	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	77
22	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	78
23	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	78
24	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	79
25	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	80
26	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	80
27	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	80
28	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	81
29	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	81
30	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	82
31	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	82
32	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	83

33	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	83
34	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	85
35	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	85
36	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	86
37	Graphs under the function <code>pca()</code> standard by <code>Multigroup0</code> package.	86
38	Graphs under the function <code>mgPCA()</code> by <code>Multigroup0</code> package.	87
39	Graphs under the function <code>MDR()</code> by <code>Multigroup0</code> package.	87

List of Tables

3.1 Methods of analysis of multigroup data.	10
3.2 R packages for multivariate analysis	11
5.1 Arguments of the simulation function of quantitative multigroup data under a multivariate normal distribution.	51
7.1 Relationship between a multivariate normal distribution and arguments of the simulation function <code>fun.sim()</code> .	84

Chapter 1

Introduction

The science of statistics has been a fundamental tool in recent decades in various areas, especially in Biomedicine, because with computational advances, biostatistics has been able to develop better algorithms to process complex data in the shortest possible time.

Study in the area of biostatistics and bioinformatics and understanding how the methodologies can be improved, or created, using statistical techniques pose a great challenge.

The realization of this thesis is motivated to advance in statistical methodologies, mainly in multivariate analysis, specifically, in dimensionality reduction.

1.1. Background and Motivation

Multivariate analysis is a topic of statistics that today is used in many areas such as: neuroscience [2, 3], biology [4, 5], molecular biology [6, 7, 8, 9, 10], medicine [11, 12], astronomy [13], atmospheric science [4], social networks [14] and others, because in this last decade there is a lot of information storage built-up which is necessary to analyze. The multivariate analysis encompasses an important set of techniques that helps to control multicollineality and irrelevant variables and missing data, reduce the high dimensionality, and capture the experimental variation.

All of these areas agree that they have a lot of information, which leads to a problem of high dimensionality in these datasets. One of the pioneering techniques in multivariate analysis is the Principal Component Analysis (PCA). The earliest literature on PCA dates from Pearson [15] and Hotelling [16]. Its idea is simple to reduce the dimensionality of the data and at the same time preserve as much variability (i.e, statistical information) as possible. This means that preserving as much variability as possible translates into finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other. Finding such new variables, the principal components (PCs), reduces to solving an eigenvalue/eigenvector problem [4].

It is quite common to study the same set of variables between individuals who are "a priori" divided into several groups. This type of data is usually found in Biology, especially in Biomedicine. For example, studies of gene expression microarrays (or any other omic data) for individuals with different experimental conditions. A very simple case is a control group vs. experimental group [17]. We refer to this configuration as a multigroup dataset.

On the other hand, the variables can also be divided into groups according to their origin or type, an example in bioinformatics are the increasingly numerous and different types of omic data that can be obtained with high performance techniques: gene expression, metabolites and proteins to name a few examples. In these cases it is said that we have multiblock schemes.

Principal Component Analysis (PCA) is one of the dimension reduction techniques most commonly used in Biomedicine to have a global view (multivariate) of the data studied. It can see if there are outliers, groupings, project the variables and detect patterns in the dataset (correlations, dependencies, etc.). Therefore, it is a basic tool to visualize the individuals and the variables. However, it is not intended to incorporate the information that the data are divided into groups, that is, the case of multigroup data. If the PCA is made with this type of data, it treats it as if were not multigroup. In addition, it does not consider the case of having variables with different types of sources of variation (e.g. genomics, phenomics and metabolomics), that is, of multiblock types. Some quotes about these problems are presented by the authors Elasmı et al 2014 [18] and Abdi et al 2013 [19].

Several authors have contributed in the literature to solve the PCA problem to incorporate the group structure and/or different types of variables. Some of them have analyzed datasets with some structure of groups in the same observations (see [20, 21, 22, 23, 24]). This technique is first introduced by Krzanowski, the so called Multigroup Principal Component Analysis and Flury in 1984, the so called Common Principal Component Analysis. Other methods under this same idea are the following: Simultaneous Component Analysis with invariant pattern [22], Dual Generalized Procrustes Analysis [25, 23] and Dual Multiple Factor Analysis [26, 19]. In addition, using an extension to multiblock analysis, some techniques such as: Constrained Principal Component Analysis [27], Joint Multidimensional Statistical Framework [28] and recently an extension of the MGPCA technique called Principal Components Analysis Multigroup and Multiblock (mbmgPCA).

This thesis presents two new methods of multigroup data analysis based on linear transformations. The first technique called multigroup PCA (mgPCA) this multivariate exploration approach that has the idea of considering the structure of groups. This framework contributes to the multivariate analysis literature, in particular, multigroup PCA is that our approach is closer to the original approach of the PCA and we do not make assumptions

about the structure of the covariance matrices of the groups. We propose the search for subspaces of representation of the data that maximize the preservation of the sum of Euclidean interdistances but we include a caveat, unlike PCA, only interdistances between observations belonging to different groups are considered.

The second proposed multivariate technique called Multigroup Dimensionality Reduction (MDR) is another multivariate exploration method that is based on projections indexes. The idea is to minimize the overlap of the projections of the observations of one group with the centroid that results from projecting the observations of another group.

On the other hand, a exploration method is also proposed called Single Dimension Exploration (SDE), which would be a modification of the MDR. This approach performs for each group we will look for a vector in sample space that minimizes the overlapping of the projected points of that group and the projected points in the rest of groups. In such a way that, for each one of the groups we obtain a centroid of the direction. We not ask that the set of directions define an orthogonal axis system that allow us to represent the data in reduced dimension. Instead, we will analyze the data in each of the dimensions separately.

1.2. Simulation

An important part of the research carried out in the thesis is based on simulation. The simulation of the dataset for quantitative variables (called matrix $\mathbf{X}$), under a multivariate normal distribution, establishes a group vector with a different sample size, a mean vector for each group, and variance and covariance matrices for each group. It should be mentioned that the mean vector is set according to a certain pattern to observe the coupling or separation between the groups. On the other hand, the matrix of variances and covariances under goes a transformation to define it based on the variances and correlations established in each group.

In addition, data with biological signal variables and variables without biological signal are simulated, which are called noise variables. These last variables are under a multivariate normal distribution with zero mean vector and correlation matrix with constant upper and lower triangular equal to zero and constant variances. More details of the data simulation are in chapter 6.

1.3. Omics Real Datasets

This investigation uses two real omics data to present our new dimensionality reduction methods. The first set of omic data is the gene expression data from Khan et al. (2001) is childhood genetic expressions of cancer diseases that classified into four types of tumors. The

small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS)[29].

The second set of real omic data is an research on acute myeloid leukemia. This global gene expression has three groups where two groups have one type of leukemia and the other group are normal individuals. The example analyzed is a simplified version of a study published in journal, PNAS (January 30, 2001 vol. 98 (3)) “Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics.”

More details can be obtained in the article available at the link : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC14719/?tool=pubmed>. The study was carried out with 48 samples that were hybridized with arrays of affymetrix of the hu6800 type and of which, for this example, a balanced subset of 15 samples, 5 of each type.

The chapter 6 and 7 describes in more detail the actual data or applications that can be used when using these multivariate analysis tools in the area of biomedicine.

1.4. Organization of the Present Study

In chapter 2, the specific objectives of this thesis. For each objective, the idea to be developed in this framework is described.

For chapter 3, some multigroup PCA methods will be presented. The similarities and differences between the different approaches that will be illustrated under a data simulation will be described.

In chapter 4, two new multivariate methods are described in this thesis, which are multigroup principal components analysis and the multigroup dimensionality reduction method. In addition, a univariate method of projection of each dimension for the groups called single dimension exploration (SDE) is described in this chapter.

Chapter 5 describes the R package **Multigroup0** that implements new algorithms multivariate statistical methods proposed in this thesis. This package by framework high-dimensional datasets analysis. Published in the CRAN.

Chapter 6 describes how the simulation data is generated using our **Multigroup0** package. In addition, the real data that is used to present our dimensionality reduction methods is presented.

Chapter 7 explains the simulations results in this investigation. Several data simulation scenarios are presented under different parameters that we are interested in controlling. On the other hand, the results of the real omic data used in this framework where these new multivariate methods are applied to high-dimensional data are presented.

Chapter 8 describes the conclusions and discussions of these two new multivariate techniques, multigroup PCA (mgPCA) and multigroup dimensionality reduction (MDR). The standard PCA is compared with our mgPCA presenting advantages and disadvantages of the new method proposed in this thesis. In addition to describing ideas for future work and the projection of this research.

Chapter 2

Goals

This chapter describes the general and specific objectives of the thesis. This project addresses some problems when analyzing a multigroup dataset in the area of biomedicine.

The general objective of this thesis is:

- Theoretically explore new approaches for dimensionality reduction in multivariate multi-group analysis.
- Implement these new methods in the R software by creating a package and publishing to CRAN.

The project proposes to address specific objectives in different areas of multigroup representation:

1. Develop a methodology that does not require or relax the need for a shared structure of covariances of the groups for the representation in reduced dimension.
2. Develop other criteria to prioritize the separation of groups in representation in reduced dimension.
3. Simulate data with multigroup structure with different scenarios for these two new approaches.
4. Application to real omic data in the area of biomedicine.

Chapter 3

Existing Methods of Analysis of Multigroup Data

In this chapter some methods of multigroup analysis are described, in particular, techniques that are a generalization of the PCA. The notation, symbology and part of the literature of this chapter are based on the article [23]. Furthermore, some these approaches are in the `multigroup` package in R proposed by Eslami et al. (2015).

A summary of some of methods of analysis of multigroup datasets is shown in the next table 3.1.

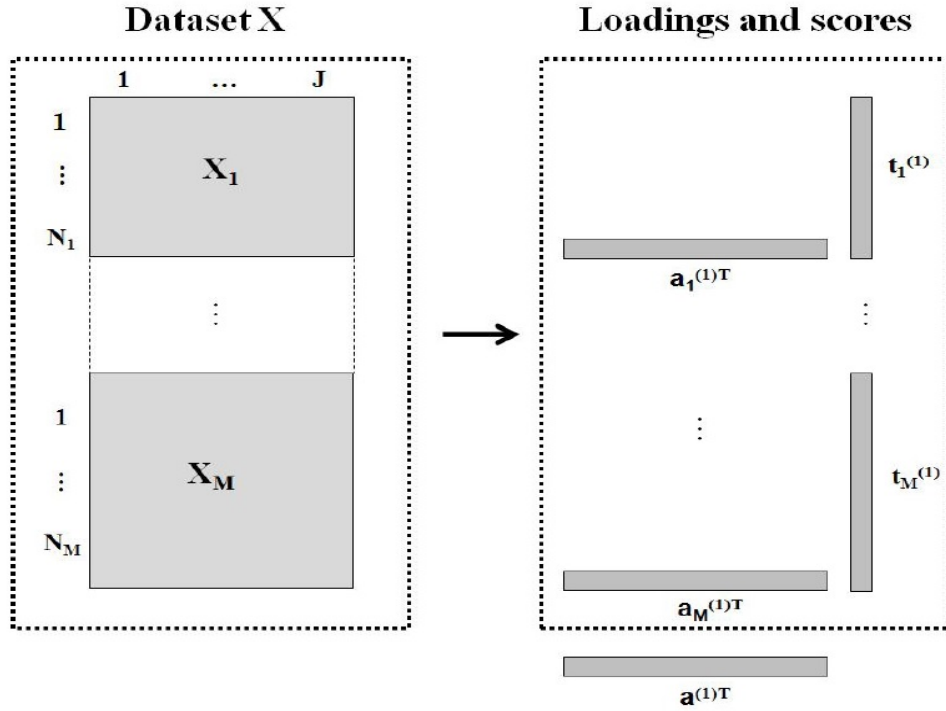


FIG. 1. Graphical display of the common vector of loadings ($\mathbf{a}^{(1)}$), the group vectors of loadings ($\mathbf{a}_1^{(1)}, \dots, \mathbf{a}_M^{(1)}$) and the group components ($\mathbf{t}_1^{(1)}, \dots, \mathbf{t}_M^{(1)}$) by Eslami et al.

Approaches	Criterion
Common Principal Component Analysis (Flury, 1984)	It uses the matrix decomposition $V_m = A\Lambda_m A^\top$. It assumes that the data follows a multivariate normal distribution and estimates the parameters with the maximum likelihood method.
PCA multigroup (Krzanowski, 1984)	It uses the matrix decomposition $V_m = A\Lambda_m A^\top$. Calculate $\mathbf{V}_W = \sum_{m=1}^M \frac{N_m}{N} \mathbf{V}_m$ the within groups variance-covariance matrix. In addition, it minimizes the following criteria: $\min_{\mathbf{V}_c} \sum_{m=1}^M N_m \ \mathbf{V}_m - \mathbf{V}_c\ ^2$.
DUAL STATIS (Lavit, 1994)	It uses the spectral decomposition $V_c = A\Lambda_m A^\top$. Minimize the following criteria: $\min_{\mathbf{V}_c, \alpha_1, \dots, \alpha_M} \sum_{m=1}^M \ \alpha_m \mathbf{V}_m - \mathbf{V}_c\ ^2$ with $\sum_{m=1}^M \alpha_m^2 = 1$ where $\mathbf{V}_c = \sum_{m=1}^M \alpha_m \mathbf{V}_m$ the compromise variance-covariance matrix.
Dual Generalized Procrustes Analysis (Eslami et al., 2013)	Minimize the following criteria: $\sum_{m=1}^M \left\ \frac{1}{\sqrt{N_m}} \mathbf{X}_m^\top \mathbf{H}_m - \mathbf{C} \right\ ^2$ where $\frac{1}{\sqrt{N_m}} \mathbf{X}_m^\top$ is orthogonally transformed towards the common matrix $\mathbf{C}$ by $\mathbf{H}_m$ the orthogonal matrix associated with group m . Use the GPA algorithm (Gower, 1975; Ten Berge, 1977).
Between Groups Comparison	Minimize the following criteria: $\sum_{m=1}^M N_m < \mathbf{a}_m^{(1)}, \mathbf{a}^{(1)} >^2 = \sum_{m=1}^M N_m ((\mathbf{a}^{(1)})^\top \mathbf{a}_m^{(1)})^2 \quad \text{with} \quad \ \mathbf{a}_m^{(1)}\ = \ \mathbf{a}^{(1)}\ = 1$ This method seek $\mathbf{a}^{(1)}$, a common vector of loadings for the first dimension ($h = 1$), and group vectors of loadings respectively associated to the various groups.
Multigroup PCA retrieved (Eslami et al., 2013)	Minimize the following criteria: $\sum_{m=1}^M N_m (\mathbf{a}^{(1)})^\top \mathbf{X}_m^\top \mathbf{X}_m \mathbf{a}^{(1)} = (\mathbf{a}^{(1)})^\top \sum_{m=1}^M N_m \mathbf{X}_m^\top \mathbf{X}_m \mathbf{a}^{(1)}$ where $\mathbf{a}^{(1)}$ is an eigenvector of matrix $\sum_{m=1}^M N_m \mathbf{X}_m^\top \mathbf{X}_m$ associated with the largest eigenvalue.
RGCCA (Tenenhaus et.al, 2011-2014)	The objective of RGCCA is to find, for each block, a weighted composite of variables (called block component) $y_j = X_j a_j$, $j = 1, \dots, J$ (where a_j is a column-vector with p_j elements) summarizing the relevant information between and within the blocks.

TABLE 3.1. Methods of analysis of multigroup data.

Packages in R that implements above existing methods are shown in table 3.2.

Package	Author
Simca-p	Wold, 1966
FactoMineR	Tayrac et al., 2009
mixOmics	Le Cao et al., 2009
lmbc	David et al., 2013
multigroup	A.Eslami et al., 2015
RGCCA	Tenenhaus et.al, 2017

TABLE 3.2. R packages for multivariate analysis

The next section will explain the methods with a multigroup configuration and analyzing the `multigroup` package where the algorithms implemented for these techniques. The iris dataset will be used. More details in <https://archive.ics.uci.edu/dataset/53/iris>.

3.1. Flury's Common Principal Component Analysis

The first idea to analyze the PCA for different groups in the data comes from the frameworks of Flury and Krzanowski in 1984 (see [20], [21]). The analysis of common principal components (CPCA) is a technique that generalizes the PCA for the case of multigroup. The CPCA is expressed in terms of the variance-covariance matrices associated with the M groups as follows:

$$(1) \quad \mathbf{V}_m = \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^\top \quad \text{with} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I} \quad \text{for} \quad m = 1, \dots, M$$

where the matrix $\mathbf{A} = (\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(H)})$ contains the vectors of loadings which are assumed to be common to the various groups and the diagonal matrix $\mathbf{\Lambda}_m$ is supposed to contain the variances associated with the group components $\mathbf{t}_m^{(h)} = \mathbf{X}_m\mathbf{a}^{(h)}$ for dimension $h = 1, \dots, H$. Thus, CPCA determines that the vector of loadings, assumed to be orthogonal, are common to the various groups but the variances associated to the group principal components are specific to each group. For determining the CPCA parameters (namely $\mathbf{A}$ and $\mathbf{\Lambda}_m$), Flury considers a multinormal setting and uses maximum likelihood estimation. This leads to the so-called F-G algorithm [20]. These authors suggest an algorithm based on pairwise minimization of the matrix $\mathbf{A}$. That is, each pair of columns or eigenvectors of $\mathbf{A}$ is updated while holding the others fixed. These updates are based on the eigendecomposition of 2×2 matrices summed across groups. Then, we are required to loop through all pairs of columns of the matrix $\mathbf{A}$ to complete a single iteration. This makes the Flury method ineffective in higher dimensions [30].

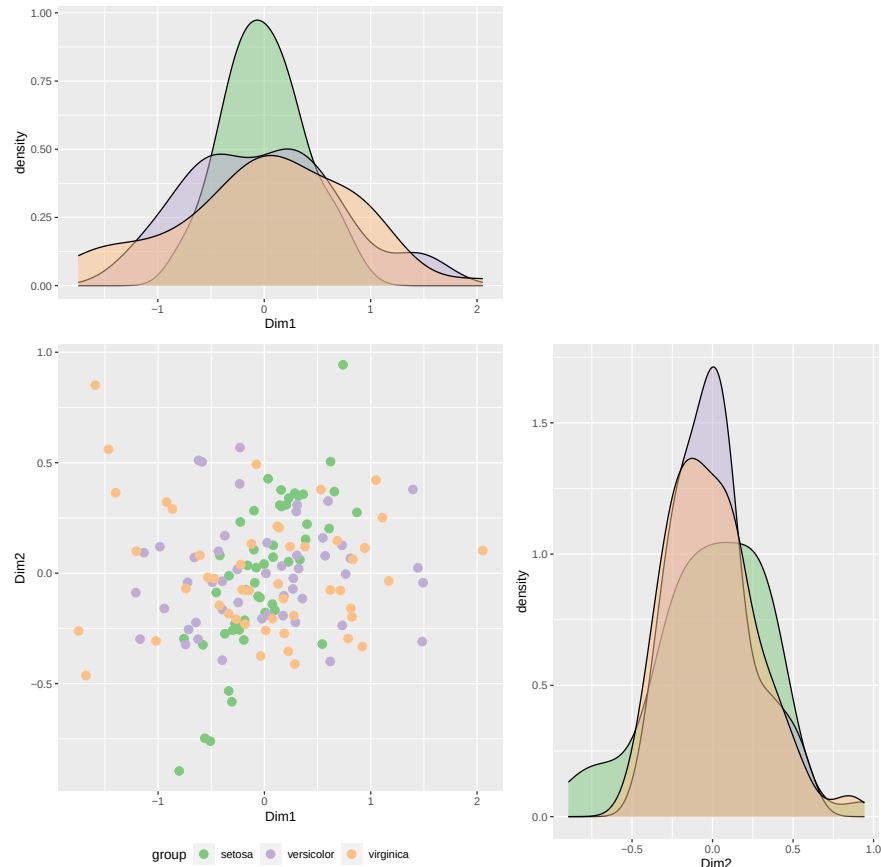


FIG. 2. Flury Common Principal Component Analysis on the iris dataset

In addition the assumption of multinormal setting may be questionable in many situations and, moreover, the algorithm is complex, time consuming and may have some convergence problems [23].

Now we will use the example code of the **multigroup** package that illustrates the iris dataset.

The following chunk code presents the Flury's method and the graphical representation of the first two main components. The graphical display in figure depicts that does not separate the groups in the iris dataset. This approach for a multigroup configuration, the main objective is to search a compromise variance and covariance matrix for to investigate predictive models in the presence of groups in structure of the dataset.

```
require(multigroup)
Data = iris[,-5]
Group = iris[,5]
res.FCPCA = FCPCA(Data, Group)
graf_generic(x=res.FCPCA, axes = c(1, 2), cex = NULL, font.lab = NULL)
```

3.2. Multigroup Principal Components Analysis

This technique was proposed by Krzanowski in 1984 called multigroup principal components analysis (MGPCA). Indeed, [21] remarks that if $\mathbf{V}_m = \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^\top$ stands for all m then it also stands for the following linear combination of the variance-covariance matrices $(\mathbf{V}_1, \dots, \mathbf{V}_M)$:

$$(2) \quad \sum_{m=1}^M \frac{N_m}{N} \mathbf{V}_m = \sum_{m=1}^M \frac{N_m}{N} \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^\top = \mathbf{A} \left(\sum_{m=1}^M \frac{N_m}{N} \mathbf{\Lambda}_m \right) \mathbf{A}^\top \quad \text{with} \quad \mathbf{A}^\top \mathbf{A} = \mathbf{I}$$

Therefore, the matrix of common loadings $\mathbf{A}$ could be derived from the eigen analysis of the matrix $\mathbf{V}_W = \sum_{m=1}^M \frac{N_m}{N} \mathbf{V}_m$ which is referred to in the literature as the within groups variance-covariance matrix. From this standpoint the matrix $\mathbf{V}_W$ can be seen as a common variance-covariance matrix to the various groups because it is the closest matrix to $(\mathbf{V}_1, \dots, \mathbf{V}_M)$, in the sense that it minimizes the following criterion:

$$(3) \quad \min_{\mathbf{V}_c} \sum_{m=1}^M N_m \|\mathbf{V}_m - \mathbf{V}_c\|^2$$

In summary, the strategy of analysis proposed by [21] consists in computing $\mathbf{V}_W$, the within groups variance-covariance matrix. The common vectors of loadings $(\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(H)})$ are the eigenvectors of $\mathbf{V}_W$. The specific variances of group m are determined as $\lambda_m^{(h)} = (\mathbf{a}^{(h)})^\top \mathbf{V}_m \mathbf{a}^{(h)}$ for $h = 1, \dots, H$.

As a matter of fact, the author Krzanowski [21] remarks that if $\mathbf{V}_m = \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^\top$ stands for all m , then it stands not only for $\mathbf{V}_W$, but also for any linear combination of $\mathbf{V}_W$ as shown in (4)

$$(4) \quad \sum_{m=1}^M \alpha_m \mathbf{V}_m = \sum_{m=1}^M \alpha_m \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^\top = \mathbf{A} \left(\sum_{m=1}^M \alpha_m \mathbf{\Lambda}_m \right) \mathbf{A}^\top \quad \text{with} \quad \alpha_m \geq 0$$

From this standpoint it is seek a common variance-covariance matrix as a solution to other optimization criteria than the problem stated above in (3). This is illustrated by means of dual STATIS which is discussed in the next section.

Note that MGPCA amounts to performing PCA on the matrix $\mathbf{X}$ obtained by stacking the group datasets $(\mathbf{X}_1, \dots, \mathbf{X}_M)$ one above the other. Indeed, it is easy to check that in this case, it is led to the eigen analysis of matrix $\frac{1}{N} \mathbf{X}^\top \mathbf{X} = \sum_{m=1}^M \frac{N_m}{N} \mathbf{V}_m = \mathbf{V}_W$ [21], [23]. In this case the Multigroup PCA algorithm is NIPALS (nonlinear iterative partial least squares).

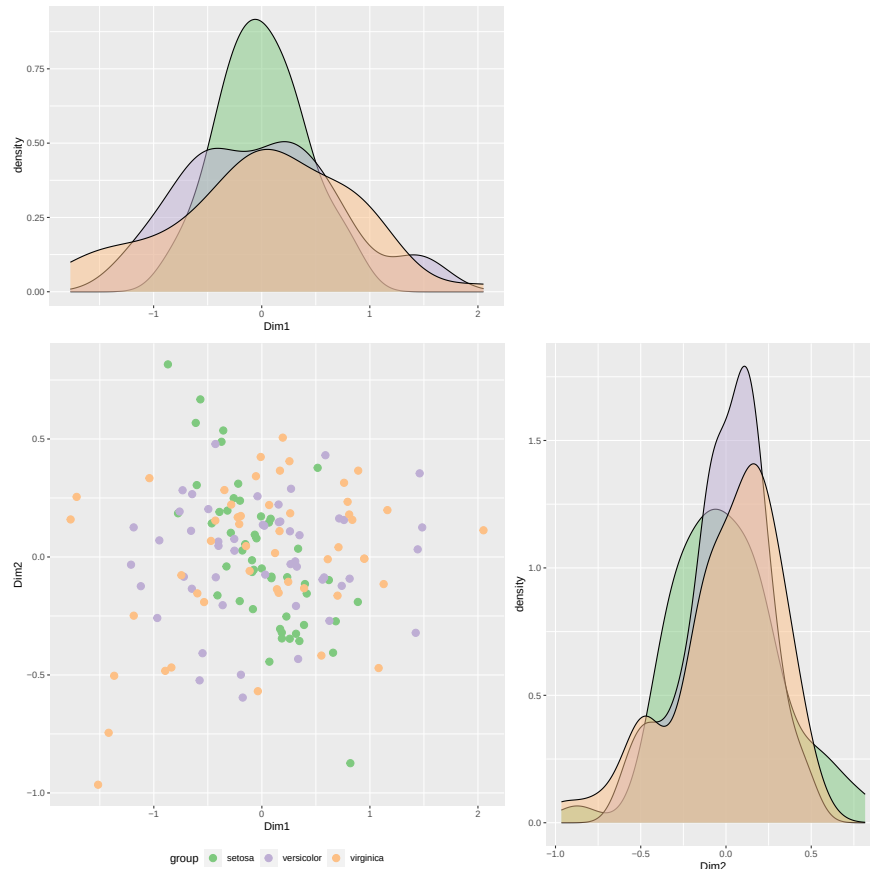


FIG. 3. Multigroup Principal Components Analysis on the iris dataset

The following chunk code presents the MGPCA method and the graphical representation of the first two main components. The figure show that does not separate the groups for the iris dataset. These methods presented in this chapter have the same objective of searching a compromise variance and covariance matrix to obtain predictive models in the presence of groups.

```
require(multigroup)
Data = iris[,-5]
Group = iris[,5]
res.mgPCA = mgPCA (Data, Group)
graf_generic(x=res.mgPCA, axes = c(1, 2), cex = NULL, font.lab = NULL)
```

3.3. DUAL STATIS

This method was proposed by Lavit in 1994 that gives a solution to the problem of optimization of (3) which is an alternative criterion presented in (5):

$$(5) \quad \min_{\mathbf{V}_c, \alpha_1, \dots, \alpha_M} \sum_{m=1}^M \|\alpha_m \mathbf{V}_m - \mathbf{V}_c\|^2 \quad \text{with} \quad \sum_{m=1}^M \alpha_m^2 = 1$$

The ACT (STATIS method) is, an exploratory technique of multivariate data analysis based on linear algebra and especially Euclidean vector spaces [31]. It has been devised for multiway data situations (a set of several matrices or each dimension is a set of data from different sources of variation with the same individuals) on the basic idea of computing Euclidean distances between configurations of points [32].

It can also be said that this technique in framework of multiblock data analysis which seeks a common configuration to several datasets measured on the same individuals. However, when these datasets pertain to the same variables instead of the same individuals, it is referred to as dual STATIS (DSTATIS). The solution to problem (5) is given by the compromise variance-covariance matrix $\mathbf{V}_c = \sum_{m=1}^M \alpha_m \mathbf{V}_m$ where $\alpha = (\alpha_1, \dots, \alpha_M)^\top$ is the eigenvector of the matrix $\mathbf{R} = (r_{ik})$ associated with the largest eigenvalue, where $r_{ik} = \text{tr}(\mathbf{V}_i \mathbf{V}_r)$ for $i, k = 1, \dots, M$. Thereafter, the spectral decomposition of $\mathbf{V}_c$ can be computed as $\mathbf{V}_c = \mathbf{A} \mathbf{\Lambda} \mathbf{A}^\top$, where $\mathbf{A} = (\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(H)})$ will stand for the matrix of common vectors of loadings $\mathbf{A}$ and is a diagonal matrix. The specific variances of group m are computed as $\lambda_m^{(h)} = (\mathbf{a}^{(h)})^\top \mathbf{V}_m \mathbf{a}^{(h)}$ for dimensions $h = 1, \dots, H$.

This method has the advantage of taking into account the similarity between the matrices of variances and covariances between the groups. That is, if the structure of one group is different from the other groups, then the contribution of this group to the determination of the common variance-covariance matrix, $\mathbf{V}_c$, will be minimized, i.e., the associated variance-covariance matrix is down weighted because its associated coefficient α_m is relatively small [23]. In addition, its graphic interpretation is easy compared to the other previous methods. More detail of this approach is in the article by Lavit [31].

