TÉCNICO LISBOA

Unveiling Interpretable Behavior in Two-Way High-Dimensional Clinical Data

Luís Bernardo de Brito Mendes Rei

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Alexandra Sofia Martins de Carvalho
Prof. Susana de Almeida Mendes Vinga Martins

Examination Committee

Chairperson: Prof. António Manuel Raminhos Cordeiro Grilo
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Prof. Adelaide de Fátima Baptista Valente Freitas

November 2018

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

For those who never cease to be curious about how the world works.

Acknowledgments

I would like to thank my supervisor, Prof. Dr. Alexandra Carvalho and co-supervisor, Prof. Dr. Susana Vinga, for their bountiful patience, guidance and time invested during the entire course of this thesis. I feel extremely fortunate to have worked under their supervision.

Many people have been, in one form or another, important in the development of this work. Notably, the Post-Doc and PhD students at CSI/IDMEC and IT, especially André Veríssimo and Marta Lopes, by giving an essential contribution in helping me produce the simulation results, and Prof. Dr. Eloísa Macedo and Prof. Dr. Adelaide Freitas, by sharing their knowledge in the field of numerical methods. I would also like to acknowledge IST and its electrotechnical engineering department for their hospitality and material.

A kind word goes to all my friends and colleagues who shared the last five years of this unique challenge with me and made the hard work more enjoyable. In particular to João Girão, Gonçalo Duarte, Catarina Campos, João Belfo and Bruna Mason who in several occasions, during the development of this thesis, have helped with their sharp and sincere judgement.

Last, but not the least, I would like to express my sincere gratitude towards my family, for their advice and support. A special word to my parents, Luís and Cidália, who made my education a priority, and who, together with my brother João and partner Margarida, have contributed more than anyone else to the person I am today.

This thesis has been performed in the framework of project PTDC/EMS-SIS/0642/2014, funded by the Portuguese Foundation of Science and Technology (FCT).

Resumo

O desenvolvimento de métodos de aprendizagem automática e a sua adequação a problemas clínicos têm possibilitado a criação de novas abordagens terapêuticas que fazem perspetivar a aplicação de soluções de engenharia para modelar integradamente sistemas fisiológicos multi-escalares, fornecendo conhecimento profundo e abrangente do funcionamento de sistemas biológicos. Os sistemas adaptativos de apoio à decisão clínica para a medicina personalizada sofrem de um problema de elevada dimensionalidade, já que contemplam o ajuste de muitos parâmetros. Este relatório apresenta o estudo teórico e a exploração prática de técnicas de aprendizagem não supervisionada, bem como a revisão de metodologias de agrupamento capazes de lidar com dados de grande dimensão. As qualidades das abordagens tandem tradicionais são debatidas através da avaliação do seu desempenho em dados sintéticos e reais. A pesquisa levada a cabo abre espaço à criação de novas estratégias integradas que conjugam a redução do espaço de variáveis com a estratificação dos objetos para maximizar a interpretabilidade dos dados e facilitar a sua análise. Neste trabalho um modelo difuso entropicamente regularizado é incorporado numa metodologia de *clustering* e análise de componentes principais disjuntos e é comparado com outras metodologias de última geração, mostrando trazer mais intuição à apreciação dos resultados fruto da palete de cores atribuída às observações com base nos graus de pertença aos respetivos grupos. É também apresentada uma nova ferramenta hierárquica capaz de desvendar ciclicamente informação oculta nas camadas mais profundas dos dados através do rearranjo dos subespaços de variáveis para reavaliação de *clusters*.

Palavras-chave: Aprendizagem automática, Estatística multivariada, Dados de alta dimensionalidade, Análise difusa de agrupamentos, Análise de Componentes Principais.

Abstract

The development of machine learning methods and their adaptation to clinical problems have enabled the creation of new therapeutic approaches that lead to the application of engineering solutions to model multi-scalar physiological systems in an integrated way, providing deep and comprehensive knowledge of how biological systems work. Adaptive clinical decision support systems for precision medicine suffer from a problem of high dimensionality, since they contemplate the adjustment of many parameters. This report presents the theoretical study and the practical exploration of unsupervised learning techniques of features, as well as the revision of clustering methodologies capable of handling large data. The qualities of traditional tandem approaches are debated by evaluating their performance in synthetic and real data. The research carried out opens space for the creation of new integrated strategies that combine the reduction of the space of variables with the stratification of the objects to maximize the interpretability of the data and to facilitate their analysis. In this work an entropy-regularized fuzzy model is incorporated into a clustering and disjoint principal component analysis method and is successfully matched against other state of the art methodologies, showing improved intuition in the appreciation of the results due to the color palette attributed to the observations based on their degrees of belonging to the respective groups. Also presented in this report is a new hierarchical tool capable of cyclically uncover hidden information in the deeper layers of the data by rearranging subspace data for re-evaluation of clusters.