The following chunk code presents the Dual STATIS method and the graphical representation of the first two main components. The figure show that does not separate the groups for the iris dataset. This method, like the other approaches, the objective is not the visualization in the separation of groups in the structure of the dataset.

```
Data = iris[,-5]
Group = iris[,5]
res.DSTATIS = DSTATIS(Data, Group, ncomp = 2)

Data = as.matrix(Data)
```

```

Group = as.factor(Group)
M = length(levels(Group))
n = as.vector(table(Group))
N = sum(n)
split.Data1<-NULL
split.Data= split(Data, Group)
for (m in 1:M) {
  split.Data1[[m]] = matrix(split.Data[[m]],nrow=n[m])
  split.Data1[[m]] = scale(split.Data1[[m]], center = TRUE,
    scale = F)
}
Con.Data<-split.Data1[[1]]
for (m in 2:M) {
  Con.Data = rbind(Con.Data, split.Data1[[m]])
}
cov.Group = vector("list", M)
for (m in 1:M) {
  cov.Group[[m]] = t(split.Data1[[m]]) %*% split.Data1[[m]]/n[m]
}
Vm<-cov.Group

res.DSTATIS$lambda[1,1]/sum(diag(Vm[[1]]))*100

## [1] 51.48175

res.DSTATIS$lambda[2,1]/sum(diag(Vm[[2]]))*100

## [1] 77.5728

res.DSTATIS$lambda[3,1]/sum(diag(Vm[[3]]))*100

## [1] 77.30301

res.DSTATIS$lambda[1,2]/sum(diag(Vm[[1]]))*100

## [1] 32.01109

res.DSTATIS$lambda[2,2]/sum(diag(Vm[[2]]))*100

## [1] 9.308736

```

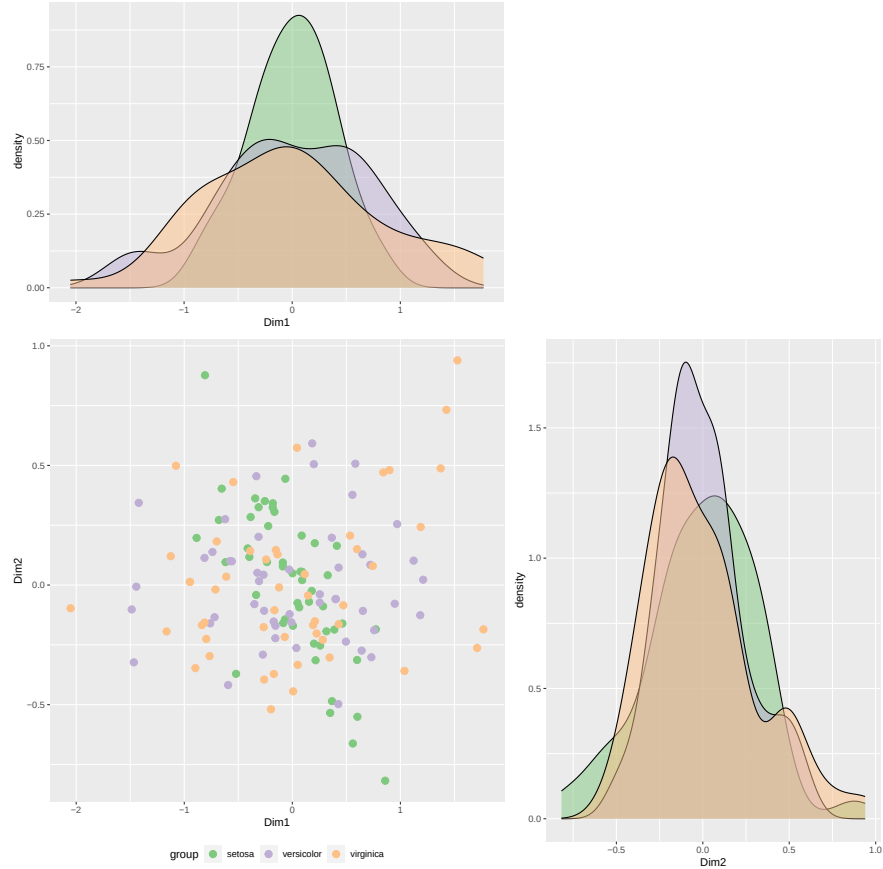


FIG. 4. DUAL STATIS on the iris dataset

```
res.DSTATIS$lambda[3,2]/sum(diag(Vm[[3]]))*100

## [1] 11.37147

graf_generic(x=res.DSTATIS, axes = c(1, 2), cex = NULL, font.lab = NULL)
```

3.4. DUAL Generalized Procrustes Analysis

Suppose that X_1 and X_2 are two centered datasets with $(n \times J)$ and $(n \times J)$, respectively, which refer to the same J variables but not necessarily to the same individuals. Without any loss of generality these two datasets are assumed to have the same number of individuals. If this is not the case then the dataset with the smallest number of rows can be augmented with the necessary number of rows containing zeroes. The case for the same individuals is proven by Glacon [33] (1981).

The technique of dual generalized procrustes analysis is to find a dataset that is the average of groups $\frac{1}{\sqrt{N_m}} \mathbf{X}_m^\top$ through an orthogonal transformation. This can be achieved by means of

generalized Procrustes analysis (GPA). Then the approach to use $\mathbf{X}_m^\top$ instead of $\mathbf{X}_m$ is called dual GPA (DGPA), that seeks to minimize the following criteria:

$$\sum_{m=1}^M \left\| \frac{1}{\sqrt{N_m}} \mathbf{X}_m^\top \mathbf{H}_m - \mathbf{C} \right\|^2$$

where $\frac{1}{\sqrt{N_m}} \mathbf{X}_m^\top$ is orthogonally transformed towards the common matrix $\mathbf{C}$ by $\mathbf{H}_m$, the orthogonal matrix associated with group m . This optimization problem can be solved by one of the several algorithms for GPA (Gower [25], 1975; Ten Berge [34], 1977). Once $\mathbf{C}$ is determined, the common vectors of loadings $\mathbf{A} = (\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(H)})$ are calculated as the left singular vectors of $\mathbf{C}$. The specific variances of group m are given as usual by $\lambda_m^h = (\mathbf{a}^{(h)})^\top \mathbf{V}_m \mathbf{a}^{(h)}$.

The following chunk code presents the DUAL Generalized Procrustes Analysis method and the graphical representation of the first two main components. The figure shows that does not achieve the separation of the groups in the iris dataset.

```
Data = iris[, -5]
Group = iris[, 5]
res.DGPA = DGPA(Data, Group)
graf_generic(x=res.DGPA, axes = c(1, 2), cex = NULL, font.lab = NULL)
```

3.5. Between Groups Comparison

This multigroup PCA technique proposed in [35] is to try to compare two groups with different individuals or observations under the same variables. It also describes an extension for more than two groups. In his article he gives a geometric approach and its interpretation.

The aim of this method is to seek $\mathbf{a}^{(1)}$, a common vector of loadings for the first dimension ($h = 1$) and group vectors of loadings respectively associated to the various groups, namely $(\mathbf{a}_1^{(1)}, \dots, \mathbf{a}_M^{(1)})$ to the common vector of loadings $\mathbf{a}^{(1)}$. The following criterion of maximization is considered:

$$(6) \quad \sum_{m=1}^M N_m < \mathbf{a}_m^{(1)}, \mathbf{a}^{(1)} >^2 = \sum_{m=1}^M N_m ((\mathbf{a}^{(1)})^\top \mathbf{a}_m^{(1)})^2 \quad \text{with} \quad \|\mathbf{a}_m^{(1)}\| = \|\mathbf{a}^{(1)}\| = 1$$

or equivalently:

$$\sum_{m=1}^M N_m \cos^2(\mathbf{a}_m^{(1)}, \mathbf{a}^{(1)})$$

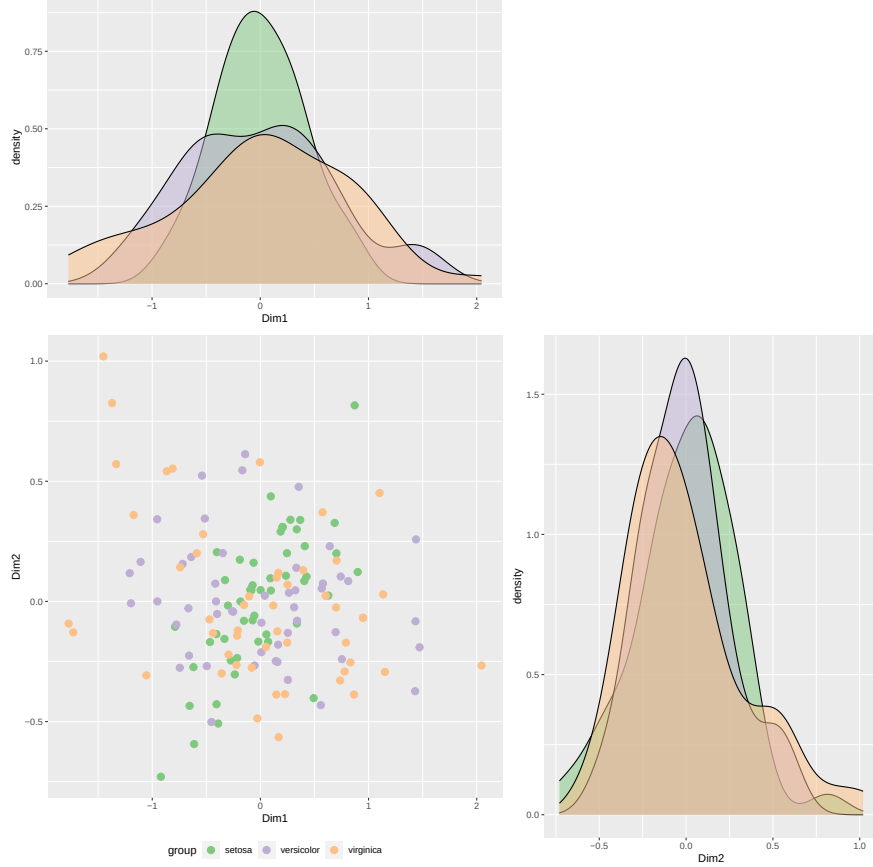


FIG. 5. DUAL Generalized Procrustes Analysis on the iris dataset

where $\cos(\cdot, \cdot)$ stands for the cosine between two vectors.

In order to solve this problem, we firstly can recall that a vector of loadings associated with group m is, as stated above, supposed to be a linear combination of the columns of matrix $\mathbf{X}_m^\top$, $\mathbf{a}_m^{(1)} = \mathbf{X}_m^\top \mathbf{t}_m^{(1)}$. Therefore, in criterion 6 the optimal solution for vector $\mathbf{a}_m^{(1)}$ is given by $\mathbf{a}_m^{(1)} = \frac{\mathbf{P}_m \mathbf{a}^{(1)}}{\|\mathbf{P}_m \mathbf{a}^{(1)}\|}$ where $\mathbf{P}_m = \mathbf{X}_m^\top (\mathbf{X}_m \mathbf{X}_m^\top)^{-1} \mathbf{X}_m$, the projector upon the subspace spanned by the columns in $\mathbf{X}_m^\top$. Then it maximizes with respect to $\mathbf{a}^{(1)}$ the following expression:

$$(7) \quad \sum_{m=1}^M N_m (\mathbf{a}^{(1)})^\top \mathbf{P}_m \mathbf{a}^{(1)} = (\mathbf{a}^{(1)})^\top \sum_{m=1}^M N_m \mathbf{P}_m \mathbf{a}^{(1)} \quad \text{with} \quad \|\mathbf{a}^{(1)}\| = 1$$

It follows that the optimal solution is given by setting $\mathbf{a}^{(1)}$ equal to the eigenvector of $\sum_{m=1}^M N_m \mathbf{P}_m$ associated with the largest eigenvalues.

The optimization problem (6) that has restriction in the common vector of loadings is related to generalized canonical correlation analysis of the M spaces spanned by the columns of $(\mathbf{X}_1^\top, \dots, \mathbf{X}_M^\top)$, respectively (Carroll [36], 1968).

The solution to the optimization problem assumes that the matrices $\mathbf{X}_m \mathbf{X}_m^\top$ are invertible, although for the case of groups where the number of variables is smaller than the number of individuals, the $(N_m \times N_m)$ matrices $\mathbf{X}_m \mathbf{X}_m^\top$ are of rank less than N_m and therefore not invertible. For this problem, the authors [23] propose to approach by using Eckart and Young (1936), [37], theorem each matrix $\mathbf{X}_m^\top$ by a matrix $(\mathbf{X}_m^*)^\top$ of rank k ($k < N_m$) such that the matrices $\mathbf{X}_m^* (\mathbf{X}_m^*)^\top$ are invertible.

Using the singular value decomposition of order k in the matrix $(\mathbf{X}_m^*)^\top$ as:

$$(\mathbf{X}_m^*)^\top = \mathbf{L}_m^k \mathbf{\Lambda}_m^k (\mathbf{Q}_m^k)^\top$$

$$\text{then, } \sum_{m=1}^M N_m (\mathbf{X}_m^*)^\top (\mathbf{X}_m^* (\mathbf{X}_m^*)^\top)^{-1} \mathbf{X}_m^* = \sum_{m=1}^M N_m \mathbf{L}_m^k (\mathbf{L}_m^k)^\top.$$

This strategy of analysis amounts to the method of analysis which was proposed by Krzanowski under the name "between-groups comparison of principal components". The authors [23] of the **multigroup** package is refer to this method as BCG.

The following chunk code presents the BCG method and the graphical representation of the first two main components. The figure shows that does not achieve the separation of the groups in the iris dataset.

```
Data = iris[, -5]
Group = iris[, 5]
res.BGC = BGC(Data, Group)
graf_generic(x=res.BGC, axes = c(1, 2), cex = NULL, font.lab = NULL)
```

3.6. Multi-group PCA retrieved

The authors Eslami et al.(2013) propose the same idea of the previous method, Between Groups Comparison (BGC), but with other limitations of the parameters. The criteria to maximize is:

$$(8) \quad \sum_{m=1}^M N_m < \mathbf{a}_m^{(1)}, \mathbf{a}^{(1)} >^2 \quad \text{with} \quad \mathbf{a}_m^{(1)} = \mathbf{X}_m^\top \mathbf{t}_m^{(1)} \quad \text{and} \quad \|\mathbf{a}_m^{(1)}\| = \|\mathbf{t}_m^{(1)}\| = 1$$

The method maximizes under the following specified constraints:

$$(9) \quad \sum_{m=1}^M N_m ((\mathbf{a}^{(1)})^\top \mathbf{X}_m^\top \mathbf{t}_m^{(1)})^2$$

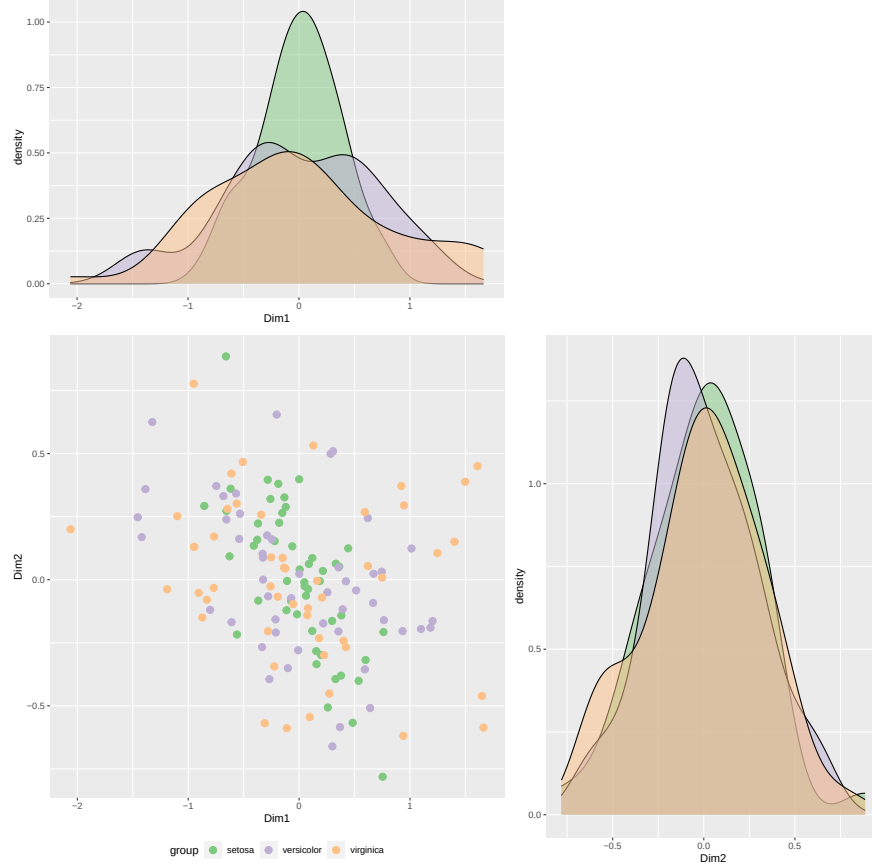


FIG. 6. Between Groups Comparison on the iris dataset

For a fixed vector $\mathbf{a}^{(1)}$ and let $\mathbf{t}_m^{(1)} = \frac{\mathbf{X}_m \mathbf{a}^{(1)}}{\|\mathbf{X}_m \mathbf{a}^{(1)}\|}$. Replacing in the equation (9), the maximizing is:

$$(10) \quad \sum_{m=1}^M N_m (\mathbf{a}^{(1)})^\top \mathbf{X}_m^\top \mathbf{X}_m \mathbf{a}^{(1)} = (\mathbf{a}^{(1)})^\top \sum_{m=1}^M N_m \mathbf{X}_m^\top \mathbf{X}_m \mathbf{a}^{(1)}$$

The $\mathbf{a}^{(1)}$ is an eigenvector of matrix $\sum_{m=1}^M N_m \mathbf{X}_m^\top \mathbf{X}_m$ associated with the largest eigenvalue. Subsequent common vectors of loadings can be found considering the same solution as the multigroup PCA (MGPCA) proposed by Krzanowski.

All these developments make it possible to highlight a striking difference between BGC developed in the previous section and MGPCA. It is clear that whereas in MGPCA the aim at recovering the total variance in the various groups, in BGC the main purpose is to find similar patterns no matter whether these patterns are linked to directions of high saliency (i.e., total variance) or not. Indeed, in MGPCA, the aim at maximizing (10),

$$(11) \quad \sum_{m=1}^M N_m (\mathbf{a}^{(1)})^\top \mathbf{X}_m^\top \mathbf{X}_m \mathbf{a}^{(1)} = \sum_{m=1}^M N_m \lambda_m^{(1)}$$

where $\lambda_m^{(1)}$ is the variance of the principal component $\mathbf{t}_m^{(1)} = \mathbf{X}_m \mathbf{a}^{(1)}$ [23].

3.7. Comparison of Methods

For the comparison of the common vectors of loadings between methods, it is set up, for each pair of methods, a sequence of indices indexed by the number of components retained in the model. Each index ranges between 0 and 1 and reflects the extent to which the common vectors of loadings obtained by the two methods at hand are similar [23]. Let $\mathbf{A} = (\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(H)})$ and $\mathbf{B} = (\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(H)})$ be the matrices of common loadings associated with two methods to be compared. It is investigated whether these methods lead to similar vectors of loadings up to a given dimension h for $h = 1, \dots, H$. Is considered the sequence of similarity indices given by:

$$S^h = \frac{1}{h} \sum_{r=1}^h |(\mathbf{a}^{(r)})^\top (\mathbf{b}^{(r)})| \quad h = 1, \dots, H$$

The indices S^h range between 0 and 1, the value 1 being reached in case of perfect agreement of the vectors of loadings up to dimension h . Contrariwise, this index takes the value 0 if $(\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(H)})$ are respectively orthogonal to $(\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(H)})$.

The methods are also compared on the basis of the total variances in the various groups recovered by the successive principal components which are given by:

$$\lambda_m^h = (\mathbf{a}^{(h)})^\top \mathbf{V}_m \mathbf{a}^{(h)} \quad h = 1, \dots, H$$

or equivalently the percentages of total variance recovered by the principal components:

$$I_m^h = \frac{\lambda_m^h}{\text{trace}(\mathbf{V}_m)} \quad h = 1, \dots, H \quad m = 1, \dots, M$$

In situations where the aim is to recover as much variation in the dataset as possible, these percentages should be large.

Now the chunk code shows the comparison of the common vectors of loadings between the methods through the similarity index that compares pairs.

```
#####
## Similarity between partial and common loadings
#####
require(multigroup)
Data = iris[,-5]
Group = iris[,5]
#####
#Between Group Comparison
#####
res.BGC = BGC(Data, Group, graph=F,ncomp = 2)
BGC01<-res.BGC$loadings.common
#####
#Dual Generalized Procrustes Analysis
#####
res.DGPA = DGPA(Data, Group, graph=F,ncomp = 2)
DGPA01<-res.DGPA$loadings.common
#####
#Multigroup Principal Component Analysis
#####
res.mgPCA = mgPCA (Data, Group,ncomp = 2, Scale = FALSE, graph = FALSE)
MGPCA01<-res.mgPCA$loadings.common
#####
#Dual STATIS
#####
res.DSTATIS = DSTATIS(Data, Group, graph=F,Scale = F,ncomp =2)
DSTATIS01<-res.DSTATIS$loadings.common
#####
#Flury's Common Principal Component Analysis(FCPCA)
#####
res.FCPCA = FCPCA(Data, Group, graph=F)
FCPCA01<-res.FCPCA$loadings.common
FCPCA01<-FCPCA01[,1:2]
#####
#PCA standard
#####
pcs<-prcomp(Data)
names(pcs)
```

```
## [1] "sdev"      "rotation" "center"   "scale"    "x"

PCA<-pcs$rotation[,1:2]
#####
loadings_matrices = list()
loadings_matrices[[1]]=DGPA01
loadings_matrices[[2]]=DSTATIS01
loadings_matrices[[3]]=MGPCA01
loadings_matrices[[4]]=FCPCA01
loadings_matrices[[5]]=BGC01
loadings_matrices[[6]]=PCA
NAMES = c("DGPA", "DSTATIS", "MGPCA", "FCPCA", "BGC", "PCA")
#####
# Similarity indices  $S^{\{h\}}$  for dimension ( $h=1,2$ )
#####
Similarity =similarity_function(loadings_matrices=loadings_matrices, NAMES)
table1<-rbind(as.data.frame(Similarity[[1]]),c(0.9412,0.9536,0.9518,
                                              0.9592,0.9378,0.7876))
rownames(table1)<-c("DGPA",      "DSTATIS", "MGPCA" ,  "FCPCA" ,
                  "BGC" ,      "PCA", "Average")
table1

##           DGPA DSTATIS  MGPCA  FCPCA    BGC    PCA
## DGPA      1.0000  0.9970 0.9980 0.9900 0.9690 0.7520
## DSTATIS  0.9970  1.0000 1.0000 0.9980 0.9790 0.7940
## MGPCA    0.9980  1.0000 1.0000 0.9970 0.9770 0.7870
## FCPCA    0.9900  0.9980 0.9970 1.0000 0.9850 0.8260
## BGC      0.9690  0.9790 0.9770 0.9850 1.0000 0.7790
## PCA      0.7520  0.7940 0.7870 0.8260 0.7790 1.0000
## Average  0.9412  0.9536 0.9518 0.9592 0.9378 0.7876

table2<-rbind(as.data.frame(Similarity[[2]]),c(0.9034,0.9218,0.9208,
                                              0.9064,0.8468,0.7116))
rownames(table2)<-c("DGPA",      "DSTATIS", "MGPCA" ,  "FCPCA" ,
                  "BGC" ,      "PCA", "Average")
table2

##           DGPA DSTATIS  MGPCA  FCPCA    BGC    PCA
## DGPA      1.0000  0.9940 0.9920 0.9400 0.9350 0.6560
```

## DSTATIS	0.9940	1.0000	0.9990	0.9670	0.9240	0.7250
## MGPCA	0.9920	0.9990	1.0000	0.9730	0.9120	0.7280
## FCPCA	0.9400	0.9670	0.9730	1.0000	0.8330	0.8190
## BGC	0.9350	0.9240	0.9120	0.8330	1.0000	0.6300
## PCA	0.6560	0.7250	0.7280	0.8190	0.6300	1.0000
## Average	0.9034	0.9218	0.9208	0.9064	0.8468	0.7116

These last results show the similarity index, S^h , between all the multigroup methods for the first two dimensions ($h = 1$ and $h = 2$). This is described in table 1 and 2 in the chunk code. The standard PCA was then added for comparison with these multigroup methods.

The output of the chunk code shows the most similar methods to multigroup configuration for dimensions 1 and 2. However, when compared with the standard PCA, considering the average similarity values, it is shown that this last approach is very different from the multigroup methods, since the standard PCA does not consider the presence of groups in the structure of the dataset. BGC analysis shows less similarity in both dimensions with the other methods. DSTATIS and FCPCA are the methods which give the most similar results in comparison with all other methods to the first two dimensions ($h = 1, 2$).

3.8. Regularized Generalized Canonical Correlation Analysis

This multivariate method analyzes data structures in blocks of variables or in groups of individuals. This approach is RGCCA package but the multigroup analysis in the individuals is not implemented, only the multiblock configuration of variables on the same group of individuals. Although the multigroup configuration is not implemented, this section will describe the multigroup analysis approach that is developing his theory.

Is considered J data matrices $\mathbf{X}_1, \dots, \mathbf{X}_J$. Each $n \times p_j$ data matrix $\mathbf{X}_j = (x_{j1}, \dots, x_{jp_j})$ is called a block and represents a set of p_j variables observed on n individuals. The number and the nature of the variables may differ from one block to another, but the individuals must be the same across blocks. It is assumed that all variables are centered. The objective of RGCCA (regularized generalized canonical correlation analysis) is to find, for each block, a weighted composite of variables (called block component) $\mathbf{y}_j = \mathbf{X}_j \mathbf{a}_j$, $j = 1, \dots, J$ (where $\mathbf{a}_j$ is a column-vector with p_j elements) summarizing the relevant information between and within the blocks.

RGCCA for multiblock and multigroup data analyses is based on the following optimization problem: It is considered I matrices $\mathbf{Q}_1, \dots, \mathbf{Q}_I$. Each matrix $\mathbf{Q}_i$ is of dimension $m \times p_i$. Is associated to each matrix $\mathbf{Q}_i$ a weight column vector $\mathbf{w}_i$ of dimension p_i and a symmetric

definite positive matrix $\mathbf{M}_i$ of dimensions $p_i \times p_i$. Moreover, a design matrix $\mathbf{C} = c_{il}$ is defined with $c_{il} = 1$ if matrices $\mathbf{Q}_i$ and $\mathbf{Q}_l$ are connected, and $c_{il} = 0$ otherwise. The central optimization problem considered, by the authors A.Tenenhaus and M.Tenenhaus [24], is defined as follows:

$$(12) \quad \begin{aligned} & \max_{\mathbf{v}_1, \dots, \mathbf{v}_I} \sum_{i,l,i \neq l} c_{il} g(\langle \mathbf{P}_i \mathbf{v}_i, \mathbf{P}_l \mathbf{v}_l \rangle) \\ & \text{s.c.} \quad \mathbf{v}_i^\top \mathbf{v}_i = 1, \quad i = 1, \dots, I \end{aligned}$$

where $\mathbf{v}_i = \mathbf{M}_i^{1/2} \mathbf{w}_i$ and $\mathbf{P}_i = \mathbf{Q}_i \mathbf{M}_i^{1/2}$. The function $g(\cdot)$ is any continuous convex function.

The RGCCA algorithm proposed in A.Tenenhaus and M.Tenenhaus [38] is a slightly modified PLS (partial least squares) algorithm with known convergence properties. The RGCCA algorithm is extended in order to cover the multigroup situation.

3.8.1. Multigroup data analysis. In the multigroup framework, a row partition $\mathbf{X} = (\mathbf{X}_1^\top, \dots, \mathbf{X}_i^\top, \dots, \mathbf{X}_I^\top)$ of the data matrix is considered. In this method, the same set of variables is observed on different groups of individuals. Each $n_i \times p$ matrix $\mathbf{X}_i$ is called a group and represents a set of p centered and normalized (i.e. unit norm) variables observed on a set of n_i individuals. This normalization implies that (i) the correlation matrix $\mathbf{R}_i$ related to $\mathbf{X}_i$ is equal to $\mathbf{X}_i^\top \mathbf{X}_i$ and (ii) the correlation matrix $\mathbf{R}$ related to the whole data matrix $\mathbf{X}$ is equal

$$\text{to } \bar{\mathbf{R}} = \frac{\sum_{i=1}^I \mathbf{R}_i}{I}.$$

The authors of paper to use RGCCA for multigroup data analysis. The objective of RGCCA for multigroup data analysis is to find group-loading vectors with small angles and high norms. However, the group-loading vectors are obtained sequentially: the related group components are constrained to be orthogonal to the previous ones.

A $p \times p$ positive definite symmetric matrix $\mathbf{M}_i$ is associated with each $\mathbf{X}_i$. A design matrix $\mathbf{C} = c_{il}$ is defined with $c_{il} = 1$ if blocks i and l are related, and $c_{il} = 0$ otherwise and a function g where is defined as the identity, the absolute value or the square function.

The following optimization problem is considered:

$$(13) \quad \begin{aligned} & \max_{\mathbf{w}_1, \dots, \mathbf{w}_I} \sum_{i,l,i \neq l} c_{il} g(\langle \mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l \rangle) \\ & \text{s.c.} \quad \mathbf{w}_i^\top \mathbf{M}_i \mathbf{w}_i = 1, \quad i = 1, \dots, I \end{aligned}$$

This optimization problem is a special case of optimization problem (12) with $\mathbf{Q}_i = \mathbf{X}_i^\top \mathbf{X}_i$ and m and p_i both equal to the number p of variables. Vector $\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i$ is called a group-loading vector. It is also introduce the standardized group-loading vector $\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i / \|\mathbf{X}_i \mathbf{w}_i\|$. Due to the unit norm of each column of $\mathbf{X}_i$, the expression $\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i / \|\mathbf{X}_i \mathbf{w}_i\|$ represents the vector of correlations between the columns of $\mathbf{X}_i$ and the group component $\mathbf{X}_i \mathbf{w}_i$.

On the other hand, this RGCCA approach considers three typical situations that give a good idea of the generality and power of the optimization problem (13) for multigroup data analysis.

3.8.1.1. *Situation 1.* A situation is considered in which all the groups are related: $c_{il} = 1$ for $i \neq l$. The matrix $\mathbf{M}_i = \mathbf{X}_i^\top \mathbf{X}_i$ is considered for each group. Therefore, in this situation, matrices $\mathbf{X}_i$ are supposed to be full rank ($\mathbf{M}_i$ is supposed to be positive definite). For g equals to the identity, problem (13) becomes:

$$(14) \quad \begin{aligned} \max_{\mathbf{w}_1, \dots, \mathbf{w}_I} \quad & \sum_{i,l,i \neq l} \cos(\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l) \times \|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i\| \times \|\mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l\| \\ \text{s.c.} \quad & \|\mathbf{X}_i \mathbf{w}_i\| = 1, \quad i = 1, \dots, I \end{aligned}$$

As the maximization of $\|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i\|$ s.c. $\|\mathbf{X}_i \mathbf{w}_i\| = 1$ yields the normalized first principal component of group $\mathbf{X}_i$, problem (14) leads to a compromise between separate one dimension PCA's of the various groups $\mathbf{X}_i$'s and group-loading vectors with small angles. Due to the chosen normalization, the group-loading vectors $\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i$ represent the vectors of correlations between the various columns $\bar{\mathbf{x}}_{ij}$ of group $\mathbf{X}_i$ and the group component $\mathbf{X}_i \mathbf{w}_i$. Due to cosines being as high as possible between the group-loading vectors, interpretations of the various group components are expected to be similar.

Higher dimension solutions are searched for by deflation on the previous components. Let $\mathbf{w}_i^1$, $i = 1, \dots, I$ the optimal solution for problem (14). Then the second dimension components are constrained to be orthogonal to the first dimension components. As any vector $\mathbf{X}_i \mathbf{w}_i$ orthogonal to $\mathbf{X}_i \mathbf{w}_i^1$ can be written as $\mathbf{X}_i^1 \mathbf{w}_i$ where $\mathbf{X}_i^1$ is the residual matrix of the regression of $\mathbf{X}_i$ on $\mathbf{X}_i \mathbf{w}_i^1$.

Therefore, searching for the second dimension group-loading vectors boils down to replacing in optimization problem (14) each matrix $\mathbf{X}_i$ by the residual matrix $\mathbf{X}_i^1$ of the regression of $\mathbf{X}_i$ on the first dimension group component $\mathbf{X}_i \mathbf{w}_i^1$. The same procedure is used for higher dimensions.

3.8.1.2. *Situation 2.* Dataset matrices $\mathbf{X}_i$ are not necessarily full rank. Then problem (14) is modified by using $\mathbf{M}_i = \mathbf{I}$ for each group and consider this new problem:

$$(15) \quad \begin{aligned} & \max_{\mathbf{w}_1, \dots, \mathbf{w}_I} \sum_{i, l, i \neq l} \cos(\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l) \times \|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i\| \times \|\mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l\| \\ & \text{s.c.} \quad \|\mathbf{w}_i\| = 1, \quad i = 1, \dots, I \end{aligned}$$

As the maximization of $\|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i\|$ s.c. $\|\mathbf{w}_i\| = 1$ yields the first principal component of group $\mathbf{X}_i$, problem (15) leads to a compromise between separate one dimension PCA's of the various groups $\mathbf{X}_i$'s and group-loading vectors with small angles. So, what is the difference between problems (14) and (15)? The criterion used in (15) can be re-written using standardized group-loading vectors and this yields the following expression:

$$(16) \quad \begin{aligned} & \sum_{i, l, i \neq l} \sqrt{n_i n_l} \cos(\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l) \times \sqrt{\text{Var}(\mathbf{X}_i \mathbf{w}_i)} \times \sqrt{\text{Var}(\mathbf{X}_l \mathbf{w}_l)} \\ & \times \sqrt{\sum_{j=1}^p \text{cor}^2(\mathbf{x}_{ij}, \mathbf{X}_i \mathbf{w}_i)} \times \sqrt{\sum_{j=1}^p \text{cor}^2(\mathbf{x}_{lj}, \mathbf{X}_l \mathbf{w}_l)} \end{aligned}$$

Therefore, group sizes are taken into account in problem (15) and not in problem (14). Higher dimension solutions will be sought by deflation as explained above.

3.8.1.3. *Situation 3.* Here is considered a super-group $\mathbf{X}_{I+1}$ equal to the row concatenation of the various groups $\mathbf{X}_i$ followed by a normalization: $\mathbf{X}_{I+1} = [\mathbf{X}_1^\top, \dots, \mathbf{X}_I^\top]^\top / \sqrt{I}$. This insures that each column of $\mathbf{X}_{I+1}$ has a norm equal to 1. From this it follows

$$\mathbf{R}_{I+1} = \mathbf{X}_{I+1}^\top \mathbf{X}_{I+1} = \frac{1}{I} [\mathbf{X}_1^\top \mathbf{X}_1 + \dots + \mathbf{X}_I^\top \mathbf{X}_I] = \frac{1}{I} [\mathbf{R}_1 + \dots + \mathbf{R}_I] = \overline{\mathbf{R}}.$$

A situation is considered in which all groups 1 to I are only related to the super-group $\mathbf{X}_{I+1}$: $c_{il} = 1$ for $i = 1, \dots, I$, $l = I + 1$ and 0 otherwise. Matrix $\mathbf{M}_i = \mathbf{X}_i^\top \mathbf{X}_i$ is considered for each group and for the super-group.

For g equals to the square function, optimization problem (13) becomes:

$$(17) \quad \begin{aligned} & \max_{\mathbf{w}_1, \dots, \mathbf{w}_{I+1}} \sum_{i=1} \cos^2(\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_{I+1}^\top \mathbf{X}_{I+1} \mathbf{w}_{I+1}) \times \|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i\|^2 \times \|\mathbf{X}_{I+1}^\top \mathbf{X}_{I+1} \mathbf{w}_{I+1}\|^2 \\ & \text{s.c.} \quad \|\mathbf{X}_i \mathbf{w}_i\| = 1, \quad i = 1, \dots, I + 1 \end{aligned}$$

Optimization problem (17) leads to a compromise between the various one-dimension PCA's of the $\mathbf{X}_i$'s, the one-dimension PCA of $\mathbf{X}_{I+1}$ and small angles between the group-loading vectors and the super-group- loading vector. Due to the chosen normalization, the loading vectors $\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i$ represent the vectors of correlations between the various columns of each

group $\mathbf{X}_i$ and their group component $\mathbf{X}_i \mathbf{w}_i$. The solution of optimization problem (17) and useful properties are given in proposition 1 below.

Proposition 1. Let $\mathbf{w}_1, \dots, \mathbf{w}_{I+1}$ be the solution of optimization problem (17). The following results are obtained:

- (a) The super-group-weight vector $\mathbf{w}_{I+1}$ is the eigenvector of $\mathbb{R}_{I+1}$ associated with the largest eigenvalue.
- (b) The group-weight vector $\mathbf{w}_i$ is equal to $\mathbf{w}_{I+1} / \|\mathbf{X}_i \mathbf{w}_{I+1}\|$.
- (c) The super-group component $\mathbf{t}_{I+1} = \mathbf{X}_{I+1} \mathbf{w}_{I+1}$ is the normalized first principal component of $\mathbf{X}_{I+1}$.
- (d) The group component $\mathbf{t}_i = \mathbf{X}_i \mathbf{w}_i$ is equal to $\mathbf{X}_i \mathbf{w}_{I+1} / \|\mathbf{X}_i \mathbf{w}_{I+1}\|$; this means that $\mathbf{t}_i$ is the normalized fragment of $\mathbf{t}_{I+1}$ related to group i .
- (e) The cosine between the group-loading vector $\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i$ and the super-group-loading vector $\mathbf{X}_{I+1}^\top \mathbf{X}_{I+1} \mathbf{w}_{I+1}$ is positive. This means that all group-loading vectors are pointing out in the direction of the super-group-loading vector.

Remarks

- (1) It is quite remarkable that optimization problem (17) produces group-weight vectors $\mathbf{w}_i$ all proportional to the super-group-weight vector $\mathbf{w}_{I+1}$, although this was not a constraint of the problem.
- (2) Optimization problem (17) has already been considered in Eslami et al. (2013a) [23].
- (3) Optimization problem (17) give the same results for $q = 1$. This property is also verified for $q > 1$ provided that the deflation procedure is carried out in optimization problem (17) on the super-group only.
- (4) If orthogonality is required for higher dimension solutions at group and super-group levels, the RGCCA algorithm can be used.

Chapter 4

Main Contribution to Multivariate Methods

In this chapter, a description is given of two new multivariate methods for multigroup datasets. The first proposal is the Multigroup Principal Component Analysis method that the aim is the maximization of the interdistances between the points of the different groups, and making it more general, can also be expanded to multiblock framework. The second method called Multigroup Dimensionality Reduction (MDR) the idea find out the direction that minimize the overlapping when comparing observations from one group with the rest of observations in other groups.

4.1. Multigroup Principal Component Analysis

Let E be an euclidean space m -dimensional about a space real $\mathbb{R}$. Let $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$ be a base of said space. Let $\mathbf{A} = (< \mathbf{e}_i, \mathbf{e}_j >)_{m \times m}$ be a symmetric and positive definite matrix associated to the scalar product with respect to said base.

Let L be a linear variety r -dimensional with $r < m$ defined by vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$, without loss of generality, can be assumed orthonormal and an arbitrary vector $\mathbf{c} \in E$. So if it is denoted by $F = < \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r >$ can be written $L = \mathbf{c} + F$.

Now, if it consider that $p_1, \dots, p_k$ are grouped in groups of size $n_1, \dots, n_q$ in form:

$$p_1, \dots, p_{n_1}, p_{n_1+1}, \dots, p_{n_1+n_2}, \dots, p_{n_1+\dots+n_{q-1}+1}, \dots, p_k$$

observe that $n_i \geq 1$ and $q \geq 1$.

Let $\mathbf{X}$ be the matrix $k \times m$ whose rows are the components of the p_i with respect to the base $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$; taking into account the q groups can be written in boxes:

$$(18) \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_q \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n_1 1} & \cdots & x_{n_1 m} \\ \vdots & \vdots & \vdots \\ x_{n_1 + \cdots + n_{q-1} + 1 \ 1} & \cdots & x_{n_1 + \cdots + n_{q-1} + 1 \ m} \\ \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{km} \end{pmatrix}$$

The dataset $\mathbf{X}$ consists in the measurements of m variables on n_k the number of individuals in the group where $\mathbf{X}_i$ is the matrix $n_i \times m$ with $\sum_i^q n_i = k$.

The coordinates of the points p_i on the variety $\mathcal{L}$ with respect to the coordinate system induced by $\mathbf{v}_1, \dots, \mathbf{v}_r$ (given $\mathbf{c}$), will be given by the rows of the matrix:

$$(19) \quad (\mathbf{X} - \mathbf{1}_k \mathbf{c}^t) \mathbf{A} (\mathbf{V}_1, \dots, \mathbf{V}_r)$$

where $\mathbf{V}_1, \dots, \mathbf{V}_r$ are the vectors $m \times 1$ whose coefficients are the components of $\mathbf{v}_1, \dots, \mathbf{v}_r$ respect the base $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$ and $\mathbf{1}_k$ is a vector $k \times 1$ whose components are all equal to 1:

$$\mathbf{1}_k = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

and $\mathbf{c}$ is a vector $m \times 1$ whose coefficients are the components of the vector $\mathbf{c}$ respect $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$. If denotes the matrix $\mathbf{V}_{m \times r} = (\mathbf{V}_1, \dots, \mathbf{V}_r)$ it can be written (19) as

$$(\mathbf{X} - \mathbf{1}_k \mathbf{c}^t) \mathbf{A} \mathbf{V}.$$

On the other hand, the euclidean distance induced by $\mathcal{L}$, introducing the following notation:

$$d_{\alpha\beta}^2 \equiv d_{\mathcal{L}}^2(P_\alpha, P_\beta)$$

Also if the vector $k \times 1$ is denoted as

$$\mathbf{e}_\alpha = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

i.e. $(\mathbf{e}_\alpha)_{i1} \equiv \delta_{\alpha i}$ is the Kronecker delta.

Thus, the distance is

$$d_{\alpha\beta}^2 = (\mathbf{e}_\alpha^t(\mathbf{X} - \mathbf{1}_k \mathbf{c}^t) \mathbf{A} \mathbf{V} - \mathbf{e}_\beta^t(\mathbf{X} - \mathbf{1}_k \mathbf{c}^t) \mathbf{A} \mathbf{V}) \cdot (\mathbf{V}^t \mathbf{A}(\mathbf{X} - \mathbf{1}_k \mathbf{c}^t)^t \mathbf{e}_\alpha - \mathbf{V}^t \mathbf{A}(\mathbf{X} - \mathbf{1}_k \mathbf{c}^t)^t \mathbf{e}_\beta)$$

since $\mathbf{v}_1, \dots, \mathbf{v}_r$ form an orthonormal basis of F and therefore:

$$d_{\alpha\beta}^2 = (\mathbf{e}_\alpha - \mathbf{e}_\beta)^t (\mathbf{X} - \mathbf{1}_k \mathbf{c}^t) \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} (\mathbf{X} - \mathbf{1}_k \mathbf{c}^t)^t (\mathbf{e}_\alpha - \mathbf{e}_\beta)$$

it is taken into account that

$$\mathbf{e}_\alpha^t \mathbf{1}_k \mathbf{c}^t - \mathbf{e}_\beta^t \mathbf{1}_k \mathbf{c}^t = 0^t.$$

This will result in the following

$$(20) \quad d_{\alpha\beta}^2 = (\mathbf{e}_\alpha - \mathbf{e}_\beta)^t \mathbf{X} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}^t (\mathbf{e}_\alpha - \mathbf{e}_\beta).$$

Next, consider the following problem of finding the linear variety $\mathcal{L}$ of dimension r such that maximizes the sum of the squares of the interdistances between points of different groups (with $q > 1$).

To express the quantity to be maximized in a simple way, determine the following sets $J_1, \dots, J_q$ defined as

- $\mathbf{J}_1 = \{1, \dots, n_1\}$
- $\mathbf{J}_i = \{n_1 + \dots + n_{i-1} + 1, \dots, n_1 + \dots + n_i\} \quad i = 2, \dots, q$

Therefore, the amount to be maximized is

$$(21) \quad \mathcal{D} = \sum_{\substack{i,j=1 \\ i \neq j}}^q \sum_{\alpha \in \mathbf{J}_i, \beta \in \mathbf{J}_j} d_{\alpha\beta}^2 = \sum_{\alpha=1}^k \sum_{\beta=1}^k d_{\alpha\beta}^2 - \sum_{i=1}^q \sum_{\alpha, \beta \in \mathbf{J}_i} d_{\alpha\beta}^2$$

If we denote as $\mathbf{G} = \mathbf{X} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}^t$ will result

$$\begin{aligned}
\sum_{\alpha=1}^k \sum_{\beta=1}^k d_{\alpha\beta}^2 &= \sum_{\alpha=1}^k \sum_{\beta=1}^k (\mathbf{e}_\alpha - \mathbf{e}_\beta)^t \mathbf{G} (\mathbf{e}_\alpha - \mathbf{e}_\beta) \\
&= \sum_{\alpha=1}^k \sum_{\beta=1}^k \{ \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\alpha + \mathbf{e}_\beta^t \mathbf{G} \mathbf{e}_\beta - \mathbf{e}_\beta^t \mathbf{G} \mathbf{e}_\alpha - \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\beta \} \\
&= 2k \sum_{\alpha=1}^k \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\alpha - 2 \sum_{\alpha=1}^k \sum_{\beta=1}^k \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\beta \\
&= 2k \sum_{\alpha=1}^k \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\alpha - 2 \sum_{\alpha=1}^k \mathbf{e}_\alpha^t \mathbf{G} \left(\sum_{\beta=1}^k \mathbf{e}_\beta \right) \\
&= 2k \sum_{\alpha=1}^k \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\alpha - 2(\mathbf{1}_k^t \mathbf{G} \mathbf{1}_k) \\
&= 2k \sum_{\alpha=1}^k \text{tr}(\mathbf{G} \mathbf{e}_\alpha \mathbf{e}_\alpha^t) - 2(\mathbf{1}_k^t \mathbf{G} \mathbf{1}_k) \\
&= 2k \text{tr} \left(\mathbf{G} \left(\sum_{\alpha=1}^k \mathbf{e}_\alpha \mathbf{e}_\alpha^t \right) \right) - 2(\mathbf{1}_k^t \mathbf{G} \mathbf{1}_k) \\
&= 2k \text{tr}(\mathbf{G}) - 2(\mathbf{1}_k^t \mathbf{G} \mathbf{1}_k)
\end{aligned}$$

Note that as $\mathbf{X}^t \mathbf{1}_k = k \bar{\mathbf{x}}$ it will result

$$\begin{aligned}
\mathbf{1}_k^t \mathbf{G} \mathbf{1}_k &= \mathbf{1}_k^t \mathbf{X} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}^t \mathbf{1}_k \\
&= k^2 \bar{\mathbf{x}}^t \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \bar{\mathbf{x}} \\
&= k^2 \text{tr}(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \bar{\mathbf{x}} \bar{\mathbf{x}}^t)
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{\alpha=1}^k \sum_{\beta=1}^k d_{\alpha\beta}^2 &= 2k^2 \text{tr} \left(\frac{1}{k} \mathbf{G} - \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \bar{\mathbf{x}} \bar{\mathbf{x}}^t \right) \\
&= 2k^2 \text{tr} \left(\frac{1}{k} \mathbf{X} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}^t - \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \bar{\mathbf{x}} \bar{\mathbf{x}}^t \right) \\
&= 2k^2 \text{tr} \left(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \left(\frac{1}{k} \mathbf{X}^t \mathbf{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}^t \right) \right) \\
&= 2k^2 \text{tr}(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{S})
\end{aligned}$$

where $\mathbf{S}$ is the matrix of variances and covariances corresponding to all $\mathbf{X}$ dataset.

On the other hand,

$$\begin{aligned}
 \sum_{\alpha, \beta \in \mathbf{J}_i} d_{\alpha\beta}^2 &= 2n_i \sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\alpha - 2 \sum_{\alpha, \beta \in \mathbf{J}_i} \mathbf{e}_\alpha^t \mathbf{G} \mathbf{e}_\beta \\
 (22) \quad &= 2n_i \text{tr} \left(\mathbf{G} \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \mathbf{e}_\alpha^t \right) \right) - 2 \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \right)^t \mathbf{G} \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \right)
 \end{aligned}$$

Observe that $\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \mathbf{e}_\alpha^t = \text{diag}(0, \dots, 0, 1, \dots, 1, 0, \dots, 0)$ and

$$\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Thus, of equation 22:

$$\begin{aligned}
 \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \right)^t \mathbf{G} \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \right) &= \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \right)^t \mathbf{X} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}^t \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \right) \\
 &= \mathbf{1}_{n_i}^t \mathbf{X}_i \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}_i^t \mathbf{1}_{n_i} \\
 &= n_i^2 \bar{\mathbf{x}}_i^t \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \bar{\mathbf{x}}_i \\
 &= n_i^2 \text{tr}(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^t)
 \end{aligned}$$

where $\bar{\mathbf{x}}_i$ is the vector $m \times 1$ of calculated averages in the group i .

Moreover,

$$\begin{aligned}
tr \left(\mathbf{G} \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \mathbf{e}_\alpha^t \right) \right) &= tr \left(\mathbf{X} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}^t \left(\sum_{\alpha \in \mathbf{J}_i} \mathbf{e}_\alpha \mathbf{e}_\alpha^t \right) \right) \\
&= tr \left(\begin{pmatrix} x_1 \\ \vdots \\ x_q \end{pmatrix} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} (x_1^t, \dots, x_q^t) \begin{pmatrix} 0 & \dots & \dots & 0 \\ & \ddots & & \\ & & \mathbf{I}_i & \\ 0 & \dots & \dots & 0 \end{pmatrix} \right) \\
&= tr \left(\begin{pmatrix} x_1 \\ \vdots \\ x_q \end{pmatrix} \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} (0, \dots, x_i^t, \dots, 0) \right) \\
&= tr(\mathbf{X}_i \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}_i^t)
\end{aligned}$$

Then the equation (22) remains as:

$$\begin{aligned}
\sum_{\alpha, \beta \in \mathbf{J}_i} d_{\alpha\beta}^2 &= 2n_i^2 tr \left(\frac{1}{n_i} \mathbf{X}_i \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{X}_i^t - \mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^t \right) \\
(23) \quad &= 2n_i^2 tr \left(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \left(\frac{1}{n_i} \mathbf{X}_i^t \mathbf{X}_i - \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^t \right) \right) \\
&= 2n_i^2 tr(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{S}_i)
\end{aligned}$$

where $\mathbf{S}_i$ is the matrix of variances and covariances corresponding to $\mathbf{X}_i$ dataset for group i .

Therefore, the equation (21) can be written as:

$$(24) \quad \mathcal{D} = 2k^2 tr(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{S}) - \sum_{i=1}^q 2n_i^2 tr(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \mathbf{S}_i)$$

In addition, it must be taken into account that

$$\mathbf{S} = \frac{1}{k} \sum_{i=1}^q n_i \mathbf{S}_i + \mathbf{B}$$

where

$$\mathbf{B} = \frac{1}{k} \sum_{i=1}^q n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^t$$

Thus, it is obtained that the equation (24) of the interdistance between the points of the different groups to be maximized can be written with the following expression:

$$\begin{aligned}
 (25) \quad \mathcal{D} &= 2tr \left(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \left(k \sum_{i=1}^q n_i \mathbf{S}_i + k^2 \mathbf{B} - \sum_{i=1}^q n_i^2 \mathbf{S}_i \right) \right) \\
 &= 2tr \left(\mathbf{A} \mathbf{V} \mathbf{V}^t \mathbf{A} \left(\sum_{i=1}^q (k - n_i) n_i \mathbf{S}_i + k^2 \mathbf{B} \right) \right)
 \end{aligned}$$

Therefore, the maximization of equation (25) is obtained through the diagonalization of

$$\mathbf{A} \left\{ \sum_{i=1}^q (k - n_i) n_i \mathbf{S}_i + k^2 \mathbf{B} \right\} \mathbf{A} \mathbf{v}_i = \lambda_i \mathbf{A} \mathbf{v}_i$$

with the condition $\mathbf{v}_i^t \mathbf{A} \mathbf{v}_j = \delta_{ij}$. This is equivalent to writing the expression as

$$\begin{aligned}
 \mathbf{A}^{\frac{1}{2}} \left\{ \sum_{i=1}^q (k - n_i) n_i \mathbf{S}_i + k^2 \mathbf{B} \right\} \mathbf{A}^{\frac{1}{2}} \mathbf{w}_i &= \lambda_i \mathbf{A} \mathbf{w}_i \\
 \mathbf{w}_i^t \mathbf{w}_j &= \delta_{ij}
 \end{aligned}$$

that is, implies having to diagonalize the symmetric matrix:

$$(26) \quad \mathbf{A}^{\frac{1}{2}} \left\{ \sum_{i=1}^q (k - n_i) n_i \mathbf{S}_i + k^2 \mathbf{B} \right\} \mathbf{A}^{\frac{1}{2}}$$

and choose the eigenvectors associated with the largest eigenvalues, that is, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \dots \geq \lambda_m \geq 0$ where $\mathbf{w}_1, \dots, \mathbf{w}_r$ are the corresponding eigenvectors.

The percentage of cumulative variability is

$$(27) \quad 100 \times \frac{\lambda_1 + \lambda_2 + \dots + \lambda_r}{tr(\mathbf{A}^{\frac{1}{2}} \left\{ \sum_{i=1}^q (k - n_i) n_i \mathbf{S}_i + k^2 \mathbf{B} \right\} \mathbf{A}^{\frac{1}{2}})}$$

Algorithm 1: Multigroup PCA

Input: Dataset, vector of classes and symmetric and positive definite matrix associated.

Output: New variance-covariance matrix.

Data: Dataset with m variables on n_k the number of individuals in the group where X_i is the matrix $n_i \times m$ with $\sum_i^q n_i = k$.

- 1 **k** $\Rightarrow$ number of rows in the dataset X ;
 - 2 **mean.x** $\Rightarrow$ Compute the vector of mean for dataset X ;
 - 3 **mean.y** $\Rightarrow$ Compute the vector of means of individuals in group q ;
 - 4 **Ss.p** $\Rightarrow$ Compute $S_1, S_2, \dots, S_q$ variance-covariance matrices for each q group;
 - 5 **Bs** $\Rightarrow (mean.y - mean.x) \times t(mean.y - mean.x) \times nrow(y)$ Calculate the variability between groups q ;
 - 6 **B** $\Rightarrow Reduce(", ", Bs)$ Sum elements of list of compute the variability between groups q ;
 - 7 $B = (1/k) \times B \longrightarrow$ The variability between groups;
 - 8 **Ss.all** $\Rightarrow$ Sum elements of list of $Ss.p$;
 - 9 $sqA = sqrtm(A) \longrightarrow$ Calculate the root of matrix A ;
 - 10 **res** $= sqA \times (Ss.all + ((k^2) \times B)) \times sqA \longrightarrow$ Matrix to diagonalize;
-

4.2. Multigroup Dimensionality Reduction

Let X be an $n \times p$ dimensional matrix, with $n = n_1 + \dots + n_g$, n_k denotes the number of data points that belong to k -th group, $k = 1, \dots, g$. Let us assume that $p > g$. Let X_k be the $n_k \times p$ dimensional matrix whose rows come from points in the k -th group.

Let $v \in \mathbb{R}^p$ be a $p \times 1$ dimensional unit vector, that is $\|v\| = 1$. Let

$$d_{ij} = X_i v - \mathbf{1}_{n_i} \frac{1}{n_j} \mathbf{1}_{n_j}^\top X_j v = (X_i - \mathbf{1}_{n_i} \frac{1}{n_j} \mathbf{1}_{n_j}^\top X_j) v$$

be denote the $n_i \times 1$ dimensional vector of differences between the projections of data points from i -th group and the mean of the projections of the data points from j -th group, with $i, j = 1, \dots, g$ and $i \neq j$, where $\mathbf{1}_{n_i}$ denotes the $n_i \times 1$ dimensional vector with 1 at each component.

Observe that, if the projections of almost all observations in the group i fall entirely to the right (or to the left) of the centroid of the projections of the group j , then almost all products $d_{l,ij} d_{s,ij}$, $l, s = 1, \dots, n_i$, will be positive, and then we could conclude that v could be a separating direction of groups i -th and j -th.

Alternatively, if the projections of the observations in the group i falls in between the centroid of the projections of the group j , then d_{ij} will have a different sign and many products $d_{l,ij} d_{s,ij}$, $l, s = 1, \dots, n_i$, should be negatives, then we could conclude that v is not a separating direction of groups i -th and j -th.

A natural criteria for choosing v is to maximize the sum

$$S_{ij} = \sum_{l=1}^{n_i} \sum_{s=1}^{n_i} d_{l,ij} d_{s,ij}$$

This criterion penalizes that the differences in d_{ij} have a different sign, on the contrary, enhances that the majority of the differences have the same sign and the greater they are.

We can write

$$S_{ij} = \mathbf{1}_{n_i}^\top (X_i - \mathbf{1}_{n_i} \frac{1}{n_j} \mathbf{1}_{n_j}^\top X_j) v v^\top (X_i - \mathbf{1}_{n_i} \frac{1}{n_j} \mathbf{1}_{n_j}^\top X_j)^\top \mathbf{1}_{n_i}$$

and setting $z_{ij} = (X_i - \mathbf{1}_{n_i} \frac{1}{n_j} \mathbf{1}_{n_j}^\top X_j)^\top \mathbf{1}_{n_i}$, we also write

$$S_{ij} = z_{ij}^\top v v^\top z_{ij}$$

This approach suggests that the vector v will tend to order the projections, whatever the group j (from which the centroid is computed), it is pursued that the projections of the observations from group i don't overlap with the centroid of the projections of the observations

of the group j . Then the objective function is

$$S = \sum_{i=1}^{g-1} \sum_{j>i}^g S_{ij} = \sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij}^T v v^T z_{ij}$$

We can express the problem that we want to solve in the following way

$$\begin{aligned} & \max_v \quad S \\ & \text{subject to} \quad v^T v = 1 \end{aligned}$$

whose solution has been found.

Applying the Lagrange multiplier technique we obtain the Lagrangian

$$L = \sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij}^T v v^T z_{ij} - \lambda(v^T v - 1)$$

Differentiating with respect to v_α , $\alpha = 1, \dots, p$, we have

$$\begin{aligned} \frac{\partial L}{\partial v_\alpha} &= \sum_{i=1}^{g-1} \sum_{j>i}^g \frac{\partial S_{ij}}{\partial v_\alpha} - 2\lambda v_\alpha \\ &= \sum_{i=1}^{g-1} \sum_{j>i}^g \frac{\partial(z_{ij}^T v v^T z_{ij})}{\partial v_\alpha} - 2\lambda v_\alpha \\ &= \sum_{i=1}^{g-1} \sum_{j>i}^g \frac{\partial(z_{ij}^T v)}{\partial v_\alpha} v^T z_{ij} + z_{ij}^T v \frac{\partial(v^T z_{ij})}{\partial v_\alpha} - 2\lambda v_\alpha \\ &= \sum_{i=1}^{g-1} \sum_{j>i}^g z_{\alpha,ij} v^T z_{ij} + z_{ij}^T v z_{\alpha,ij} - 2\lambda v_\alpha \\ &= \sum_{i=1}^{g-1} \sum_{j>i}^g 2z_{\alpha,ij} v^T z_{ij} - 2\lambda v_\alpha \end{aligned}$$

Thus

$$\frac{\partial L}{\partial v} = \sum_{i=1}^{g-1} \sum_{j>i}^g 2z_{ij} z_{ij}^T v - 2\lambda v$$

Then, we have the equation

$$(28) \quad \left(\sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij} z_{ij}^T \right) v = \lambda v$$

Then, solutions are the eigenvector of the symmetric matrix $\sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij} z_{ij}^T$.

As we will see, if v denotes the solution of equation (28) associated with λ , then the value of S is λ ;

$$\begin{aligned} S &= \sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij}^T v v^T z_{ij} \\ &= \sum_{i=1}^{g-1} \sum_{j>i}^g v^T z_{ij} z_{ij}^T v \\ &= v^T \left(\sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij} z_{ij}^T \right) v \\ &= v^T (\lambda v) = \lambda (v^T v) \\ &= \lambda \end{aligned}$$

Then, to maximize S we have to take the eigenvector associated with the highest eigenvalue.

Implementing the same criterion we can find a second vector v_2 orthogonal to v_1 . We define

$$\begin{aligned} S_{ij}^{(1)} &= z_{ij}^T v_1 v_1^T z_{ij} \\ S_{ij}^{(2)} &= z_{ij}^T v_2 v_2^T z_{ij} \end{aligned}$$

And,

$$\begin{aligned} S_1 &= \sum_{i=1}^{g-1} \sum_{j>i}^g S_{ij}^{(1)} \\ S_2 &= \sum_{i=1}^{g-1} \sum_{j>i}^g S_{ij}^{(2)} \end{aligned}$$

We want to maximize

$$S = S_1 + S_2$$

subject to $v_1^T v_1 = 1$ and $v_2^T v_2 = 1$.

The Lagrange function is

$$L = S_1 + S_2 - \lambda_1(v_1^\top v_1 - 1) - \lambda_2(v_2^\top v_2 - 1)$$

Since L is a function of v_1 and v_2 that can be expressed as a sum of two functions depending separately each one on v_1 or v_2 , we have the equations

$$\begin{aligned} \left(\sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij} z_{ij}^\top \right) v_1 &= \lambda_1 v_1 \\ \left(\sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij} z_{ij}^\top \right) v_2 &= \lambda_2 v_2 \end{aligned}$$

Then, v_1 and v_2 are eigenvectors of $\sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij} z_{ij}^\top$ associated with two leading eigenvalues.

Observe that, by construction, the matrices $z_{ij} z_{ij}^\top$ in (28) are $p \times p$ dimensional matrices with rank 1. Then the matrix $\sum_{i=1}^{g-1} \sum_{j>i}^g z_{ij} z_{ij}^\top$, which results from the sum of $g(g-1)/2$ symmetric matrices having rank 1, will have the rank at most equal to $g(g-1)/2$. If none of the vectors are collinear, then the rank of the array, which coincides with the number of nonzero eigenvalues, will be $g(g-1)/2$ and in case there are collinear vectors the rank will be lower. In general, there will be multiple solutions of the equation (28) that will provide us orthogonal vectors with which to define the axes of the representation in low dimension.

Algorithm 2: Multigroup Dimensionality Reduction (MDR)

Input : Factor groups, dataset, maximum dimension, plot**Output:** List of 3 elements, y , of the MDR method.**Data** : Let X dataset $n \times p$ dimensional, with $n = n_1 + n_2 + \dots + n_g$ where n_k denotes the number of data points that belong to k -th group for $k = 1, \dots, g$. Dataset X_k be the $n_k \times p$ dimensional matrix whose rows come from points in the k -th group.

```

1  $g \leftarrow$  number of groups;
2 for  $k \leftarrow 1$  to  $g$  do
3    $mg[[k]] \Rightarrow$  list of data matrix by group;
4    $ln1[[k]] \Rightarrow$  list of vector of ones;
5 for  $w \leftarrow 1$  to  $g$  do
6   for  $r \leftarrow 1$  to  $g$  do
7     if  $w > r$  then
8        $/*$  All differences between the projections of data points from i-th group
9       and the mean of the projections of the data points from j-th group  $*/$ 
10       $A[[s]] \Rightarrow$  list of matrices;
11       $s = s + 1;$ 
12    else
13       $W=1;$ 
14    for  $w \leftarrow 1$  to  $g$  do
15      for  $r \leftarrow 1$  to  $g$  do
16        if  $w > r$  then
17           $z[[W]] \Rightarrow$  list of vector of differences the projections ;
18           $W = W + 1;$ 
19        else
20           $C[[q]] \Rightarrow$  list of symmetric matrices;
21           $CC \Rightarrow$  Sum elements of list of  $C$ 
22           $v \Rightarrow$  Eigenvectors;
23          for  $p \leftarrow 1$  to  $g$  do
24             $ys \Rightarrow$  list
25          return  $y$ 

```

4.2.1. Multigroup Single Dimension Exploration. We propose a modification of the MDR method for data exploration. For each group we will look for a vector in sample space that minimizes the overlapping of the projected points of that group and the projected points in the rest of groups. In such a way that, for each one of the groups we obtain a particular direction. We not ask that the set of directions define an orthogonal axis system that allow us to represent the data in reduced dimension. Instead, we will analyze the data in each of the dimensions separately. We can statistically analyze the projections to acquire knowledge about the differences between each group and the other groups.

For example, given the vector determined by the i -th group, we can calculate the sample correlation coefficient of each variable with the projection values in that vector. Therefore, we can range the input variables according the correlation with the i -th vector and characterize the i -th group. In the context of gene expression studies we can use the correlation values for gene set enrichment analysis.

4.2.1.1. Single Dimension Exploration. Let X be an $n \times p$ dimensional matrix, with $n = n_1 + \dots + n_g$, n_k denotes the number of data points that belong to k -th group, $k = 1, \dots, g$. Let us assume that $p > g$. Let X_k be the $n_k \times p$ dimensional matrix whose rows come from points in the k -th group.

Let $v \in \mathbb{R}^p$ be a $p \times 1$ dimensional unit vector, that is $\|v\| = 1$. Let

$$d_i = X_i v - \mathbf{1}_{n_i} \frac{1}{n_{-i}} \mathbf{1}_{n_{-i}}^\top X_{-i} v = (X_i - \mathbf{1}_{n_i} \frac{1}{n_{-i}} \mathbf{1}_{n_{-i}}^\top X_{-i}) v$$

be denote the $n_i \times 1$ dimensional vector of differences between the projections of data points from i -th group and the mean of the projections of the rest of the data points not in i -th group, with $i \in \{1, \dots, g\}$, where $\mathbf{1}_{n_k}$ denotes the $n_k \times 1$ dimensional vector with 1 at each component, $\mathbf{1}_{n_{-i}}$ denotes the $n_{-i} \times 1$ dimensional vector with 1 at each component, being $n_{-i} = n - n_i$ and X_{-i} is the $n_{-i} \times p$ dimensional matrix that is obtained when we remove from X the rows from i -th group.

A natural criteria for choosing v is to maximize the sum

$$S_i = \sum_{l=1}^{n_i} \sum_{s=1}^{n_i} d_{l,i} d_{s,i}$$

This criterion penalizes that the differences in d_i have a different sign, on the contrary, enhances that the majority of the differences have the same sign and the greater they are.

We can write

$$S_i = \mathbf{1}_{n_i}^\top (X_i - \mathbf{1}_{n_i} \frac{1}{n_{-i}} \mathbf{1}_{n_{-i}}^\top X_{-i}) v v^\top (X_i - \mathbf{1}_{n_i} \frac{1}{n_{-i}} \mathbf{1}_{n_{-i}}^\top X_{-i})^\top \mathbf{1}_{n_i}$$

and setting

$$z_i = (X_i - \mathbf{1}_{n_i} \frac{1}{n_{-i}} \mathbf{1}_{n_{-i}}^\top X_{-i})^\top \mathbf{1}_{n_i},$$

We also write

$$S_i = z_i^\top v v^\top z_i$$

This approach suggests that the vector v will tend to order the projections, it is pursued that the projections of the observations from group i -th don't overlap with the centroid of the projections of the observations outside of the group i -th. We can express the problem to solve in the following way

$$\begin{aligned} \max_v \quad & S_i \\ \text{subject to} \quad & v^\top v = 1 \end{aligned}$$

Alternatively, we can write

$$S_i = v^\top z_i z_i^\top v$$

Therefore, the problem we need to solve is to maximize a quadratic form with symmetric and rank 1 associated matrix. Thus the solution will be the vector $v = \frac{z_i}{\|z_i\|}$.

Algorithm 3: Single Dimension Exploration (SDE)

Input : Dataset and vector factor groups.

Output: A list of lists of g elements, Y , of the SDE method.

Data : Let X dataset $n \times p$ dimensional, with $n = n_1 + n_2 + \dots + n_g$ where n_k denotes the number of data points that belong to k - th group for $k = 1, \dots, g$. Assume that $p > g$. Dataset X_k be the $n_k \times p$ dimensional matrix whose rows come from points in the k - th group.

```
1 for  $k \leftarrow 1$  to  $g$  do
2    $X[[k]] \Rightarrow$  Element of  $k$  -  $th$  group of list data matrix  $X$ ;
3    $Xn[[k]] \Rightarrow$  Element of list, not in  $k$  -  $th$  group , of data matrix  $X$ ;
4    $n[k] \Rightarrow$  Number of row in data matrix  $k$  group;
5    $nn[k] \Rightarrow$  Number of row in data matrix different  $k$  group;
6    $u[[k]] \Rightarrow$  Element of list, vector of one  $n_k \times 1$  dimensional;
7    $un[[k]] \Rightarrow$  Element of list, vector of one  $nn_k \times 1$  dimensional;
   /* the projections of data points from  $k$  -  $th$  group and the mean of the projections
      of the rest of the data points not in  $k$  -  $th$  group */
8    $A[[k]] \Rightarrow$  List of matrices;
9    $z[[k]] \Rightarrow$  List of vector of differences the projections;
10   $B[[k]] \Rightarrow$  List of symmetric matrices  $p \times p$  dimensional;
11   $BB \Rightarrow$  Matrix of each component of list  $B$ ;
12   $Y \Rightarrow$  A list of lists;
13 return  $Y$ 
```

Chapter 5

MultiGroupO Package

The idea of creating a package is to organize and present the most dynamic information and better understand the statistical methods proposed in this thesis to the scientific community.

The **MultiGroupO** package has all the algorithms worked on this thesis, and also, it has vignettes to facilitate the understanding to the user of the new statistical methodologies in the multivariate analysis and high-dimensional data. These techniques implemented in the package are under an unsupervised and linear approach.

5.1. R Package

Two method new of multigroup and simulation of data. The first technique called multigroup PCA (mgPCA) this multivariate exploration approach that has the idea of considering the structure of groups and/or different types of variables. On the other hand, the second multivariate technique called Multigroup Dimensionality Reduction (MDR) it is another multivariate exploration method which seeks to find out the direction that minimizes overlap when comparing the observations in one group with the rest of the observations in other groups. In addition, a method called Single Dimension Exploration (SDE) was incorporated for to analyze the exploration of the data. It could help us in a better way to observe the behavior of the multigroup data with certain variables of interest.

The **MultiGroupO** package is available under the general public license (GPL-3) from the Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/package=MultiGroupO>, and can be installed in R the following code:

```
#install.packages('MultiGroupO')  
library('MultiGroupO')
```

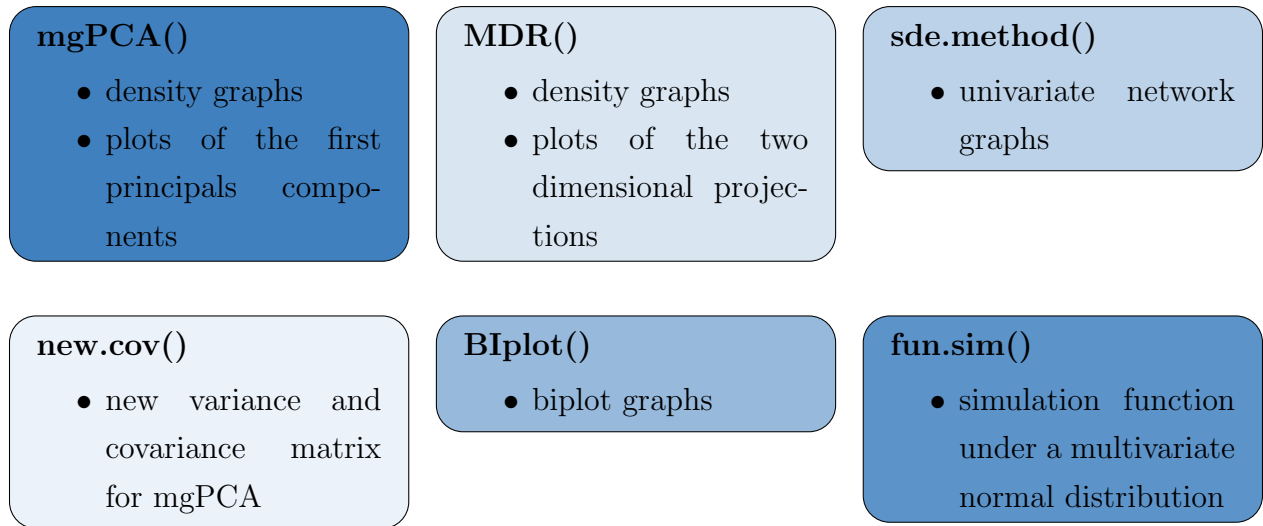


FIG. 1. Structure of the **MultiGroup0** package: the implemented functions of the methods proposed in this thesis.

The figure 1 shows the functions that are in the **MultiGroup0** package, which are the implemented algorithms proposed in chapter 4. This package provides visualization and exploration of high-dimensionality data in multivariate analysis, mainly in the area of biomedicine.

5.1.1. Real Dataset. The real data is from the **plsgenomics** package, in which it describes the dataset as gene expression data from the microarray experiments of Small Round Blue Cell Tumors (SRBCT) of childhood cancer study of Khan et al. (2001). This data set contains 83 samples with 2308 genes: 29 cases of Ewing sarcoma (EWS), coded 1, 11 cases of Burkitt lymphoma (BL), coded 2, 18 cases of neuroblastoma (NB), coded 3, 25 cases of rhabdomyosarcoma (RMS), coded 4. A total of 63 training samples and 25 test samples are provided in Khan et al. (2001). Five of the test set are non-SRBCT and are not considered here. The training sample indexes correspond to 1:65 and the test sample indexes (without non-SRBCT sample) correspond to 66:83.

This real dataset was applied to the methods: **multigroup pca (mgPCA)** and **multigroup dimensionality reduction (MDR)** proposed in this thesis. It was also applied to the **standard PCA**.

Here is the **standard pca** method:

```
library(plsgenomics)
library(MultiGroup0)
data(SRBCT)
```

```
mydata<-SRBCT$X
groups<-as.factor(SRBCT$Y)
pca(datos=mydata,grupos=groups,Plot=TRUE,center=TRUE,scale=FALSE)
```

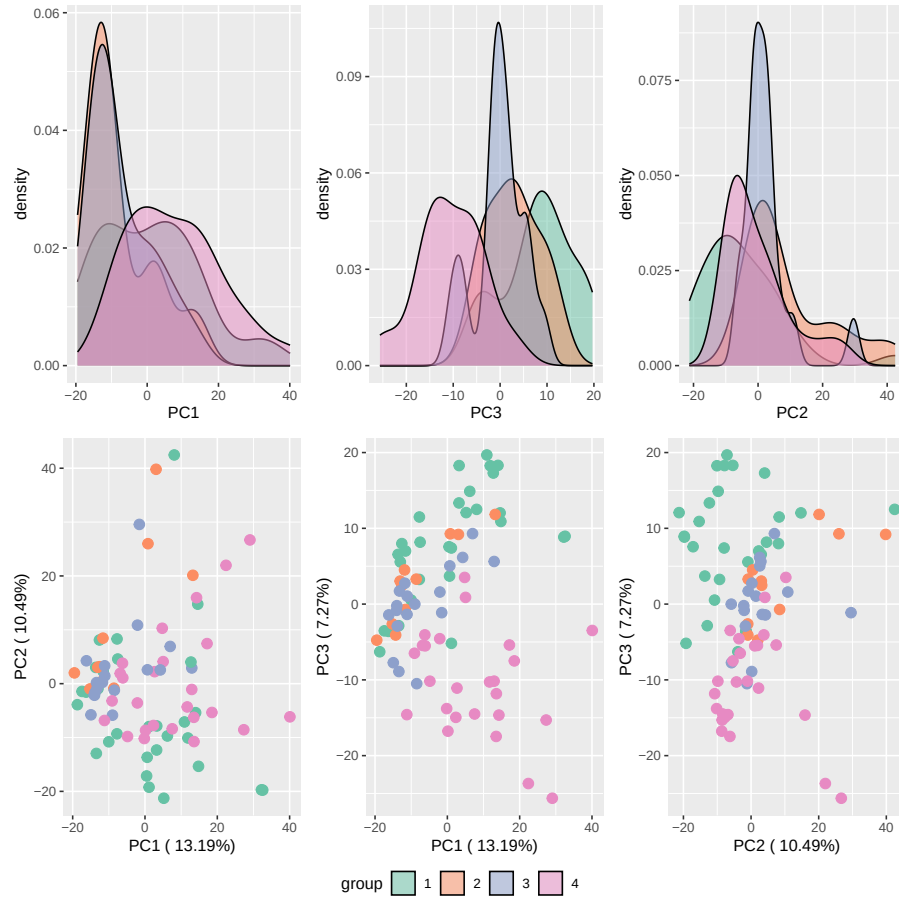


FIG. 2. Output by method of the function `pca()` using SRBCT dataset in `plsgenomics` package.

The visualization of graph by the standard PCA method show to can not separated well the clusters (tumour type) with the first two principal components and the points (genes samples) are more dispersed in the figure 2.

For the proposed **multigroup pca (mgPCA)** method, the following output is shown:

```
require(formatR)
mat.to.diag1 <- new.cov(x = mydata, cls = groups, A = diag(ncol(mydata)))
mgpca(mat.to.diag = mat.to.diag1, mat.x = as.matrix(mydata),
      cls = groups, Plot = TRUE, ncomp = 2, center = TRUE, scale = TRUE)
```

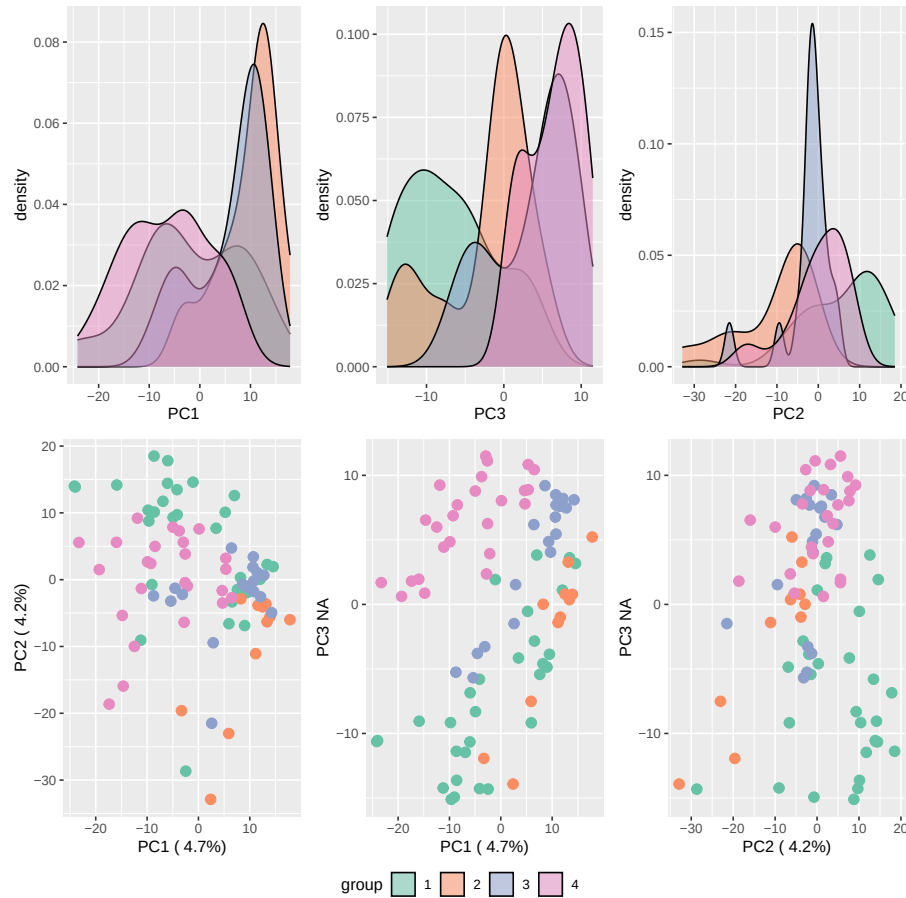


FIG. 3. Output by method of the function `mgPCA()` using SRBCT dataset in `plsgenomics` package.

Our multigroup PCA method that is implemented in the function `mgPCA()` shows us in the figure 3 that can separated better the clusters across the first two principal components than standard PCA method.

The **multigroup dimensionality reduction (MDR)** has the following output:

```
mdr(group=groups, data.x=mydata, c=2)
```

In figure 4, our MDR method implemented in the function `MDR()` of proposed `MultiGroup0` package shows us that separated well clusters with two dimensions and the points are close within of the groups, except the cluster 2 (Burkitt lymphoma (BL)) that has more variability between the points (individuals samples).

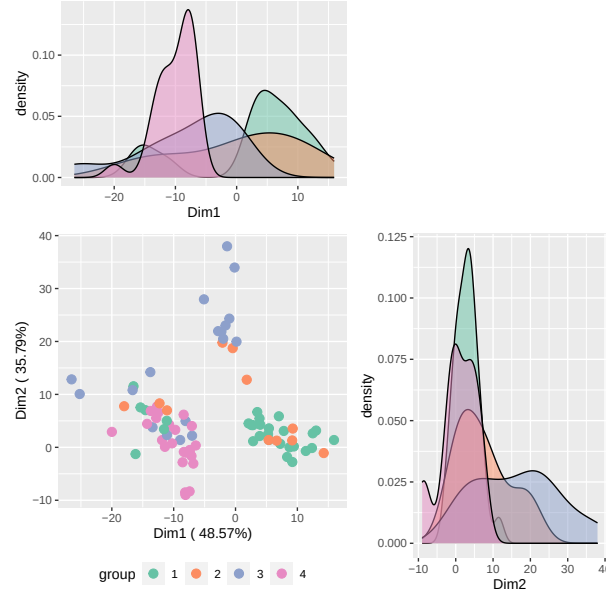


FIG. 4. Output by method of the function `MDR()` using SRBCT dataset in `plsgenomics` package.

5.1.2. Simulation function. The simulation function `fun.sim()` allows us to validate and evaluate the statistical methodologies proposed in this thesis and that have been developed and implemented in the `MultiGroup0` package. This function is defined as follows:

`fun.sim(g, mean1, d, n.var, sds2, corr)`

Arguments	
<i>g</i>	A vector for size of each group
<i>mean1</i>	A vector of population means structure
<i>d</i>	Distance <i>d</i> associated with population mean structure
<i>n.var</i>	2×1 dimension vector associated with dataset structure
<i>sds2</i>	Vector of the variances of groups
<i>corr</i>	Vector of correlation associated with each group

TABLE 5.1. Arguments of the simulation function of quantitative multigroup data under a multivariate normal distribution.

An example of the simulation function described in the table 5.1 is the following:

```
require(base)
require(MultiGroup0)
simList <- fun.sim(g = c(5, 10), mean1 = 2, d = 0, sds2 = c(2,
  4, 1), corr = c(0.5, 0.5, 0), n.var = c(7, 5))
```

The idea of the function is to simulate a data set

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}_{15 \times 12}$$

where have two groups and each of them have 5 and 10 observations, respectively, and the number of variables is 12 by the function `fun.sim()`. These variables, 7 have biological signal and 5 variables noise under a multivariate normal distribution. These variables, 7 have biological signal and 5 variables noise under a normal multivariate

The output of the simulation function gives us the simulated data matrix $\mathbf{X}_1$ of order 5×12 under a multivariate normal for first group:

```
simList$slisx[[1]]
```

##		V1	V2	V3	V4	V5	V6	V7
## 1		3.759057	4.2013268	4.7826183	1.9173354	5.1908317	5.6144988	3.25612571
## 2		1.307670	-0.2928946	1.3913962	1.5446127	1.1776332	2.9045423	1.56574413
## 3		1.763361	3.7535541	2.4065526	3.0583788	2.5088053	0.7176376	2.11575888
## 4		2.647546	4.2319350	2.7058426	2.2060390	1.9085591	2.1407169	1.80621370
## 5		2.155506	3.0361131	0.9493491	0.9928214	0.7337421	1.2603689	0.09726324
##		V8	V9	V10	V11	V12		
## 1		0.7528606	-0.5373580	0.5908362	-1.6784722	-0.6695191		
## 2		-1.4234026	0.1216406	-1.5913039	-0.2040965	-0.5038735		
## 3		1.3897623	0.2624373	-0.7782212	0.4061528	0.7191360		
## 4		0.6170223	1.7270114	0.1065836	-1.2353417	-1.0320205		
## 5		0.1447799	0.2899469	-1.1391845	0.5704799	0.8700978		

For second group, the simulation function gives us the simulated data matrix $\mathbf{X}_2$ of order 10×12 under a multivariate normal:

```
simList$slisx[[2]]
```

##		V1	V2	V3	V4	V5	V6	V7
## 6		0.7752979	1.4991006	4.0648327	0.8886554	0.2466033	1.63566069	3.0840216
## 7		5.2634814	2.6011960	4.7399455	0.9439288	2.9788025	5.28529103	3.5289593
## 8		2.4744191	3.6154862	5.5904213	1.8961370	4.0033288	1.65568122	4.1525921
## 9		2.9466944	1.5028537	4.3273817	3.8321056	1.2634796	3.55185377	4.2798952
## 10		2.4386214	3.3280983	2.8365354	3.4854572	4.3236701	3.15383313	2.1273140
## 11		2.9675637	2.0405305	0.7938322	3.4272147	5.3096527	0.02487396	3.3641560
## 12		2.1469187	1.1669303	4.2820890	4.8556246	4.7088043	3.55381868	5.7967657
## 13		1.7932771	2.9364307	2.0247033	4.4409759	2.1007672	-0.52220808	2.6743597

```
## 14  1.4157344 -1.4331043 0.2636505  1.0453784 3.1605802 -0.88707051  2.3455483
## 15 -1.7774920 -0.4396925 0.9330775 -1.6231762 0.5870846 -2.86624269 -0.3880846
##           V8           V9           V10           V11           V12
## 6    0.68659240  1.01403851  0.58933187  0.4396519  0.7349312
## 7    1.28510182 -0.30292560 -0.61040044 -0.2203358  1.3791266
## 8    -1.25237163  0.87693819  0.88145022 -1.5855831  0.1845668
## 9     0.25160622  1.51045500 -0.38392613  0.0584346  0.6793156
## 10   -0.93029306 -0.38863259  0.04336037  1.0953100  1.4722134
## 11   -1.09755949 -1.02154893 -0.52545508  0.7758873  0.6995225
## 12    0.16002044  0.34655082 -0.81033167  0.6533931 -0.5814139
## 13   -0.09914291 -0.09447033 -0.84345088 -1.0744276 -0.3622101
## 14    0.37448973 -0.62663551  1.42807635 -0.7848811  0.4046244
## 15   -0.63424736  1.74851588  0.12143461  0.1999020  0.3425511
```

The next output is the factor group of two levels (balanced or unbalanced) :

```
simList$grp
##  [1] 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
## Levels: 1 2
```

Then the simulation function `fum.sim()` concatenates by row of the two simulated matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ forming the $\mathbf{X}$ matrix of simulated data 5.1.2 with multigroup structure.

Then we have the output of the vectors of means for each group. A structure of means was simulated under an arithmetic progression with a distance d and a starting mean called `mean1` in the function `fum.sim()`. Thus, the vectors de means μ_1 y μ_2 of order 1×7 can be extracted with the following code:

```
simList$mu[[1]]
## [1] 2 2 2 2 2 2 2

simList$mu[[2]]
## [1] 2 2 2 2 2 2 2
```

The variance and covariance matrices for each group Σ_1 and Σ_2 by following code:

```
head(simList$discovar[[1]])
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
```

```
## [1,] 2 1 1 1 1 1 1
## [2,] 1 2 1 1 1 1 1
## [3,] 1 1 2 1 1 1 1
## [4,] 1 1 1 2 1 1 1
## [5,] 1 1 1 1 2 1 1
## [6,] 1 1 1 1 1 2 1
```

```
head(simList$discovar[[2]])
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,] 4 2 2 2 2 2 2
## [2,] 2 4 2 2 2 2 2
## [3,] 2 2 4 2 2 2 2
## [4,] 2 2 2 4 2 2 2
## [5,] 2 2 2 2 4 2 2
## [6,] 2 2 2 2 2 4 2
```

Finally, for the noise variables in the simulation function `fun.sim()` under a multivariate normal distribution with variance and covariance matrix $\Sigma = \mathbf{I}$. Then, we have to simulate a dataset of variable noises such as $\mathbf{X}_l^* \sim \mathcal{N}_5(\mathbf{0}_{1 \times 5}, \Sigma = \mathbf{I}_{5 \times 5})$ where $l = 1, \dots, 12$. Thus, the vector of means for noise variables can be extracted as:

```
head(simList$mu2)
```

```
## [1] 0 0 0 0 0
```

The simulated data matrix $\mathbf{X}$ is a concatenation by column of the matrix $\mathbf{X}_i$ where $i = 1, 2$ with group structure and biological signal variables, and the matrix $\mathbf{X}_l^*$ without group structure with noise variables.

Finally, the final dataset simulate by the `fun.sim()` function show following output of the matrix 5.1.2:

```
head(simList$x[, -13])
```

```
##      V1      V2      V3      V4      V5      V6      V7
## 1 3.7590571 4.2013268 4.7826183 1.9173354 5.1908317 5.6144988 3.25612571
## 2 1.3076699 -0.2928946 1.3913962 1.5446127 1.1776332 2.9045423 1.56574413
## 3 1.7633610 3.7535541 2.4065526 3.0583788 2.5088053 0.7176376 2.11575888
## 4 2.6475455 4.2319350 2.7058426 2.2060390 1.9085591 2.1407169 1.80621370
## 5 2.1555062 3.0361131 0.9493491 0.9928214 0.7337421 1.2603689 0.09726324
```

```
## 6 0.7752979 1.4991006 4.0648327 0.8886554 0.2466033 1.6356607 3.08402158
##          V8          V9          V10          V11          V12
## 1 0.7528606 -0.5373580 0.5908362 -1.6784722 -0.6695191
## 2 -1.4234026 0.1216406 -1.5913039 -0.2040965 -0.5038735
## 3 1.3897623 0.2624373 -0.7782212 0.4061528 0.7191360
## 4 0.6170223 1.7270114 0.1065836 -1.2353417 -1.0320205
## 5 0.1447799 0.2899469 -1.1391845 0.5704799 0.8700978
## 6 0.6865924 1.0140385 0.5893319 0.4396519 0.7349312
```


Chapter 6

Simulation Study and Omic Real Data

The study of the simulation of data in statistical methodologies helps us to observe the parameters that we want to control, and also, allows us in a parsimonious way to choose a method than others.

Data simulation is a very important tool in statistical methodologies, since we can explore, observe and control the scenarios to be studied and check the validation of the new statistical methods proposed in scientific research.

Simulation studies have had an important boom in the last decade thanks to the advancement of computers. This relationship helps us to improve and/or apply statistical methodologies in different areas in real life. Mainly, in the area of biomedicine, the study of data simulation allows us to observe and better understand the nature of data and try to control experimental studies in health science.

In addition, data simulation allows us to reduce costs and be able to study on a large scale by performing simulations in experimental trials, without the need to use real subjects or depend on real data, but simulating the data to what can happen in real life. This is a great advantage as a scientific methodology in the science of biostatistics because it helps us to obtain advances in studies more quickly and efficiently using computers.

This chapter describes the real omic data that were used for the methods proposed in this investigation on the high dimensionality of data sets. These omic data are of great interest in the area of biomedicine, since there is a peak in this 21st century to investigate cell lines, genes, microRNA, proteins, etc. for complex diseases such as cancer, neurological, infections and others. It is interesting to study the groups that can be formed in the data set and to implement and give ideas about related techniques in the formation of clusters.

6.1. The Data Simulation Function

The first part of simulation is generate quantitative variables under a multivariate normal distribution, i.e., $\mathbf{X}_q \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ where q are groups in datasets. In addition, generate in the simulation quantitative variables with noise, that is, $\mathbf{X}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$ where $k = n_1 + \dots + n_q$ are total observations the all q groups of the datasets and n_i are number of observations by group with $i = 1, \dots, q$. Then in the data simulation function the variables with signal and noise variables are concatenated and form a large data matrix $\mathbf{X}_{k \times m}$ where m is number of variables of datasets. Then to this matrix $\mathbf{X}_{k \times m}$ the standard PCA, multigroup PCA and MDR methods are applied.

For the multigroup PCA proposed in this thesis, the variance and covariance matrix, that is, $\boldsymbol{\Sigma}_i$, is calculated for each group. Therefore, this method takes into account the variability within the i groups and the variability between the i groups.

The second methods proposed in this thesis, which is the MDR, uses the projections between the centroid of one group i and the observations of another group j where $i \neq j$.

One way to explain this simulation function is to give us an example and describe in a simple mathematical way how is our data simulation function.

- Let a dataset is simulated under a normal multivariate distribution as:

(1) **Variables with signal:**

- $\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and $\mathbf{X}_3 \sim \mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$.
- The mean vectors: $\boldsymbol{\mu}_1 = (2, 0, \dots, 0)_{1 \times 100}$, $\boldsymbol{\mu}_2 = (0, 2, 0, \dots, 0)_{1 \times 100}$ and $\boldsymbol{\mu}_3 = (1, 0, 2, -2, 0, \dots, 0)_{1 \times 100}$.
- Given a variance and correlation the matrix of variances and covariances is constructed:

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.5 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 0.5 & 1 & \cdots & 0.5 \\ \vdots & & & \ddots & \\ 0.5 & 0.5 & 0.5 & \cdots & 1 \end{pmatrix}_{100 \times 100}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.2 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 0.2 & 1 & \cdots & 0.2 \\ \vdots & & & \ddots & \\ 0.2 & 0.2 & 0.2 & \cdots & 1 \end{pmatrix}_{100 \times 100}$$

$$\Sigma_3 = \begin{pmatrix} 1 & 0.1 & 0.1 & \cdots & 0.1 \\ 0.1 & 1 & 0.1 & \cdots & 0.1 \\ 0.1 & 0.1 & 1 & \cdots & 0.1 \\ \vdots & & & \ddots & \\ 0.1 & 0.1 & 0.1 & \cdots & 1 \end{pmatrix}_{100 \times 100}$$

(2) **Noise variables:**

- $\mathbf{X}^* \sim \mathcal{N}(\boldsymbol{\mu}^*, \Sigma^*)$
- The mean vector: $\boldsymbol{\mu}^* = (0, 0, \dots, 0)_{1 \times 10}$
- The variance and covariance matrix is the identity matrix:

$$\Sigma^* = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{10 \times 10}$$

- Then the variables with signal and the noise variables in a given matrix $\mathbf{X}$ are concatenated as:

$$\mathbf{X} = \left(\begin{array}{c|c} \begin{matrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{matrix} & \mathbf{X}^* \end{array} \right)_{90 \times 110} = \left(\begin{array}{c|c} \begin{matrix} \mathbf{X}_{30 \times 100} \\ \mathbf{X}_{30 \times 100} \\ \mathbf{X}_{30 \times 100} \end{matrix} & \mathbf{X}_{90 \times 10}^* \end{array} \right)_{90 \times 110}$$

Then we present the following graphs with the methods incorporated in our `Multigroup0` package: PCA standard, new Multigroup PCA and MDR that is demonstrated in this thesis.

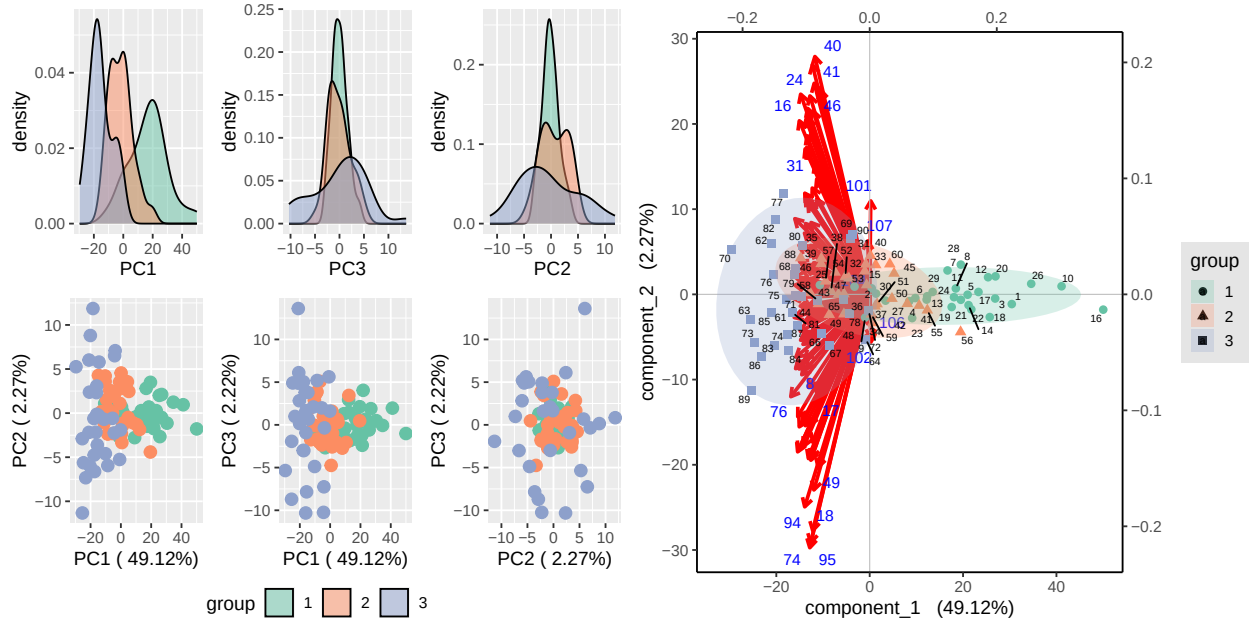


FIG. 1. PCA with 30 observations and correlation (0.5,0.2,0.1) for each group, respectively; in the upper and lower triangular. Different variances for each group (3,3,5). The variables with signal are 100 and the variable noise are 10. Graph Biplot PCA.

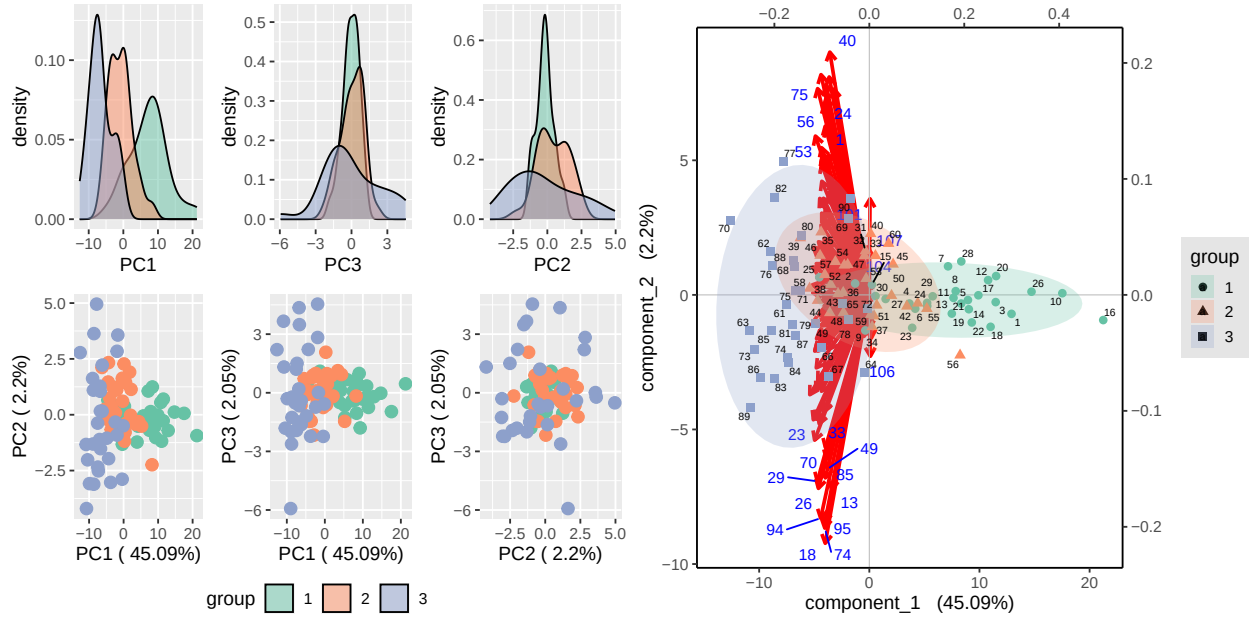


FIG. 2. Multigroup PCA with 30 observations and correlation (0.5,0.2,0.1) for each group, respectively; in the upper and lower triangular. Different variances for each group (3,3,5). The variables with signal are 100 and the variable noise are 10. Graph Biplot Multigroup PCA.

Algorithm 4: Pseudo code of `fun.sim()` algorithm for generating datasets under a multivariate normal distribution.

Input : Number of groups, Vector of the population means, Distance d for the structure of population means, Number of random variables and noise variables, Vector of the variances to simulate for each group and noise, Vector of the correlation to simulate for each group and noise variables

Output: Dataset X_k , Factor group, Dataset $X_{n \times p}$, Vector of means, Matrices of variance-covariance by group

Data : Let X dataset $n \times p$ dimensional, with $n = n_1 + n_2 + \dots + n_g$ where n_k denotes the number of data points that belong to k -th group for $k = 1, \dots, g$. Dataset X_k be the $n_k \times p$ dimensional matrix whose rows come from points in the k -th group.

```

1 for j ← 1 to length(g) do
2   if j = 1 then
3     means[j] ← mean1;
4     // Mean start
5   else
6     means[j] ← mean1 + (j - 1) * d;
7     // Arithmetic progression for the distance between the vectors of means ;
8     means[j] ;
9     // Vector of means
10
11 for i ← 1 to length(g) + 1 do
12   if i ← length(g) then
13     covar ← matrix(NA, nrow = n.var[1], ncol = n.var[1], byrow = FALSE);
14     // Create matrix variance-covariance using random variables with signal ;
15     covar[lower.tri(covar, diag = F)] ← corr[i] * sqrt(sds2[i] * sds2[i]);
16     covar[upper.tri(covar)] ← corr[i] * sqrt(sds2[i] * sds2[i]);
17     diag(covar) ← sqrt(sds2[i] * sds2[i]);
18     liscovar[[i]] ← covar ;
19     // List of matrices of variance-covariance by group ;
20     mu[[i]] ← rep(means[i], n.var[1]);
21     // Vector of means by group ;
22     x1[[i]] ← rmvnorm(g[i], mean = mu[[i]], sigma = liscovar[[i]]);
23     // Simulation dataset under a multivariate normal distribution by group ;
24   else
25     covar ← matrix(NA, nrow = n.var[2], ncol = n.var[2], byrow = FALSE);
26     // Create matrix variance-covariance using noise variables ;
27     covar[lower.tri(covar, diag = F)] ← corr[i] * sqrt(sds2[i] * sds2[i]);
28     covar[upper.tri(covar)] ← corr[i] * sqrt(sds2[i] * sds2[i]);
29     diag(covar) ← sqrt(sds2[i] * sds2[i]);
30     mu2 ← rep(means[i], n.var[2]);
31     x1[[i]] ← rmvnorm(sum(g), mean = mu2, sigma = covar);
32     // Simulation dataset with noise variables under a multivariate normal
33     // distribution by group ;
34
35 x2 ← list.remove(x1, c(length(g) + 1));
36 grp ← factor(rep(LETTERS[1 : length(g)], g));
37 output ← do.call(rbind, lapply(x2, matrix, ncol = n.var[1], byrow = F));
38 x3 ← x1[[length(g) + 1]];
39 x4 ← as.data.frame(cbind(output, x3, grp));
40 for i ← 1 to length(g) do
41   lisx[[i]];
42   // List dataset with structure multigroup, no group factor ;
43
44 return list()
```

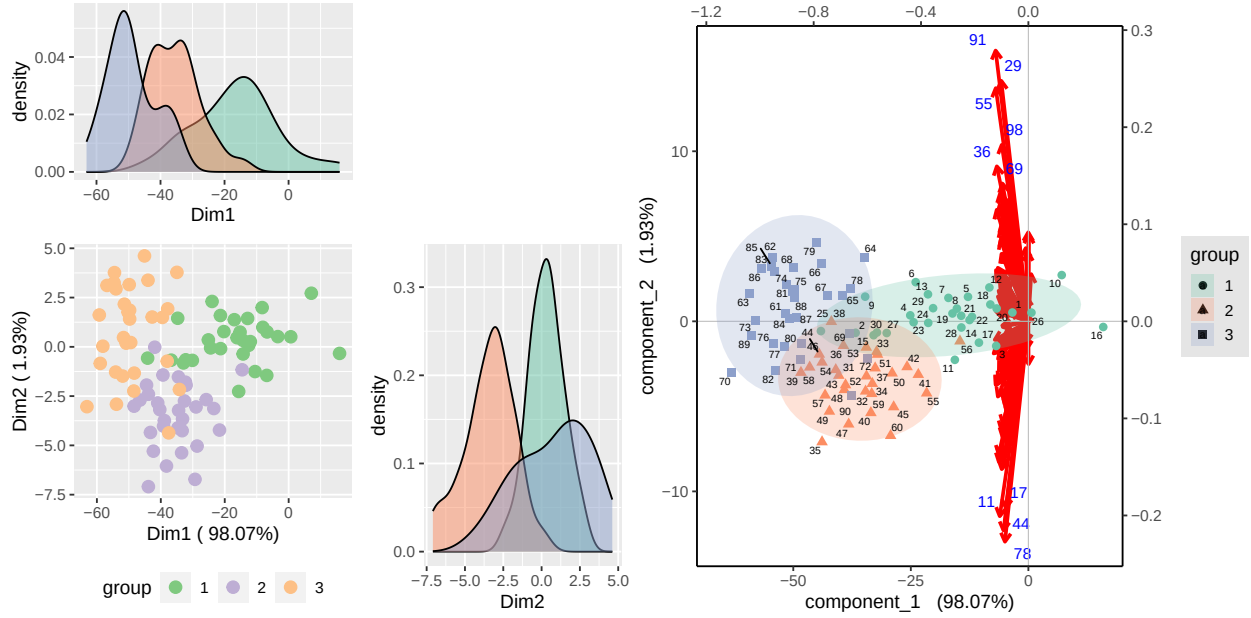


FIG. 3. MDR with 30 observations and correlation (0.5,0.2,0.1) for each group, respectively; in the upper and lower triangular. Different variances for each group (3,3,5). The variables with signal are 100 and the variable noise are 10. Graph display MDR.

6.2. Real Dataset

Real dataset is always important to study together with new proposals for statistical methodologies, as it allows us to observe and understand the nature of real dataset. Also validate the new statistical approaches to be able to apply them in real life and can serve the advancement of studies of high-dimensional data in the area of biomedicine and others.

However, we are going to focus on the study of omics data, since in the area of biomedicine they are real data of great importance in the last decade. The omics data has allowed us to understand the nature of diseases that before, humanity, could not fully understand.

This has allowed us to improve treatments and drugs for certain diseases that were once very complicated for humanity. For example, the types of cancer, viral diseases such as covid-19 and others, diabetes, VIH, pharmacology and other areas of health science. All these examples have been studied omics data.

Omic data are interesting to explore because they have great difficulty analyzing them with statistical and computational methodologies, since they are of high dimensionality, it has a lot information or missing data and/or groups and the complexity of studying several sources of variation together such as transcriptomics, metabolomics, proteomics, metagenomics, etc.

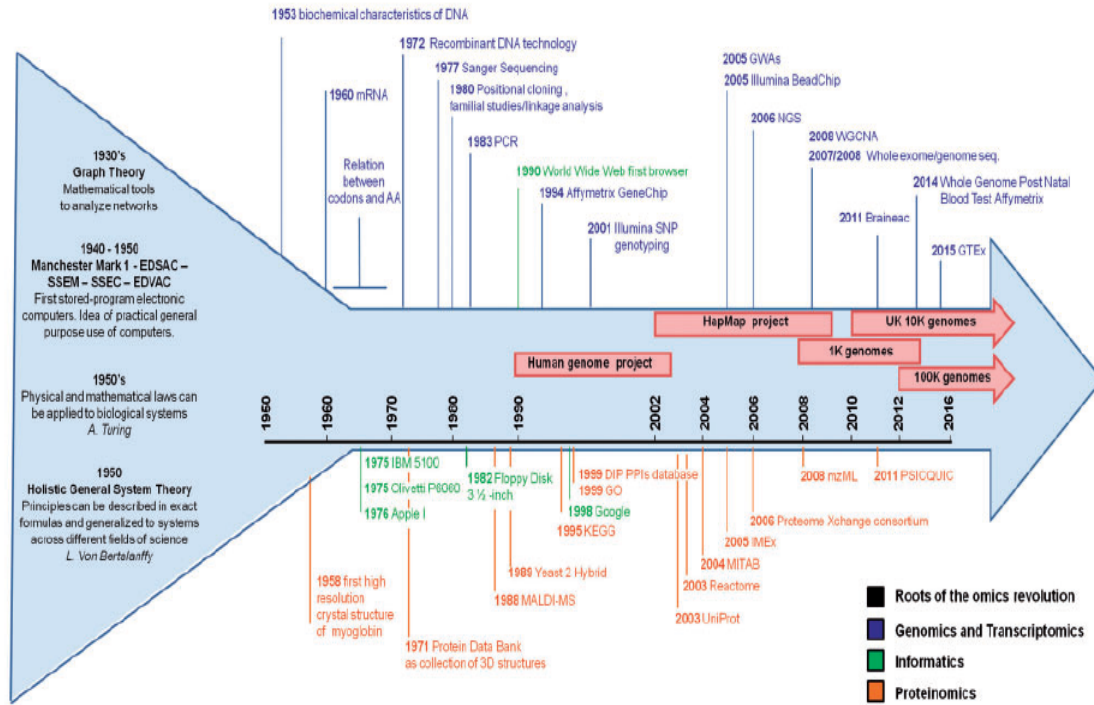


FIG. 4. Overview of the progressive advance in the methods to study genes, transcripts and proteins in the computer science [1].

For these reasons, it is necessary to advance in statistical methodologies to improve the analysis of these data in biomedicine, and increasingly to understand and explore them in a better computational way along with new statistical methods. In this way, genes that are biomarkers of diseases such as some type of cancer can be discovered even more.

6.2.1. Introduction Omic Dataset. In recent decades there has been progress and interest in the development of omic dataset research. In addition to the technological and computer advancement that has helped generate these omic dataset of high dimensionality in the area of biomedicine. It is important to analyze this data in a multidisciplinary way with the help of different experts worldwide. In the figure 4 shows the development, over time, of the many disciplines now involved in biomedical science accompanied by the fundamental advances in informatics and community resources. The broad roots of the omics revolution are represented by the wider start of the arrow before the year 1950, when the foundations for a paradigm shift in science (from single observations to systems dynamics) were laid [1].

On the other hand, the rapid development of microarray and RNA-seq technologies in the past few years has enabled researchers to analyse thousands of genes simultaneously and obtain biological information for various purposes, especially for cancer classification. However, gene

expression data obtained by throughput technologies could bring difficulties to methods due to the fact that usually the number of genes is very big, while the number of samples is small. This fact is known as the curse of dimensionality in data mining [39]. In this thesis R software was used to analyze these real data, mainly in the area of biomedicine. In bioinformatics and biostatistics there are packages to process biological databases such as: Bioconductor, genetics, gap, ape, pheatmap and mixOmics.

6.2.2. Gene expression data from Khan et al. (2001). This dataset is childhood genetic expressions of cancer diseases that classified into four types of tumors. The small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS)[29]. The data have 2308 genes for 83 samples with unbalanced groups. These classifications according to their clinical diagnosis used artificial neural networks (ANNs).

It is necessary to study these types of childhood cancers, since as the authors Khan and others say, they are types of tumors that are difficult to detect clinically and also complicated to analyze statistically and to obtain an optimal classification in each group.

These microarray expression genes are in the `plsgenomics` package in R and `made4` (requires `ade4` package in R) in Bioconductor.

6.2.3. Genes associated with Myeloid leukemia in patients with cytogenetics normal or trisomy on chromosome 8. This dataset is an investigation of leukemia disease, specifically acute myeloid leukemia. Acute myeloid leukemia (AML) is a heterogeneous group of diseases. Normal cytogenetics (CN) constitutes the single largest group, while trisomy 8 (+8) as a sole abnormality is the most frequent trisomy.

The oligonucleotide-based DNA microarrays to study global gene expression in AML+8 patients with +8 as the sole chromosomal abnormality and AML-CN patients. CD34+ cells purified from normal bone marrow (BM) were also analyzed as a representative heterogeneous population of stem and progenitor cells. Expression patterns of AML patients were clearly distinct from those of CD34+ cells of normal individuals[40].

This global gene expression has three groups where two groups have one type of leukemia and the other group are normal individuals. The age of the patients is a median of 70 years.

The study was carried out with 48 samples that were hybridized with arrays of affymetrix of the hu6800 type and of which, for this example, a balanced subset of 15 samples, 5 of each type.

Chapter 7

Results

This chapter presents the methods proposed in this thesis, carrying out the simulation scenarios and applying them to real dataset in the area of molecular biomedicine.

Some results are presented in the case of simulations applying our `MultiGroup0` package. Up to 5 groups are carried out in the simulation with observations, or small samples and with many variables. Thus, applying our multigroup PCA and MDR methods. Then is observed if they can generate groups when the standard PCA has many problems in grouping under certain parameters in the data simulation.

On the other hand, the real dataset is used to apply our multigroup PCA and MDR methods as new clustering approaches that can help to find cluster in omic data with high dimensionality where the samples are small and have a large number of variables.

7.1. Simulation Scenarios

This section, several data simulation scenarios with different numbers of groups are presented and visualized with the proposed methods that are multigroup PCA and MDR. The PCA standard is also plotted.

```
#install.packages('MultiGroup0')  
library('MultiGroup0')
```

7.1.1. Simulation scenarios for 3 groups.

- (1) This scenario shows that the standard PCA (see figure 1) has complications in grouping, since it is a data set with few samples and has many variables. In addition, these variables have a large amount without a biological signal, that is, noise variables and few variables with a biological signal. Within the latter, the standard PCA does not have a good behavior when separating the groups.

Contrary to what happens with the standard PCA, our Multigroup PCA method (see figure 2) has a better behavior in separating the groups. Although it does not succeed to separate the three groups proposed in this simulation, but does manage to separate two of them. This can also be observed in the biplot where it is possible to improve the separation of the groups in this case of data simulation.

On the other hand, we have another new method that we propose in this thesis, different from the PCA approach, which is the MDR method. This technique clearly observes the three groups in this data simulation scenario. It has no problem achieving separation with its first two dimensions (see figure 3).

```
require(MultiGroup0)
set.seed(12569)
sim.list<-fun.sim(g=c(50,50,50),mean1=1,d=0,sds2=c(1,1,1,1),
                 corr=c(0.5,0.5,0.5,0),
                 n.var=c(500,1000))
```

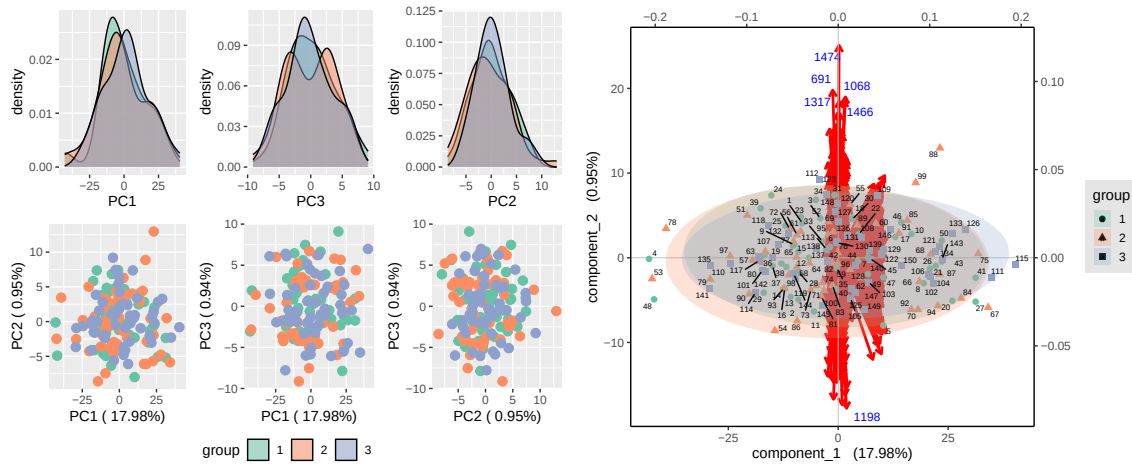
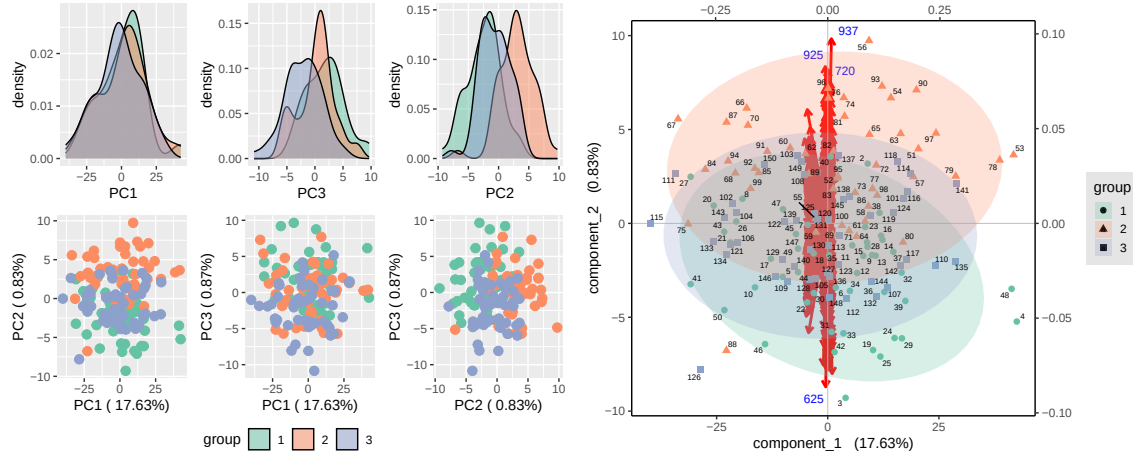
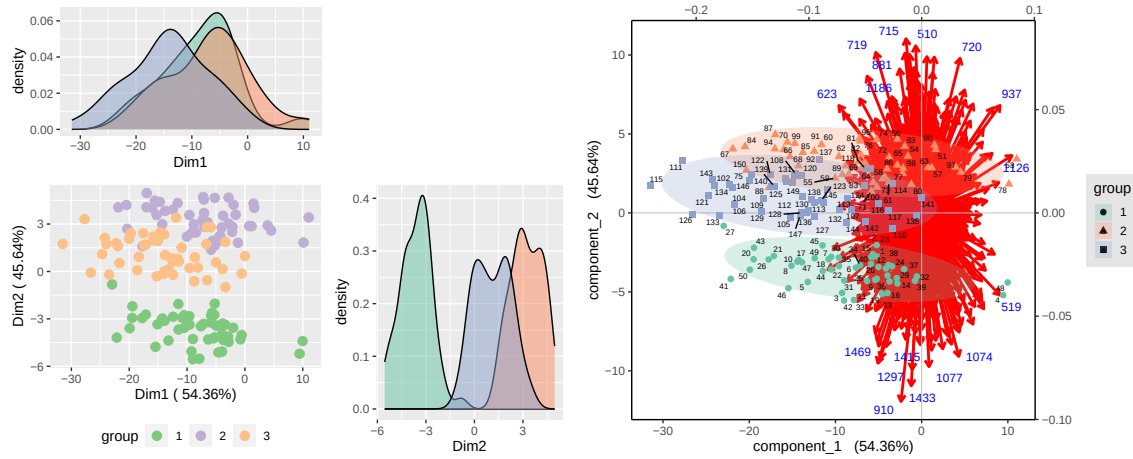


FIG. 1. Graphs under the function `pca()` standard by `Multigroup0` package.

FIG. 2. Graphs under the function `mgPCA()` by Multigroup0 package.FIG. 3. Graphs under the function `MDR()` by Multigroup0 package.

- (2) In this second case of simulation, it is observed that in the standard PCA (see figure 4) it is not possible to separate the three simulated groups with its first two main components. This case has many noise variables and few biological signal variables, but this time, the variances for each group were increased and low correlations for two groups and the other moderate.

The multigroup PCA tries to separate the groups with the first two principal components; managing to observe only two groups of the three simulated. This is best seen in the biplot where the ellipses manage to be further apart than in the standard PCA case (see figure 5).

For our MDR method (see figure 6), it separated the simulated groups with the first two dimensions. We can also see in the graph where it shows the three separated ellipses with the first dimensions.

```
require(MultiGroup0)
set.seed(12569)
sim.list<-fun.sim(g=c(100,100,100),mean1=1,d=0,sds2=c(2.5,3.0,3.5,1),
  corr=c(0.5,0.3,0.2,0),
  n.var=c(500,1000))
```

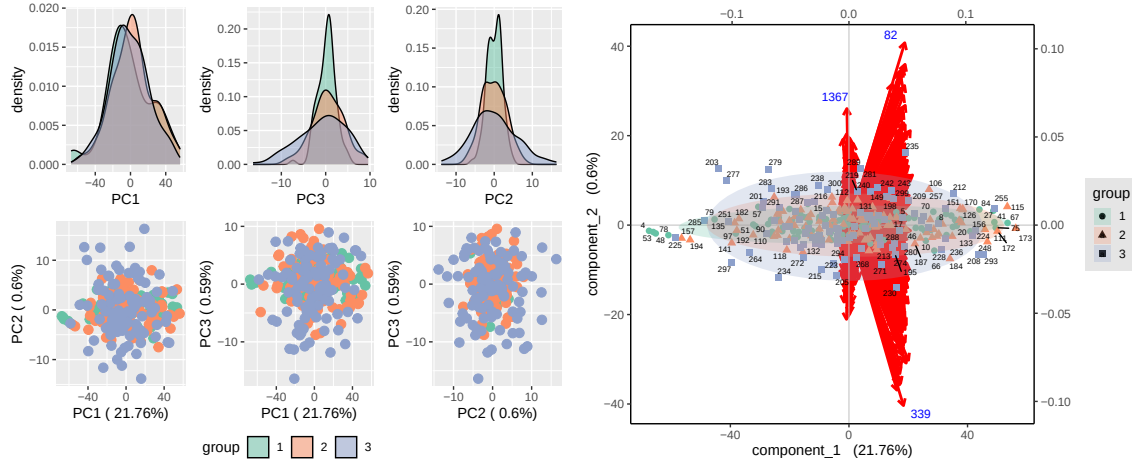


FIG. 4. Graphs under the function `pca()` standard by `MultiGroup0` package.

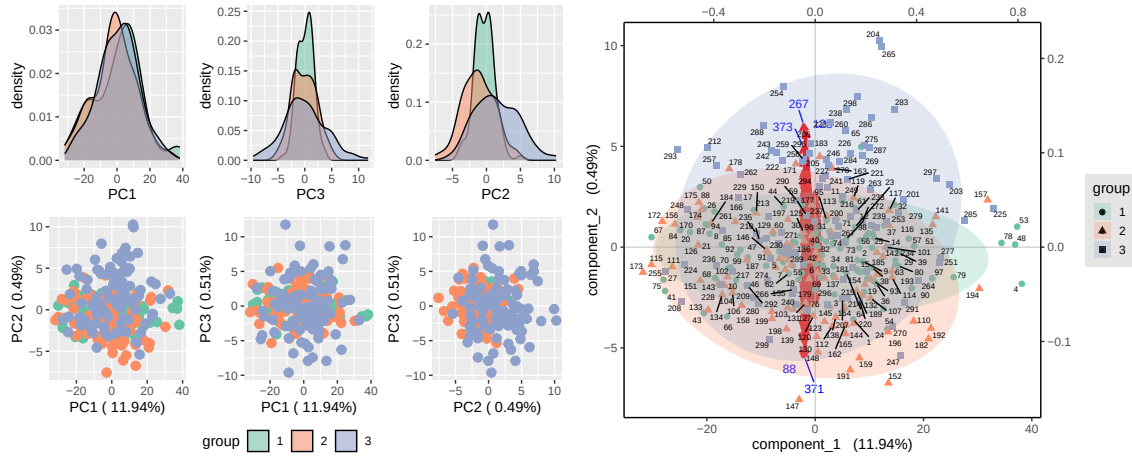


FIG. 5. Graphs under the function `mgPCA()` by `MultiGroup0` package.

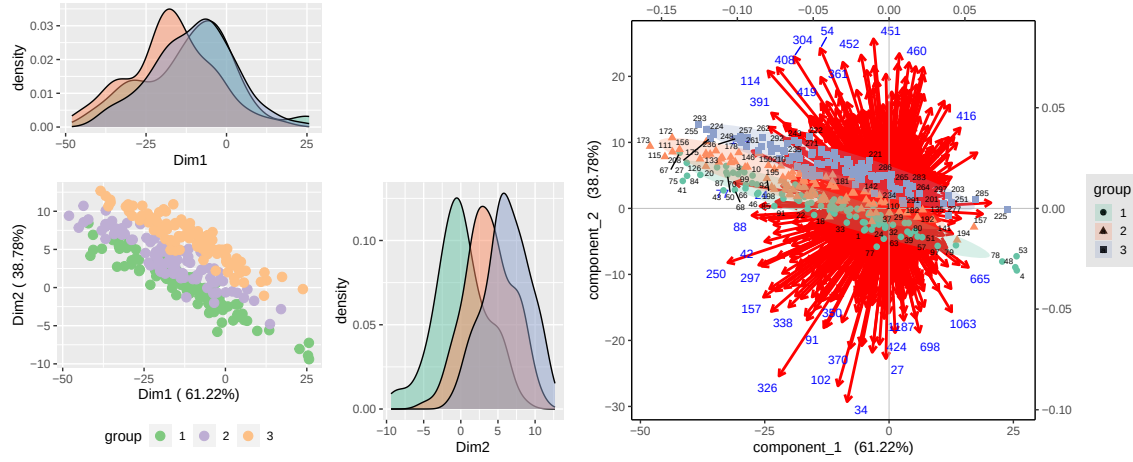


FIG. 6. Graphs under the function `MDR()` by `Multigroup0` package.

- (3) The third simulation presents us with many variables with biological signals and few noise variables. It also has high correlations in the three simulated groups and high variances in all of the groups. In this scenario, the classical PCA method fails to separate the groups with their first two principal components (see figure 7).

The new multigroup PCA approach proposed in this thesis shows that for this simulation case it manages to separate the three simulated groups with their first two principal components. Also, we see it in the biplot where the ellipses are separated quite well with the first two dimensions (see figure 8).

The MDR method shows two separate clusters of the three simulated (see figure 9). For this case, the MDR fails to separate the three groups, but can classify two groups with its two dimensions. We also observe this in the biplot where the ellipses are very elongated, possibly due to the high correlations and variances in the groups.

```
set.seed(12569)
sim.list<-fun.sim(g=c(30,30,30),mean1=2,d=0,sds2=c(4.5,3.2,1.6,1),
  corr=c(0.8,0.9,0.65,0),
  n.var=c(3000,500))
```

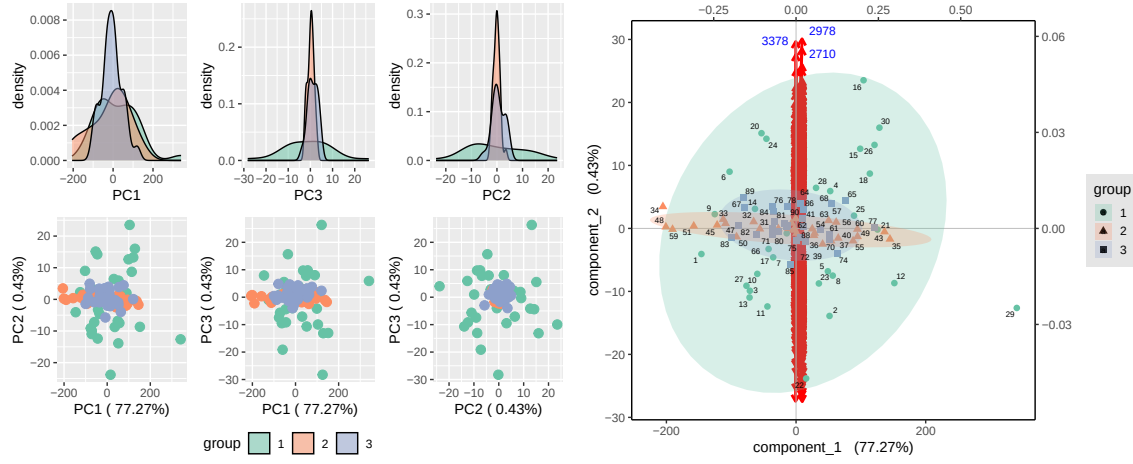


FIG. 7. Graphs under the function `pca()` standard by `Multigroup0` package.

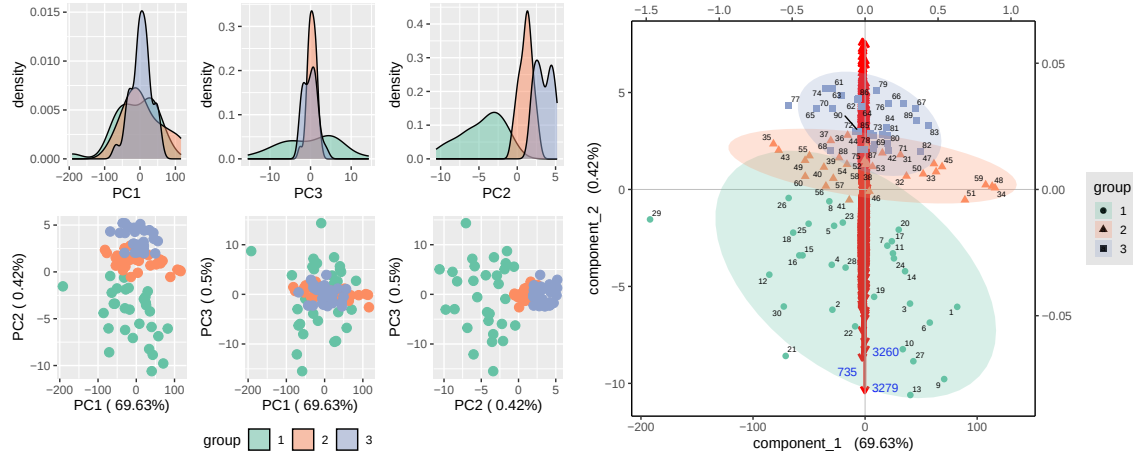


FIG. 8. Graphs under the function `mgPCA()` by `Multigroup0` package.

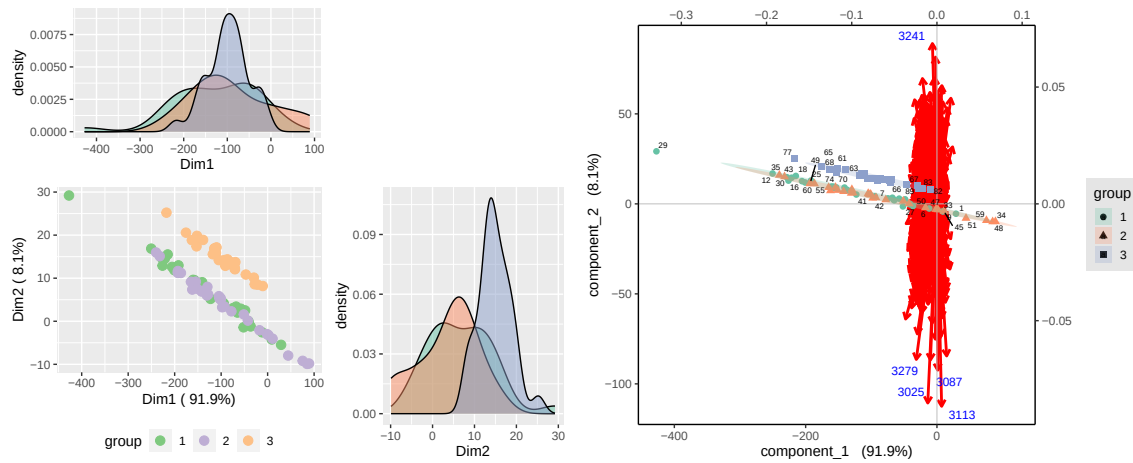


FIG. 9. Graphs under the function `MDR()` by `Multigroup0` package.

- (4) The fourth simulation scenario presents us with three groups with heterogeneous variances, low correlations, and many variables with biological signals and very few noise variables. Under this scenario, the standard PCA manages to separate the three groups with their first two principal components, but the three groups are observed to be quite coupled (see figure 10).

For our multigroup PCA approach, it is observed that it manages to separate the three groups, although they are also coupled. Although, in the biplot it is observed that the ellipses look more separated and less stretched than in the standard PCA (see figure 11).

The MDR method manages to separate the three simulated groups very well and they are not coupled. We also see it in the biplot where the ellipses are better separated with the two dimensions (see figure 12).

```
set.seed(12569)
sim.list<-fun.sim(g=c(50,50,50),mean1=4,d=2,sds2=c(2.1,2.3,2.5,1),
  corr=c(0.5,0.3,0.2,0),
  n.var=c(2000,50))
```

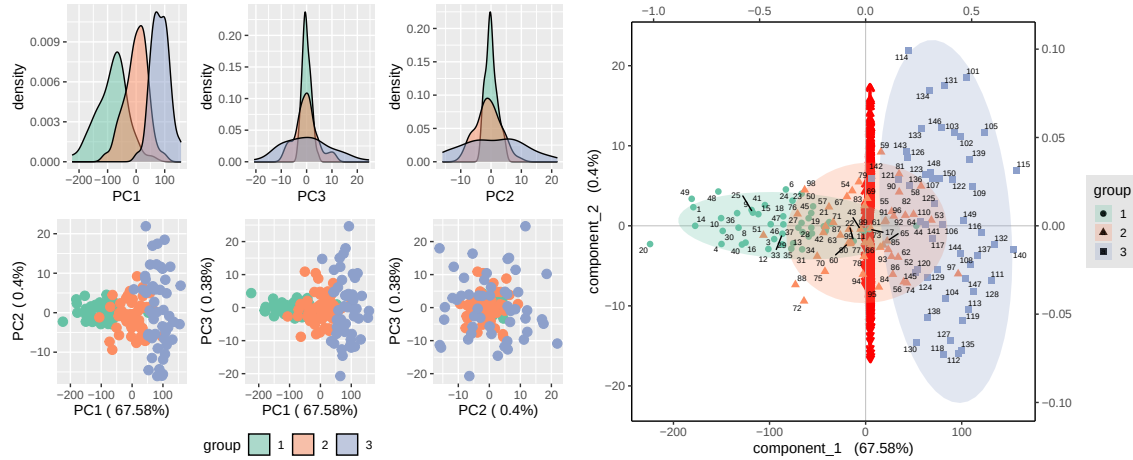


FIG. 10. Graphs under the function `pca()` standard by `Multigroup0` package.

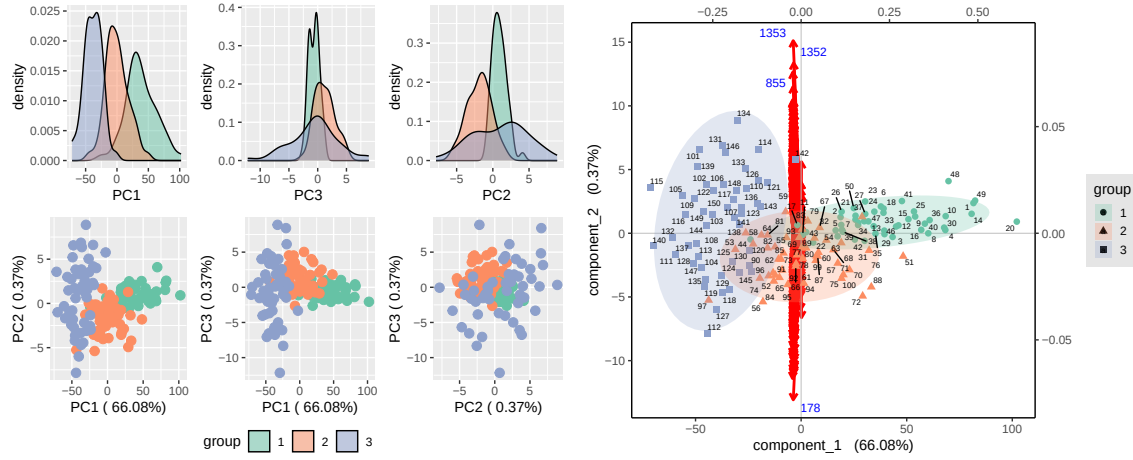


FIG. 11. Graphs under the function `mgPCA()` by `Multigroup0` package.

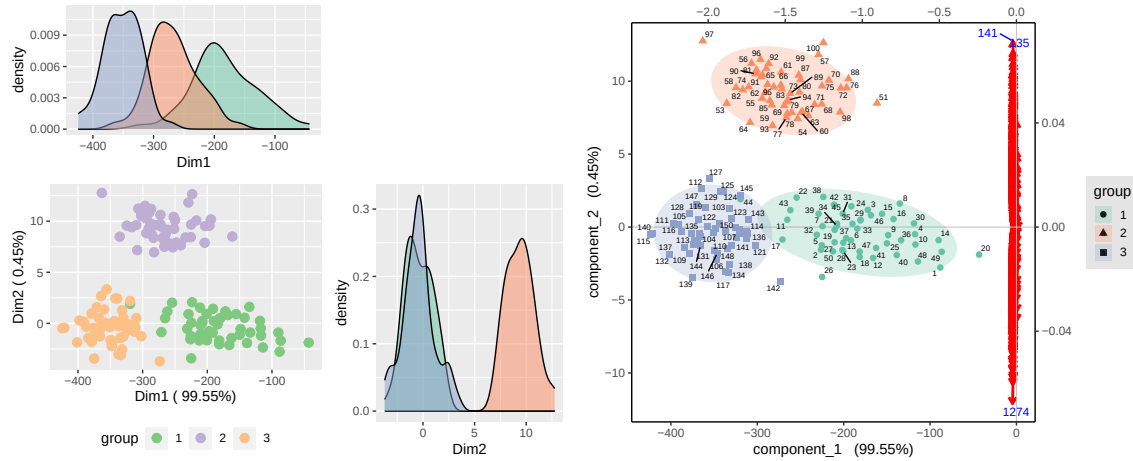


FIG. 12. Graphs under the function `MDR()` by `Multigroup0` package.

- (5) The fifth simulation with three groups increased the noise variables, although there are still more variables with biological signals. Furthermore, the three groups have the same but moderate correlation and heterogeneous variance.

The classic PCA fails to separate the groups or classify the samples into clusters (see figure 13). The biplot also shows no improvement from separating the ellipses with the first two principal components.

The multigroup PCA manages to separate two groups of the three simulated with the first two principal components, which also indicates that it tries to classify the samples in the clusters. The biplot shows (see figure 14) that the ellipses are trying to separate and form the three groups even though one of them is very close to the other group.

The MDR approach shows us that it manages to separate the three groups (see figure 15), although there is one that is coupled to the other group, but still, it achieves a

certain distance between the three groups. It is also observed in the graph where the ellipses are separated the groups quite well.

```
set.seed(12569)
sim.list<-fun.sim(g=c(50,50,50),mean1=4,d=0,sds2=c(2.1,2.3,2.5,1),
  corr=c(0.5,0.5,0.5,0),
  n.var=c(2000,1000))
```

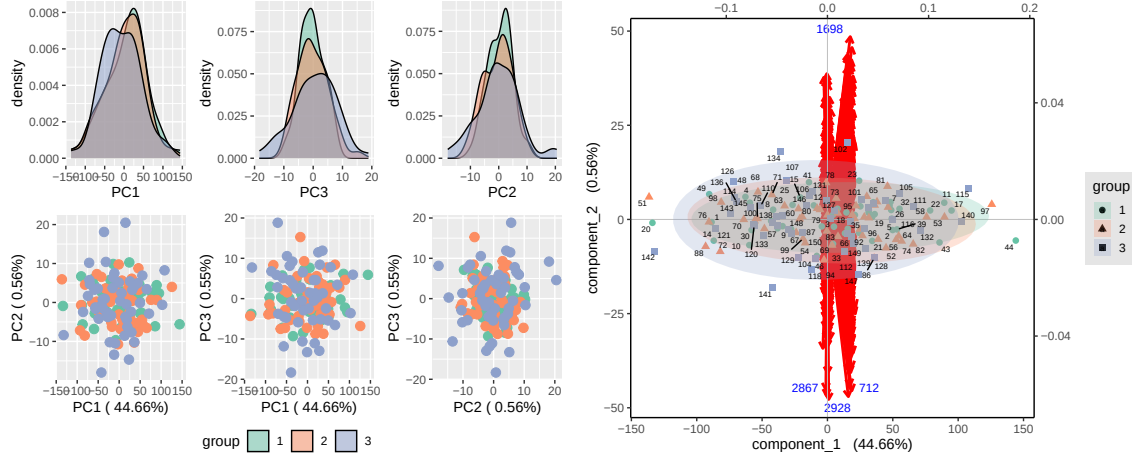


FIG. 13. Graphs under the function `pca()` standard by `Multigroup0` package.

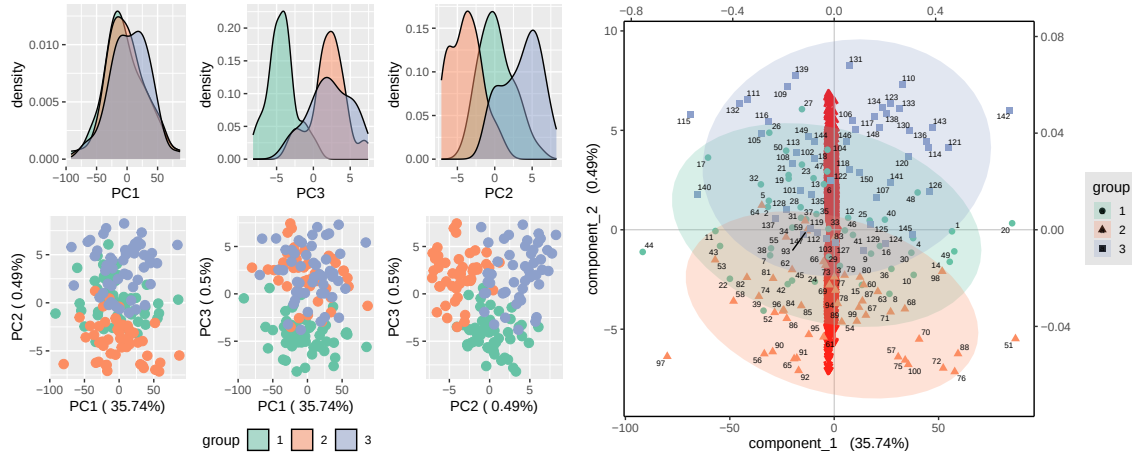


FIG. 14. Graphs under the function `mgPCA()` by `Multigroup0` package.

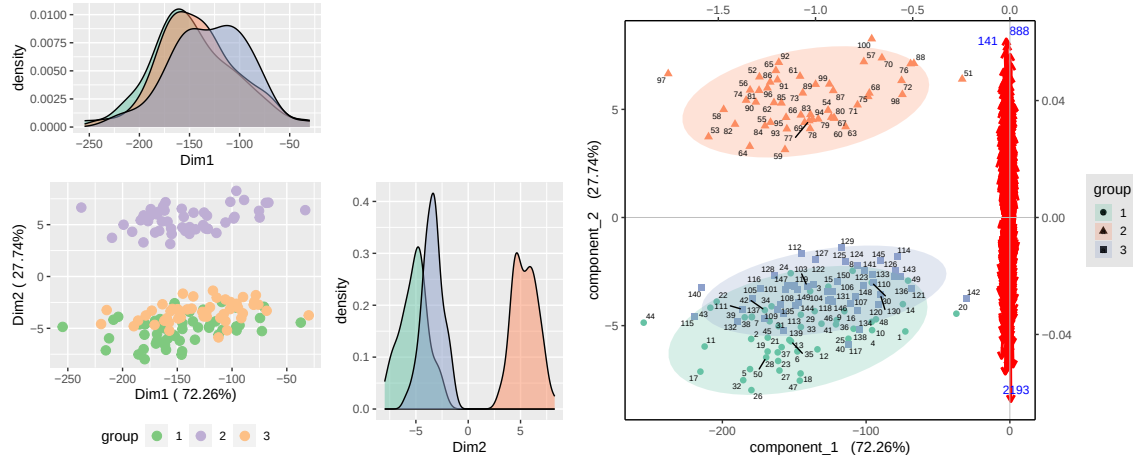


FIG. 15. Graphs under the function `MDR()` by `Multigroup0` package.

7.1.2. Simulation scenarios for 4 groups.

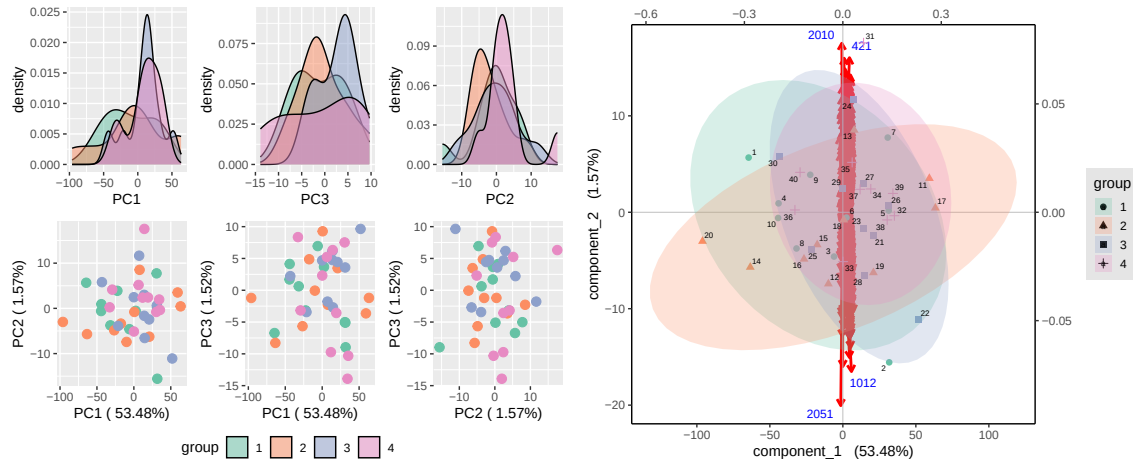
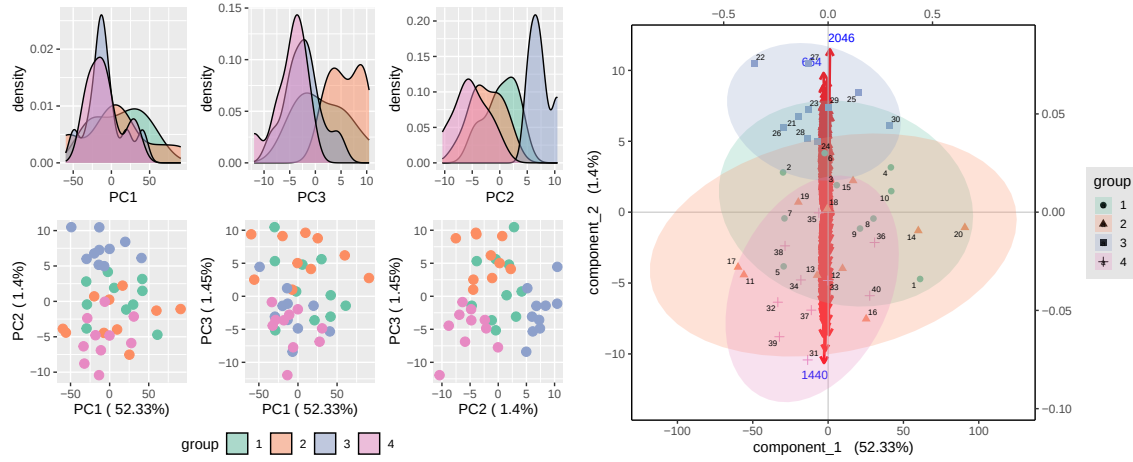
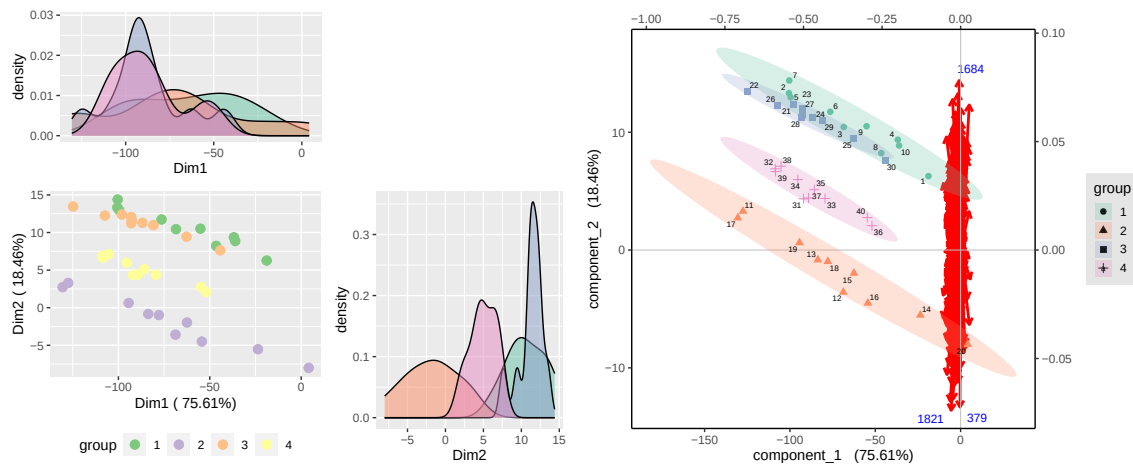
- (1) The sixth simulation scenario shows us a dataset with four groups, homogeneous variances, moderate correlations for each group, many variables with biological signals and few noise variables.

The standard PCA shows us that it cannot separate the groups with the first two principal components. This also implies that you cannot classify the samples into the groups, respectively. This observation is also seen in the biplot where it is not possible to separate the ellipses (see figure 16).

Our multigroup PCA manages to separate three groups out of the four simulated with the first two principal components. We also see this in the biplot where the ellipses try to separate, although they are coupled between the groups, but the four simulated groups can be observed (see figure 17).

Figure 18 shows that our MDR method manages to separate the four simulated groups with two dimensions. This is also observed in the graph, since the ellipses are separated, there is only one group that is somewhat coupled to another group, but the separation of the four groups is much better in this simulation case.

```
set.seed(12569)
sim.list<-fun.sim(g=c(10,10,10,10),mean1=2,d=0,sds2=c(1,1,1,1,1),
                 corr=c(0.5,0.5,0.5,0.5,0),n.var=c(2000,100))
```

FIG. 16. Graphs under the function `pca()` standard by Multigroup0 package.FIG. 17. Graphs under the function `mgPCA()` by Multigroup0 package.FIG. 18. Graphs under the function `MDR()` by Multigroup0 package.

- (2) The seventh simulation shows us four groups with heterogeneous variances, some high, moderate correlations in the groups and many biological signal variables and few noise variables.

The standard PCA fails (see figure 19) to separate the groups with the first two principal components. The biplot also shows that the ellipses fail to separate, or identify each other.

Figure 20 shows us that the multigroup PCA manages to separate three of the four groups with the first two principal components. The biplot also shows us that the ellipses can be separated, although there is a group that couples another, but at least this method manages to identify the samples in each group and manages to separate the groups with two main components.

The MDR method shows us that it manages to separate the four simulated groups. It is observed that groups 3 and 4 are more coupled, but it is possible to separate with the first two dimensions (see figure 21).

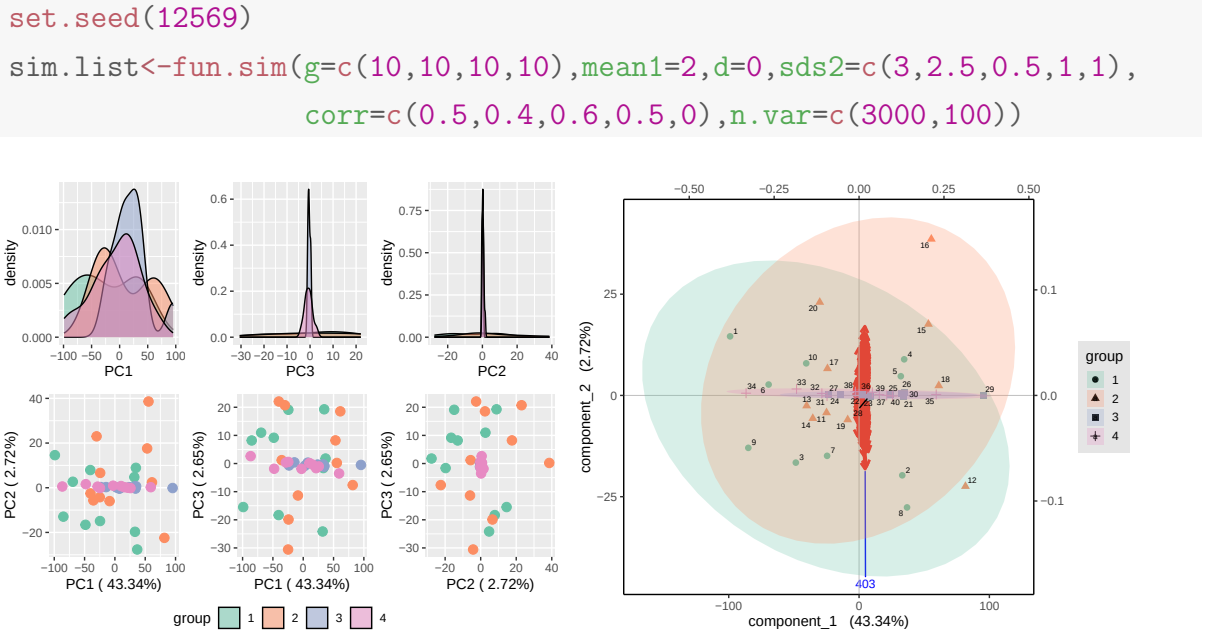
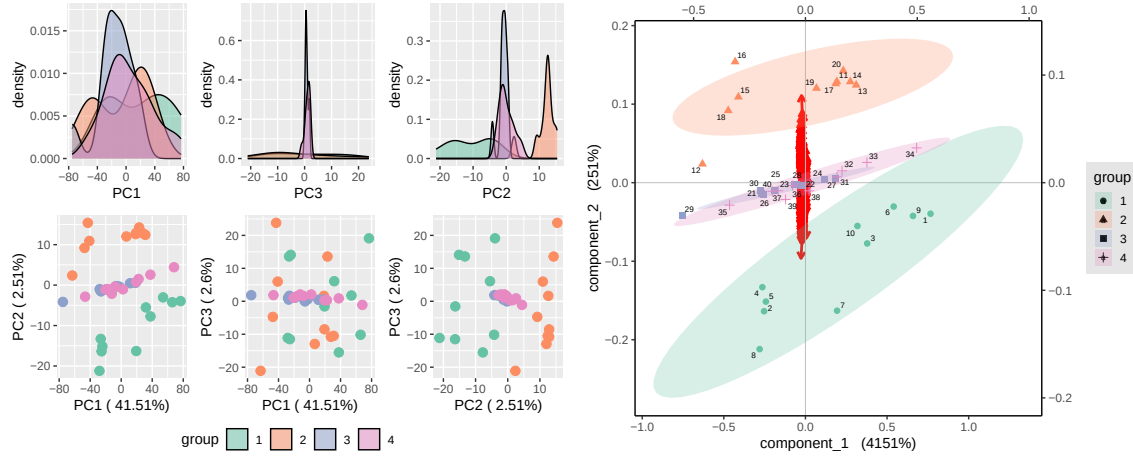
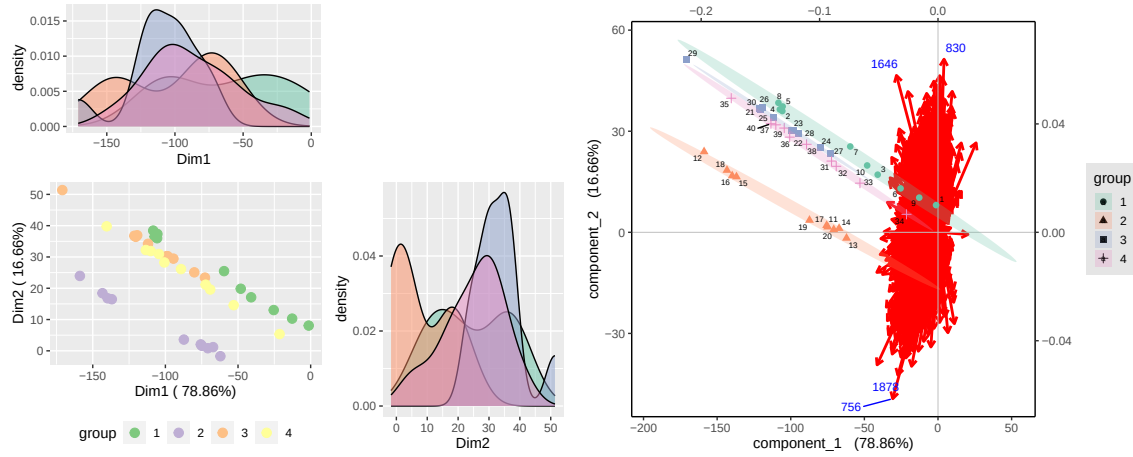


FIG. 19. Graphs under the function `pca()` standard by `Multigroup0` package.

FIG. 20. Graphs under the function `mgPCA()` by `Multigroup0` package.FIG. 21. Graphs under the function `MDR()` by `Multigroup0` package.

- (3) The eighth simulation presents us four groups with heterogeneous variances, some high, moderate correlations in the groups and few biological signal variables and many noise variables.

Figure 22 the standard `pca` fails to separate and identify the groups with the first two principal components.

The multigroup PCA method manages to separate three of the four simulated groups, since group 3 and 4 are highly coupled. However, our multigroup PCA manages to identify the samples for each group and separate much better than the classical PCA under this simulation scenario (see figure 23).

On the other hand, our MDR approach shows us that it manages to separate the 4 groups with two dimensions, although it is also observed that group 3 and 4 are closer but it is well identified in this method (see figure 24).

```
set.seed(12569)

sim.list<-fun.sim(g=c(10,10,10,10),mean1=2,d=0,sds2=c(3,2.5,0.5,1,1),
                 corr=c(0.5,0.4,0.6,0.5,0),n.var=c(100,3000))
```

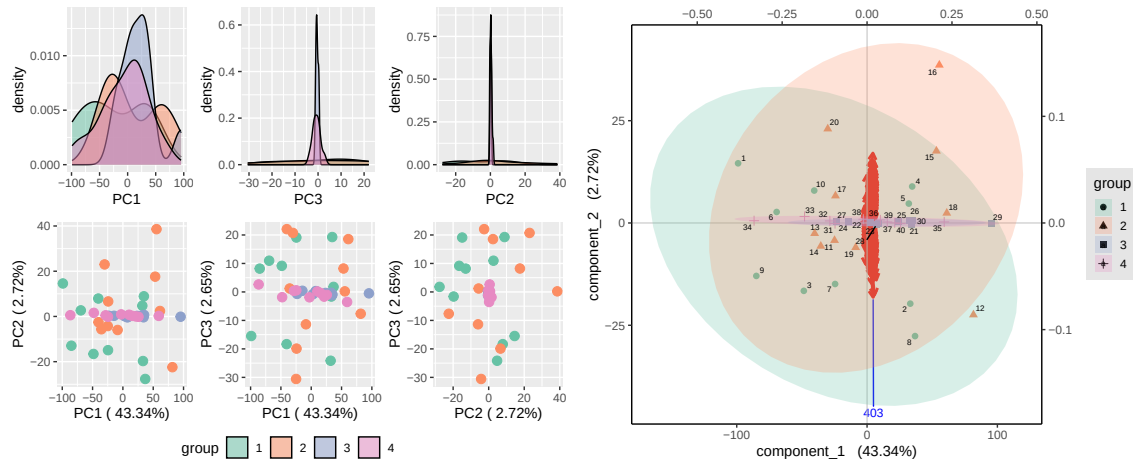


FIG. 22. Graphs under the function `pca()` standard by `Multigroup0` package.

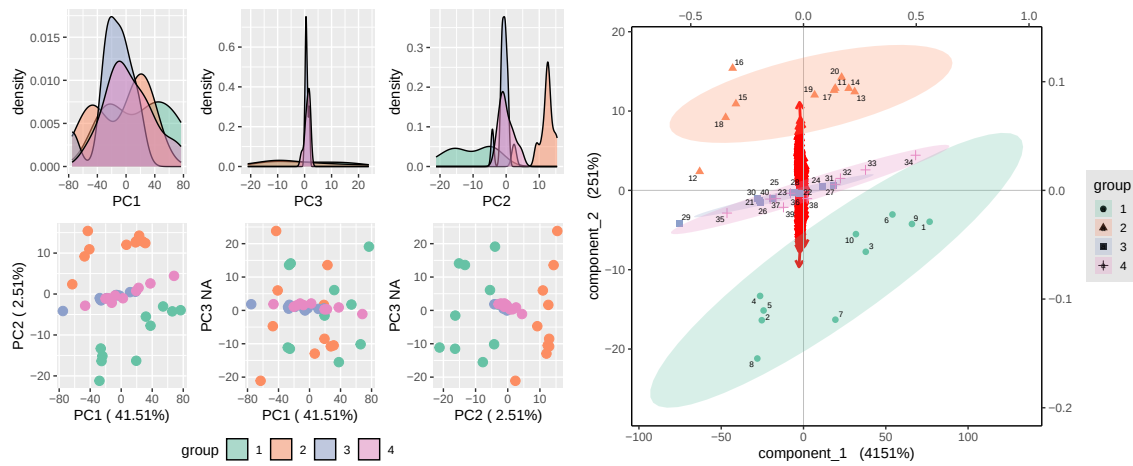


FIG. 23. Graphs under the function `mgPCA()` by `Multigroup0` package.

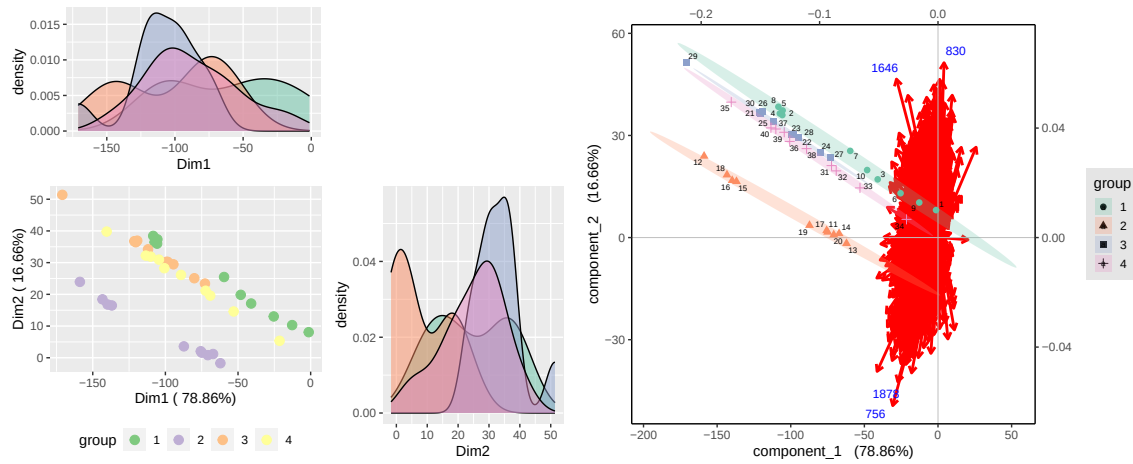


FIG. 24. Graphs under the function `MDR()` by `Multigroup0` package.

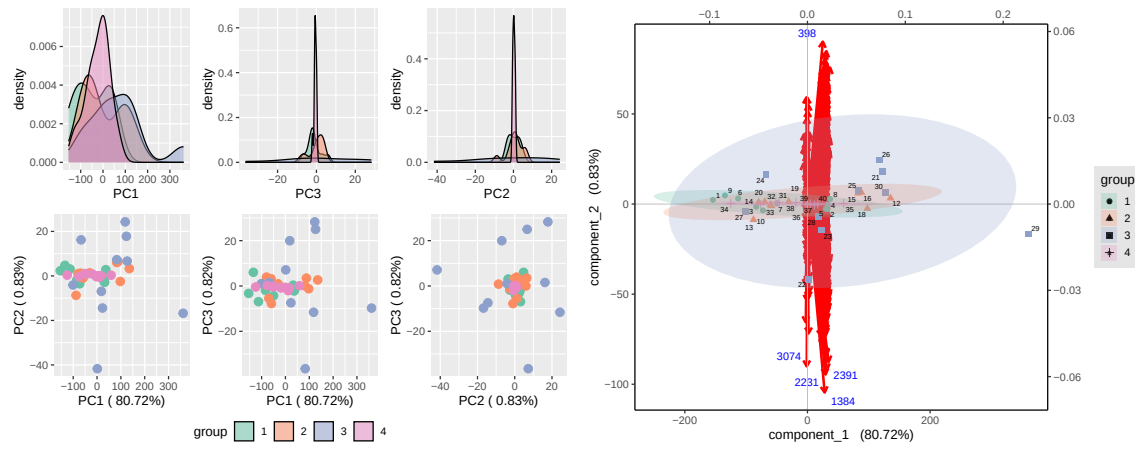
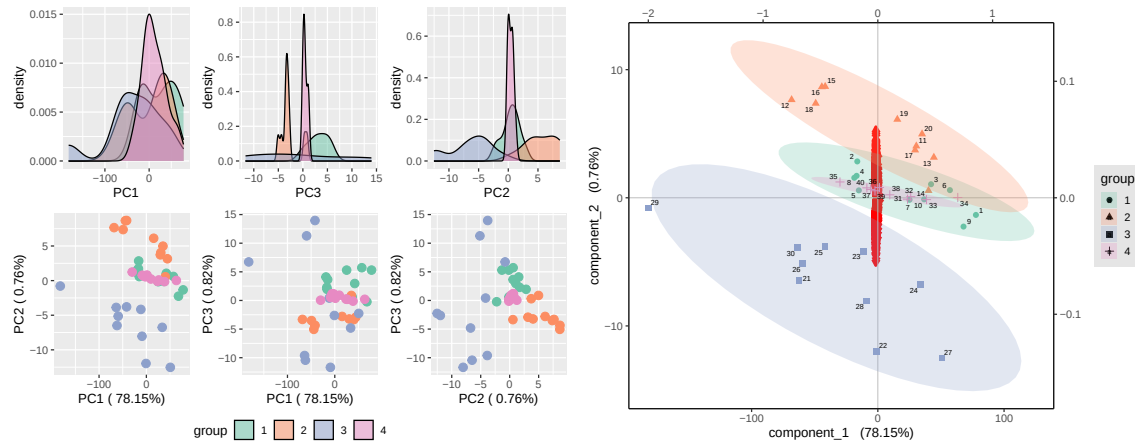
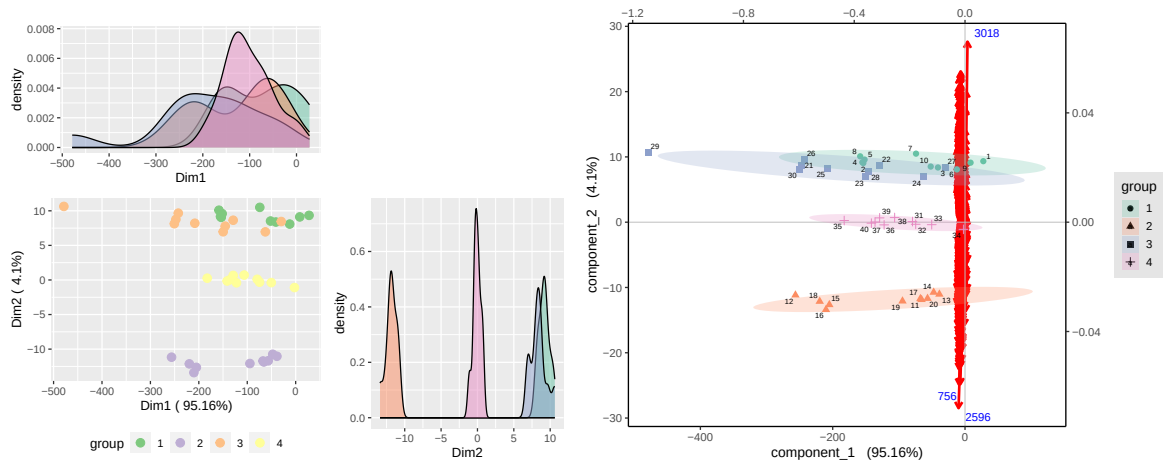
- (4) The ninth simulation shows us four groups with high heterogeneous variances, high correlations and many variables with biological signals and few noise variables.

Figure 25 show that the standard PCA fails to separate the groups with the first two principal components.

The multigroup PCA is observed to be able to separate the groups with the first two principal components, although group 1 and 4 are more coupled with each other. However, the separation of the clusters is better appreciated with our multigroup PCA approach (see figure 26). The biplot shows the separation of the ellipses but those of group 1 and 4 are coupled.

The figure 27 shows that the MDR method manages to separate the groups with two dimensions. The graph shows the ellipses separated from each other, although it is observed that groups 1 and 4 are somewhat coupled.

```
set.seed(12569)
sim.list<-fun.sim(g=c(10,10,10,10),mean1=2,d=0,sds2=c(4,4.5,5.5,1,1),
                 corr=c(0.8,0.8,0.8,0.8,0),n.var=c(3000,100))
```

FIG. 25. Graphs under the function `pca()` standard by Multigroup0 package.FIG. 26. Graphs under the function `mgPCA()` by Multigroup0 package.FIG. 27. Graphs under the function `MDR()` by Multigroup0 package.

7.1.3. Simulation scenarios for 5 groups.

- (1) The tenth simulation shows us five groups with high heterogeneous variances, high correlations and many variables with biological signals, and few noise variables, 3000 and 100, respectively.

The standard PCA fails to identify groups with their first two principal components. However, our multigroup pca manages to identify the groups and tries to separate them with the first two principal components. This is seen in the figure 29.

The MDR is a technique that manages to separate the groups quite well with their first two dimensions (see figure 30).

```
set.seed(12569)
sim.list<-fun.sim(g=c(10,10,10,10,10),mean1=2,d=0,sds2=c(4,4.5,5.5,5.2,1,1),
  corr=c(0.8,0.8,0.8,0.8,0.8,0),n.var=c(3000,100))
```

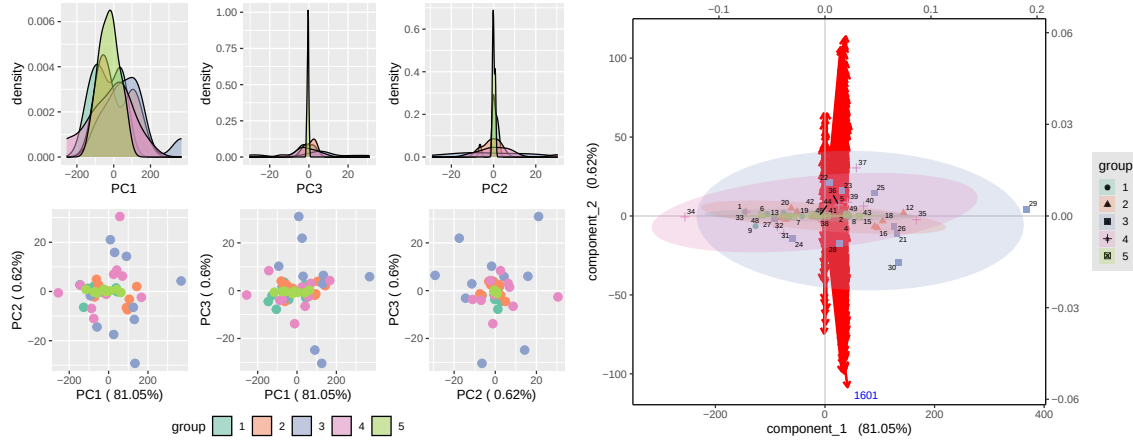


FIG. 28. Graphs under the function `pca()` standard by `Multigroup0` package.

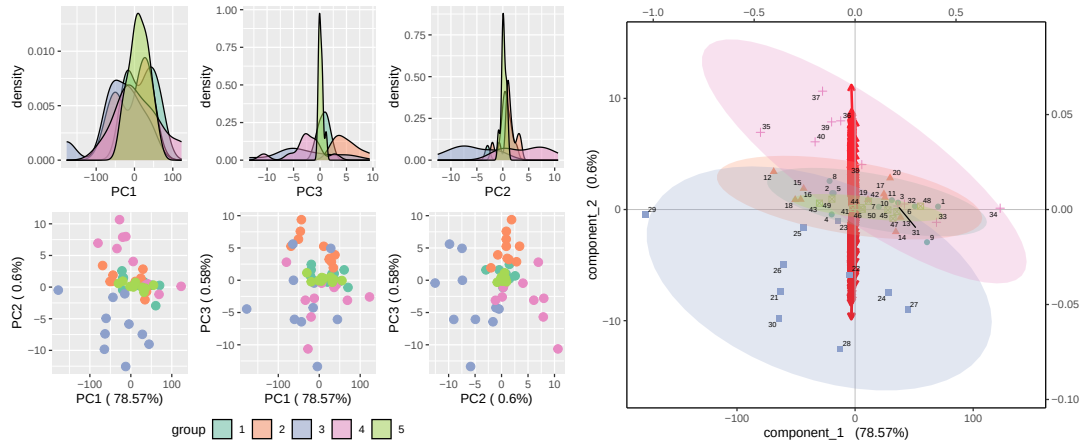


FIG. 29. Graphs under the function `mgPCA()` by `Multigroup0` package.

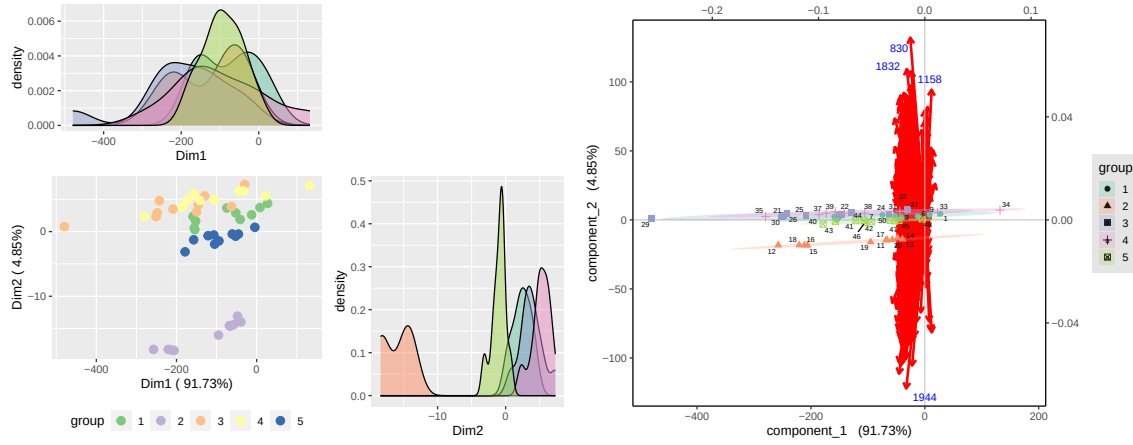


FIG. 30. Graphs under the function MDR() by Multigroup0 package.

- (2) The eleventh simulation shows us five groups with high heterogeneous variances, low and other moderate correlations and many variables with biological signals, and few noise variables, 3000 and 100, respectively.

The multigroup PCA proposed in this thesis is observed to try to separate the groups, although they are still coupled, but the standard PCA fails to identify the groups in the simulated data set with their first two principal components (see figure 31 and 32).

On the other hand, the MDR method is observed in the figure 33 that manages to separate the five groups with the first two dimensions.

```
set.seed(12569)
sim.list<-fun.sim(g=c(50,50,50,50,50),mean1=2,d=0,sds2=c(4,4.5,5.5,5.2,1,1),
  corr=c(0.1,0.2,0.3,0.4,0.4,0),n.var=c(3000,100))
```

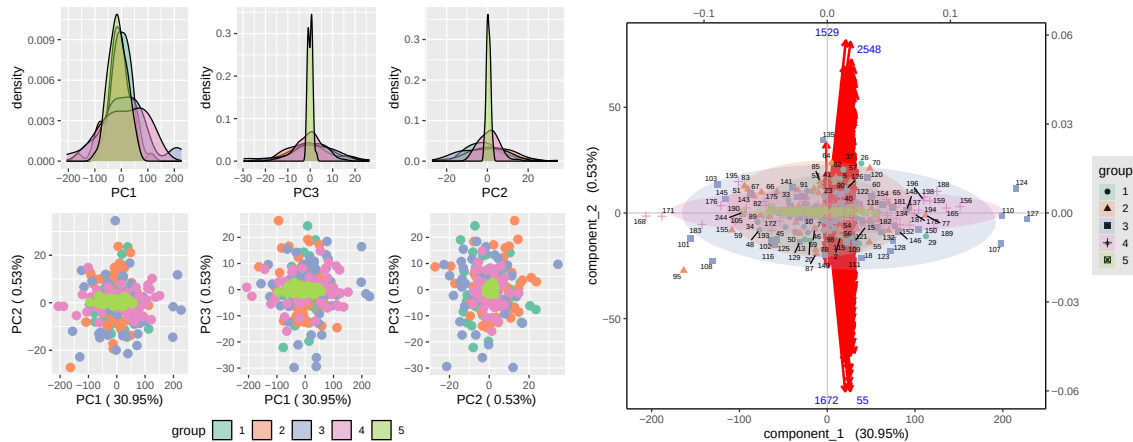
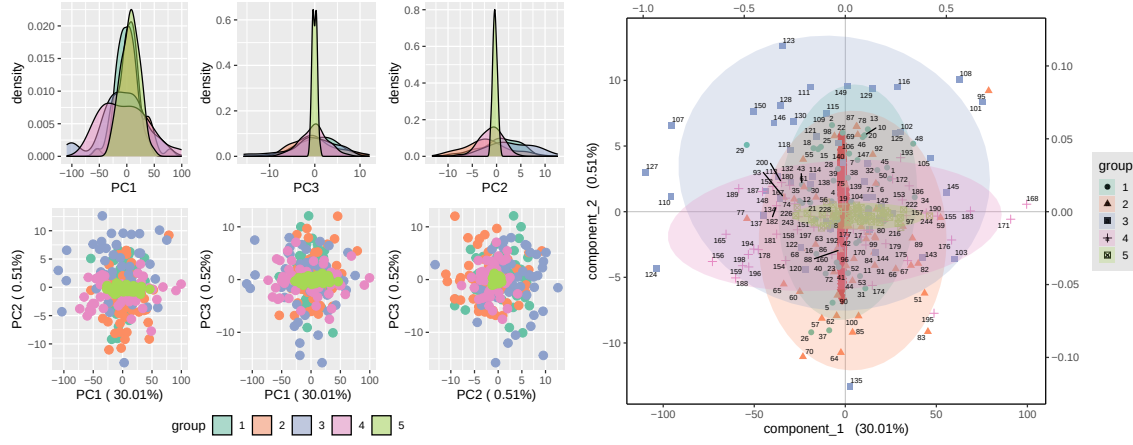
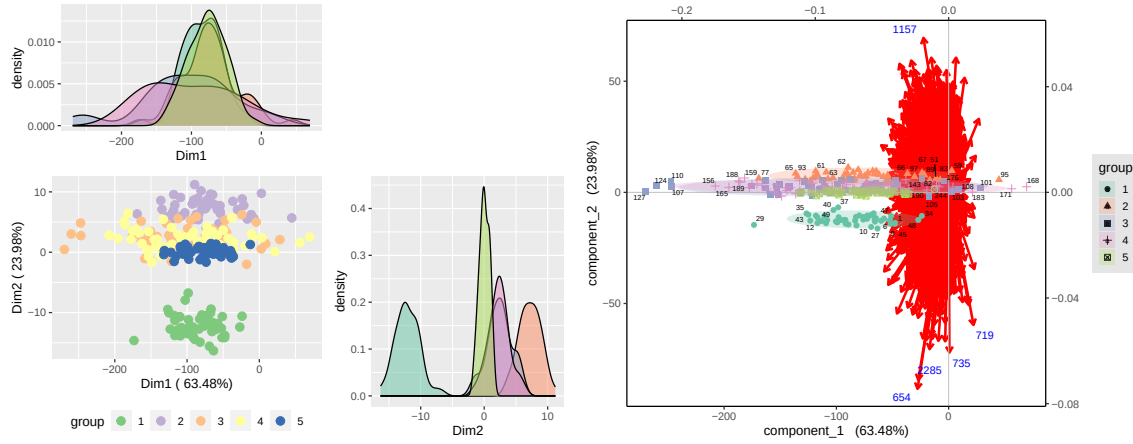


FIG. 31. Graphs under the function pca() standard by Multigroup0 package.

FIG. 32. Graphs under the function `mgPCA()` by Multigroup0 package.FIG. 33. Graphs under the function `MDR()` by Multigroup0 package.

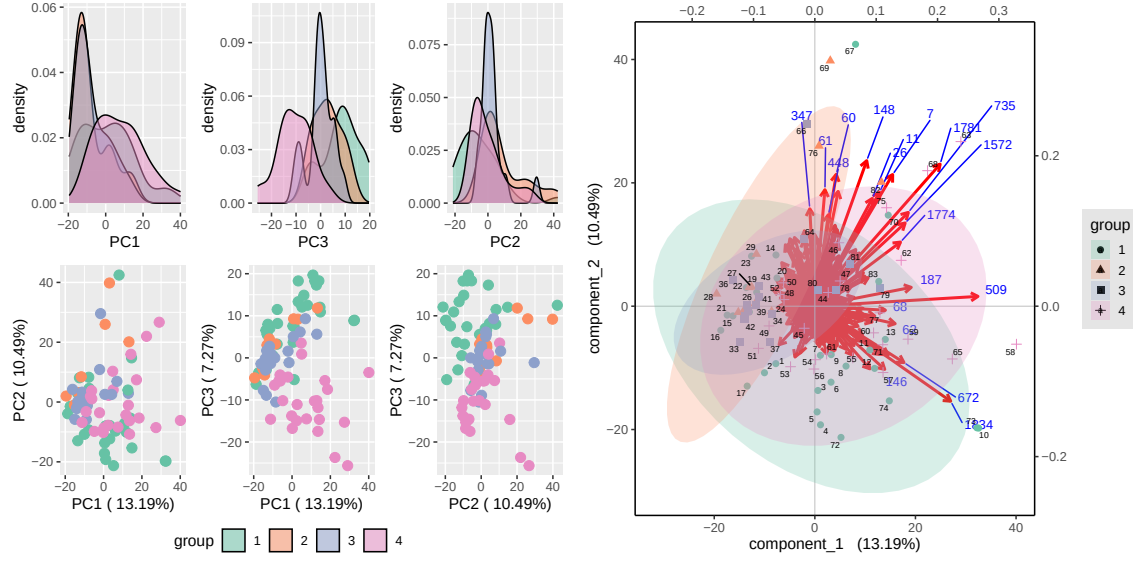
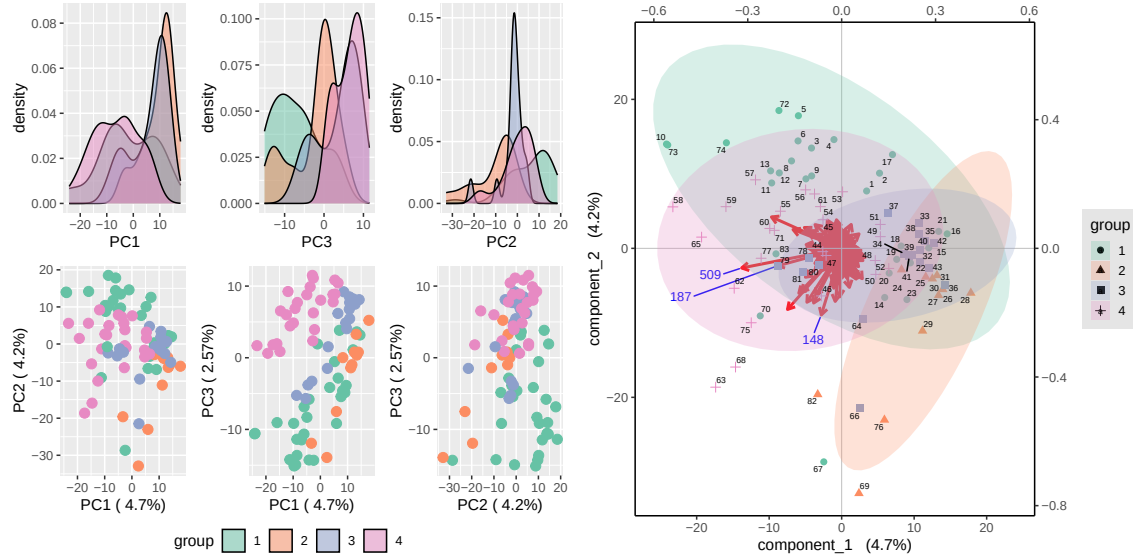
7.2. Real Dataset

7.2.1. Gene expression data from Khan et al. (2001). Gene expression data (2308 genes for 83 samples) from the microarray experiments of Small Round Blue Cell Tumors (SRBCT) of childhood cancer study of Khan et al. (2001).

This dataset contains 83 samples with 2308 genes: 29 cases of Ewing sarcoma (EWS), coded 1, 11 cases of Burkitt lymphoma (BL), coded 2, 18 cases of neuroblastoma (NB), coded 3, 25 cases of rhabdomyosarcoma (RMS), coded 4. A total of 63 training samples and 25 test samples are provided in Khan et al. (2001). Five of the test set are non-SRBCT and are not considered here. The training sample indexes correspond to 1:65 and the test sample indexes (without non-SRBCT sample) correspond to 66:83.

Arguments					Distribution	
cases	g	$mean$	d	σ^2	$\hat{\rho}$	$n.var$
1	(50, 50, 50)	1	0	(1, 1, 1)	(0.5, 0.5, 0.5, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 500, 500, 500, 1000$ $i = 1, 2, 3, 4$
2	(100, 100, 100)	1	0	(2.5, 3.0, 3.5, 1)	(0.5, 0.3, 0.2, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 500, 500, 500, 1000$ $i = 1, 2, 3, 4$
3	(30, 30, 30)	2	0	(4.5, 3.2, 1.6, 1)	(0.8, 0.9, 0.65, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 3000, 3000, 3000, 500$ $i = 1, 2, 3, 4$
4	(50, 50, 50)	4	2	(2.1, 2.3, 2.5, 1)	(0.5, 0.3, 0.2, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 2000, 2000, 2000, 50$ $i = 1, 2, 3, 4$
5	(50, 50, 50)	4	0	(2.1, 2.3, 2.5, 1)	(0.5, 0.5, 0.5, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 2000, 2000, 2000, 1000$ $i = 1, 2, 3, 4$
6	(10, 10, 10, 10)	2	0	(1, 1, 1, 1)	(0.5, 0.5, 0.5, 0.5, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 2000, 2000, 2000, 100$ $i = 1, 2, 3, 4$
7	(10, 10, 10, 10)	2	0	(3, 2.5, 0.5, 1, 1)	(0.5, 0.4, 0.6, 0.5, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 3000, 3000, 3000, 100$ $i = 1, 2, 3, 4, 5$
8	(10, 10, 10, 10)	2	0	(3, 2.5, 0.5, 1, 1)	(0.5, 0.4, 0.6, 0.5, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 100, 100, 100, 100, 3000$ $i = 1, 2, 3, 4, 5$
9	(10, 10, 10, 10)	2	0	(4, 4.5, 5.5, 1, 1)	(0.8, 0.8, 0.8, 0.8, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 100, 100, 100, 100, 3000$ $i = 1, 2, 3, 4, 5$
10	(10, 10, 10, 10, 10)	2	0	(4, 4.5, 5.5, 5.2, 1, 1)	(0.8, 0.8, 0.8, 0.8, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 100, 100, 100, 100, 3000$ $i = 1, 2, 3, 4, 5$
11	(50, 50, 50, 50, 50)	2	0	(4, 4.5, 5.5, 5.2, 1, 1)	(0.1, 0.2, 0.3, 0.4, 0.4, 0)	$\mathbf{X}_i \sim N_l(\mu_l, \Sigma_{l \times l})$ $l = 100, 100, 100, 100, 3000$ $i = 1, 2, 3, 4, 5$

TABLE 7.1. Relationship between a multivariate normal distribution and arguments of the simulation function `fun.sim()`.

FIG. 34. Graphs under the function `pca()` standard by Multigroup0 package.FIG. 35. Graphs under the function `mgPCA()` by Multigroup0 package.

7.2.2. Genes associated with Myeloid leukemia in patients with cytogenetics normal or trisomy on chromosome 8. The main objective of this real data set is to do microarray analysis with Bioconductor and R and apply the methods of the MultiGroupO package. The example analyzed is a simplified version of a study published in journal, PNAS (January 30, 2001 vol. 98 (3)) “Expression profiling reveals fundamental biological differences in acute myeloid leukemia.” In which the profiles of expression characteristics of the disease in patients with trisomy 8 are studied and compared with those of patients with normal cytogenetics. Purified CD34+ cells from bone marrow were also analyzed. normal as representative

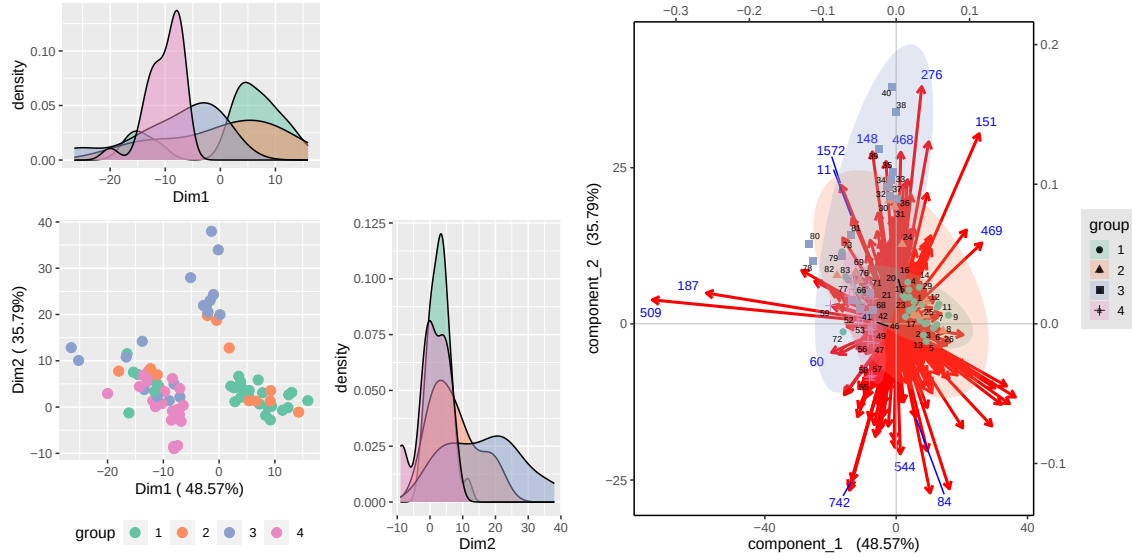


FIG. 36. Graphs under the function `MDR()` by `Multigroup0` package.

of the population not affected by the disease. More details can be obtained in the article available at the link : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC14719/?tool=pubmed> The study was carried out with 48 samples that were hybridized with arrays of affymetrix of the hu6800 type and of which, for this example, a balanced subset of 15 samples, 5 of each type.

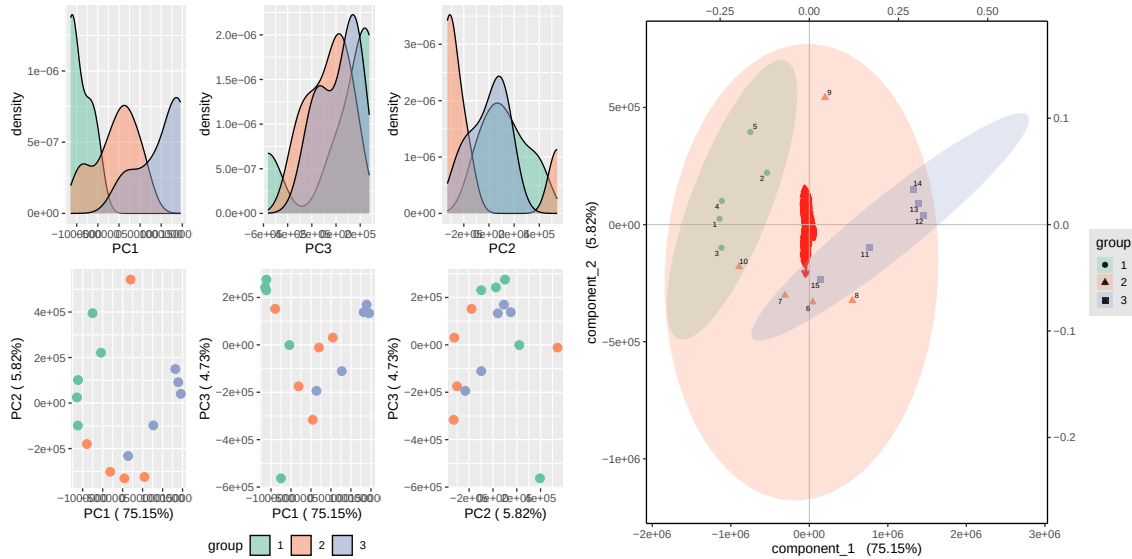
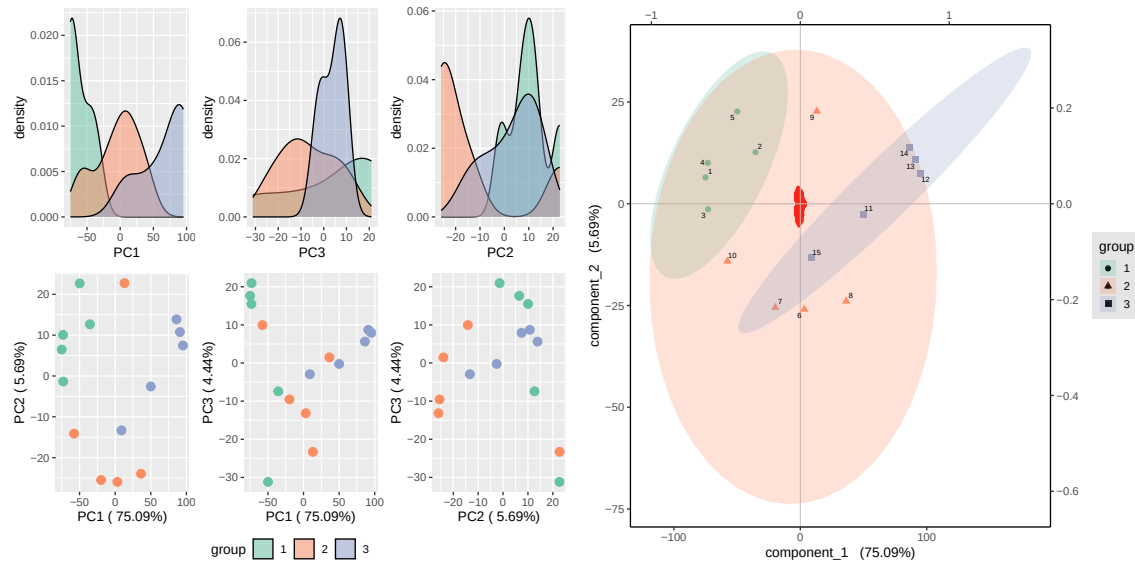
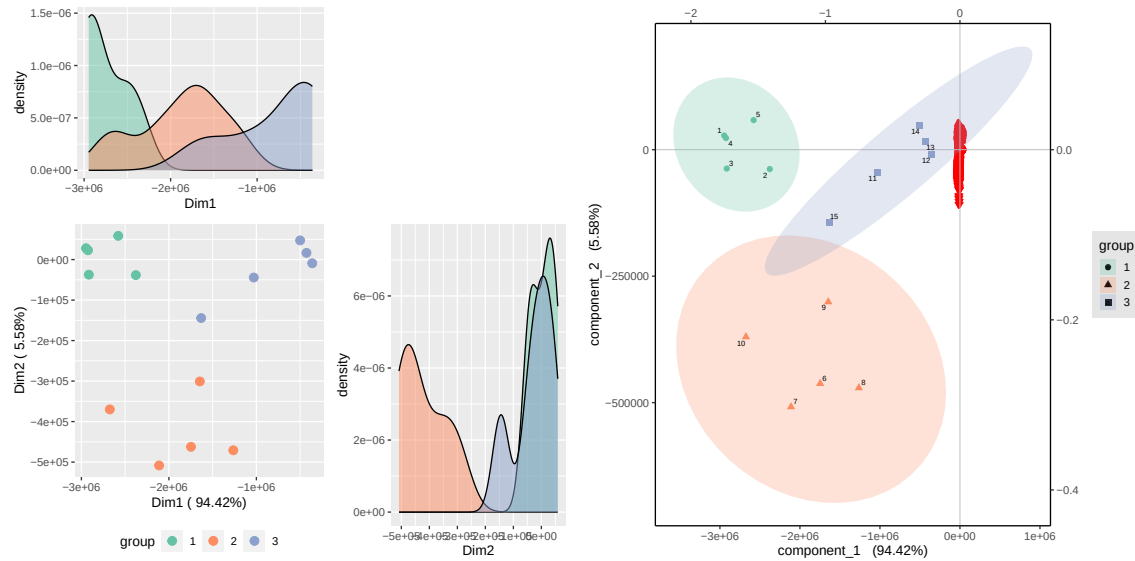


FIG. 37. Graphs under the function `pca()` standard by `Multigroup0` package.

FIG. 38. Graphs under the function `mgPCA()` by Multigroup0 package.FIG. 39. Graphs under the function `MDR()` by Multigroup0 package.

Chapter 8

Discussion and Conclusion

8.1. Discussion

Chapter 3 describes techniques that exist in the literature, e.g., Common Principal Component Analysis (Flury , 1984 [20]), PCA multigroup (Krzanowski, 1984 [21]), DUAL STATIS (Lavit, 1994 [31]), Dual Generalized Procrustes Analysis (Eslami et al., 2013 [23]), Between Groups Comparison [35], Multigroup PCA retrieved (Eslami et al., 2013 [23]) and which are implemented in the `multigroup` package in R (Eslami et al. (2015)). On the other hand, the multigroup RGCCA (Tenenhaus et.al,2011-2014 [38], [24]) method exists in the literature, but it is not implemented in his `RGCCA` package, only the multiblock configuration.

The objective of these methods is to find a compromise variance-covariance matrix and then predict regression models where the groups are considered. Furthermore, existing techniques fail to distinguish groups in reduced dimension representation (see figures Fig. 2., Fig. 3., Fig. 4., Fig. 5. and Fig. 6. in Chapter 3). On the contrary, our methods aim to explore and visualize the data set, under a multigroup configuration, that two objectives have been sought. On the one hand, maximize the variability represented but taking into account the group structure, and also not focus so much on the explained variability but on the ability to separate or distinguish the groups.

The simulation scenarios of our methodologies seek to control the parameters of our interest such as the separation of groups, the degree of noise variables (without biological signal) and the degree of variability and covariability where these types of cases greatly complicate the methodology of the standard PCA.

Our results show us that the classical PCA has difficulties in separating its first two principal components when we simulate under conditions of variables without biological signal, few observations in each group, and large variances in the groups (see some figures 1, 7, 13, 28).

On the other hand, our multigroup PCA method proposed in this thesis shows an improvement in separating several groups with the first two principal components under simulation

scenarios with different conditions, such as having many, or few, variables without biological signals and high variances in the groups. However, it is not a perfect separation in the groups, since there are some that are still coupled, but it is an advance that our multigroup PCA method can detect the groups and separate some of them, as we have seen in the results section (see some figures 2, 8, 14, 29).

However, the standard PCA manages to separate the groups very well with its first two principal components when there are no variables without biological signals and/or high variances in the groups. In addition, it has a low computational cost. Otherwise, our multigroup PCA method has a higher computational cost when analyzing data with high dimensionality, but it provides us with valuable information for visualization and exploration in data sets that have greater complexities in the analysis (see some figures 8, 14, 17, 20, 23, 26).

Our second MDR multigroup method manages to separate the groups quite well under scenarios with many, or few, variables without biological signals and/or high and heterogeneous variances in the groups. This method manages to separate the groups with only two dimensions, which makes it a good method of dimensionality reduction. It is also worth mentioning that it has a high computational cost, but the information it provides is important for more complex datasets (see some figures 12, 15, 24, 27, 30, 33).

Real omics datasets were analyzed, such as gene expression data from Khan et al. (2001) which are microarrays found in the `plsgenomics` package in R and the genes associated with Myeloid leukemia in patients with cytogenetics normal or trisomy on chromosome 8 where we can see that our dimensionality reduction methods manage to detect the groups and separate them better than the standard PCA (see figures 35, 39). It is worth mentioning that it was very difficult to find an omic dataset that represents what we observe under simulation scenarios where one has more control over the dataset.

Our future work would be to analyze datasets that have missing data in each group, since this condition was not analyzed in this thesis. These methods can also be extended to multiblocks, since datasets with different sources of omic variation could be analyzed, in addition to having missing data complexities in the multigroup and multiblock methods.

Other future research could extend these multigroup methods to a kernel methods approach and may give us an improvement to our methods, but in a nonlinear transformations approach.

8.2. Conclusion

This thesis contributes in multigroup approaches in the analysis of high dimensionality data. The main conclusions are:

1. We introduce two multigroup methods that contribute to the literature on multivariate analysis for analyzing complex datasets. That is, they have many variables without biological signals and high variances in the groups. These datasets are very common in the area of Biomedicine, and therefore, would be a great contribution for the visualization and exploration of these omic datasets.
2. An R package was made and published in the CRAN with these new multigroup approaches.
3. Data simulations under certain controlled parameters: our multigroup methods give us good results in achieving separation of several groups with the first two dimensions.
4. The real data that we present in this thesis shows that our multigroup methods achieve improvements in the separation of groups with the first two dimensions and the purpose of dimensionality reduction is achieved. However, our multigroup PCA shows that the groups are not very coupled, but it is not a separation of groups that stop being coupled, however, compared to the classic PCA it achieves a slight improvement in the distances between several groups.
5. Also the real data that is presented when applying our second MDR multigroup method manages quite well to separate the groups with its two dimensions. We see in the figure 39 what makes this multigroup method a good dimensionality reduction approach.

This thesis achieves objectives in presenting and demonstrating contributions in multivariate analysis, specifically, in the multigroup approach with linear transformations. These methods, multigroup PCA and MDR, are techniques that can help us in the analysis of omic datasets when they are presented in a very complex way, that is, many noise variables, high variance in the groups, and few observations with many variables. This research can be extended to many approaches that can be analyzed and followed in the contribution and advances in this area of data science.

My research about the multigroup PCA and MDR methods shows that they are a good alternatives that both contribute and help with the standard PCA in the visualization and exploration of omic data in Biomedicine.

References

- [1] Manzoni, C., D. Kia, J. Vandrovcova, J. Hardy, N. Wood, P. Lewis, and R. Ferrari. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2):286–302, mar 2018.
- [2] Fernández-Arjona, M., J. Grondona, P. Granados-Durán, P. Fernández-Llebrez, and M. López-Ávalos. Microglia Morphological Categorization in a Rat Model of Neuroinflammation by Hierarchical Cluster and Principal Components Analysis. *Frontiers in Cellular Neuroscience*, 11(August):1–22, 2017.
- [3] Soltysik, D., D. Thomasson, S. Rajan, and N. Biassou. Improving the use of principal component analysis to reduce physiological noise and motion artifacts to increase the sensitivity of task-based fMRI. *Journal of Neuroscience Methods*, 241:18–29, 2015.
- [4] Jolliffe, I. and J. Cadima. Principal component analysis : a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.
- [5] Vasseur, F., T. Bontpart, M. Dauzat, C. Granier, and D. Vile. Multivariate genetic analysis of plant responses to water deficit and high temperature revealed contrasting adaptive strategies. *Journal of Experimental Botany*, 65(22):6457–6469, 2014.
- [6] Wang, Ch., J. Degnan, T. Pemberton, N. Rosenberg, Z. Szpiech, M. Jakobsson, J. Hardy, and A. Singleton. Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.
- [7] Xiong, H., D. Zhang, C. Martyniuk, V. Trudeau, and X. Xia. Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data. *BMC Bioinformatics*, 9:25, 2008.
- [8] R. Powers Worley, B. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.
- [9] Bartenhagen, C., H.U. Klein, C. Ruckert, X. Jiang, M. Dugas. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics*, 11:567, 2010.
- [10] Lenz, M., F.J. Muller, M. Zenke, A. Schuppert. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Scientific Reports*, 6:25696, 2016.
- [11] Costa, J., P. Da-Silva, R. Almeida, and A. Infantosi. Validation in Principal Components Analysis Applied to EEG Data. *Computational and Mathematical Methods in Medicine*, 2014:1–10, 2014.
- [12] Yesavage, J., J. Taylor, L. Friedman, P. Rosenberg, L. Lazzeroni, J. Leoutsakos, L. Kinoshita, M. Perlow, C. Munro, D. Devanand, L. Drye, J. Mintzer, B. Pollock, A. Porsteinsson, L. Schneider, D. Shade, D. Weintraub, C. Lyketsos, A. Noda, and the CitAD Research Group. Principal components analysis of agitation outcomes in Alzheimer’s disease. *Journal of Psychiatric Research*, 79:4–7, 2016.
- [13] Fraix-burnet, D., M. Thuillard, and A. Chattopadhyay. Multivariate Approaches to Classification in Extragalactic Astronomy. *Frontiers in Astronomy and Space Sciences, Frontiers Media*, 2(3), 2015.

- [14] Saukani, N. and N. Azina. Identifying the Components of Social Capital by Categorical Principal Component Analysis (CATPCA). *Social Indicators Research*, 2018.
- [15] Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- [16] Hotelling, H. Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology*, 24:417–441, 1933.
- [17] Mishra, P., A. Medlar, L. Holm, and P. Törönen. Robust multi-group gene set analysis with few replicates. *BMC Bioinformatics*, 17(1):526, 2016.
- [18] Eslami, A., E. Qannari, A. Kohler, and S. Bougeard. Multivariate analysis of multiblock and multigroup data. *Chemometrics and Intelligent Laboratory Systems*, 133:63–69, 2014.
- [19] Abdi, H., L. Williams, and D. Valentin. Multiple factor analysis: Principal component analysis for multi-table and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179, 2013.
- [20] Flury, B.N. Common principal components in k groups. *Journal of the American Statistical Association*, 79:892–898, 1984.
- [21] Krzanowski, W.J. Principal Component Analysis in the Presence of Group Structure. *Journal of the Royal Statistical Society*, 33:164–168, 1984.
- [22] Niesing, J. *Simultaneous component and factor analysis methods for two or more groups: A comparative study*. PhD thesis, Leiden University, 1997.
- [23] Eslami, A., E. Qannari, A. Kohler, and S. Bougeard. General overview of methods of analysis of multi-group datasets. *Hdsda*, pages 108—123, 2011.
- [24] Tenenhaus, A. and Tenenhaus, Michel. Regularized generalized canonical correlation analysis for multi-block or multigroup data analysis. *European Journal of Operational Research*, 238(2):391–403, 2014.
- [25] Gower, J. C. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.
- [26] Lê, S. and J. Pagés. DMFA: Dual multiple factor analysis. *Communications in Statistics - Theory and Methods*, 39(3):483–492, 2010.
- [27] Takane, Y. and T. Shibayama. Principal component analysis with external information on both subjects and variables. *Psychometrika*, 56(1):97–120, 1991.
- [28] Chen, X., X. Chen, R. Ward, and Z. Wang. A Joint Multimodal Group Analysis Framework for Modeling Corticomuscular Activity. *IEEE - Institute of Electrical and Electronics Engineers, Inc.*, 15:309–312, 2013.
- [29] Khan, J., Jun S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.
- [30] Browne, R. and P. McNicholas. Estimating Common Principal Components in High Dimensions. 2013.
- [31] Lavit, C., Y. Escoufier, R. Sabatier, and P. Traissac. The ACT (STATIS method). *Computational Statistics Data Analysis*, 18:97–117, 1994.
- [32] Escoufier, Y. Le traitement des variables vectorielles. *Biometrics*, 29:751–760, 1973.
- [33] Glacon, F. *Analyse conjointe de plusieurs matrices de données: comparaison de différentes méthodes*. PhD thesis, University of Grenoble, 1981.
- [34] Ten Berge, J. Generalized procrustes analysis. *Psychometrika*, 42:267–276, 1977.
- [35] Krzanowski, W.J. Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association*, 74:703–707, 1979.

- [36] Carroll, J. D. Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 3:227–228, 1968.
- [37] Eckart, C. and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.
- [38] Tenenhaus, A. and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76:257–284, 2011.
- [39] Alanni, R., J. Hou, H. Azzawi, and Y. Xiang. A novel gene selection algorithm for cancer classification using microarray datasets. *BMC Medical Genomics*, 12(1), jan 2019.
- [40] Virtaneva, K., F. Wright, S. Tanner, B. Yuan, W. Lemon, M. Caligiuri, C. Bloomfield, A. Chapelle, and R. Krahe. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Biological Sciences*, 3:1124–1129, 2001.