Keywords: Machine learning, Multivariate statistics, High-dimensional data, Fuzzy cluster analysis, Principal Component Analysis.

Contents

List of Figures

List of Tables

List of Abbreviations

ALS Alternating Least-Squares.

ARI Adjusted Rand Index.

CDPCA Clustering and Disjoint Principal Component Analysis.

EFA Exploratory Factor Analysis.

FA Factor Analysis.

FCM Fuzzy C-Means.

FE Feature Extraction.

FKM Factorial K-Means.

FOM Figure of Merit.

FS Feature Selection.

LDA Linear Discriminant Analysis.

MEI Maximum-Entropy Inference.

NCDPCA Nested Clustering and Disjoint Principal Component Analysis.

PCA Principal Component Analysis.

RCDPCA Relaxed Clustering and Disjoint Principal Component Analysis.

RKM Reduced K-Means.

TCGA The Cancer Genome Atlas.

Notation

For the convenience of the reader the notation and terminology common to all sections is listed here.

I Number of objects, indexed $i = 1, ..., I$.

J Number of variables, indexed $j = 1, ..., J$.

P Desired number of clusters of objects, indexed $p = 1, ..., P$.

Q Desired number of subsets of variables, indexed $q = 1, ..., P$.

$\mathbf{X}$ Standardized data matrix with I objects in rows and J variables in columns.

$\mathbf{U}$ $I \times P$ binary and row stochastic matrix defining an allocation of the I objects into P clusters. Matrix entry is 1 if the i-th object belongs to the cluster p, 0 otherwise.

$\mathbf{V}$ $J \times Q$ binary and row stochastic matrix defining an allocation of the J variables into Q subsets. Matrix entry is 1 if the j-th object belongs to the cluster q, 0 otherwise.

$\bar{\mathbf{X}}$ $P \times J$ object centroid matrix in the original space defined by $\bar{\mathbf{X}} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{X}$.

$\mathbf{Z}$ $I \times J$ centroid-based data matrix where each object is identified by the corresponding centroid, $\mathbf{Z} = \mathbf{U}\bar{\mathbf{X}}$.

$\mathbf{W}$ $I \times K^{(q)}$ submatrix extracted from the centroid-based data matrix $\mathbf{Z}$ where only the original variables assigned into the q-th column of $\mathbf{V}$ are considered.

$$w_{ik}^{(q)} = z_{ij}, \quad \text{if } v_{jq} = 1, \quad \text{with } k = \text{rank}_{J^{(q)}}(j),$$

where $J^{(q)} = j : v_{jq} = 1$, $K^{(q)} = \#J^{(q)}$ and $k = 1, ..., K^{(q)}$.

$\mathbf{A}$ $J \times Q$ matrix of the component loadings where the Q columns are identifying the coefficients of Q linear combinations such that rank($\mathbf{A}$) = Q, $\mathbf{A}^T\mathbf{A} = I_Q$ and $\sum_{j=1}^{J}(a_{jq}a_{jr})^2$, for any q and r ($q \neq r$).

$\mathbf{Y}$ $I \times Q$ component score matrix given by $\mathbf{Y} = \mathbf{XA}$ where y_{iq} is the value of the i-th object for the q-th CDPCA component.

$\bar{\mathbf{Y}}$ $P \times Q$ object centroid matrix in the reduced space given by $\bar{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{A}$.

Furthermore, boxes containing definitions, theorems and algorithms are utilized throughout the document in order to introduce important mathematical concepts and information. These boxes can be easily distinguished by their background color: yellow, green and blue, respectively.

1

Introduction

Contents

The aim of this thesis is to develop novel statistical and machine learning methodologies to deal with the issues prevalent with precision medicine - a medical model for disease treatment that takes into account individual variability in genes, environment, and lifestyle for each person. This approach will allow doctors and researchers to predict more accurately which treatment and prevention strategies for a particular disease will work better in which groups of people; contrary to a one-size-fits-all approach, in which strategies take less consideration for the differences between individuals and are developed for the average person.

This work has a strong relation with real-life applications as discussed in Section 1.1. The layout of the main goals of this thesis is done in Section 1.2, followed by the main original contributions in Section 1.3 and a general outline of this document in Section 1.4.

1.1 Context and Motivation

Due to recent progress in data storage and acquisition, an increasing number of databases are emerging, as computerization in health care services and the amount of available digital data grows at an unprecedented rate [1]. Considering that the use of computer and information technologies in health care services can help achieve efficiency and effectiveness in diagnostic decision making, cost economy, and better risk management and strategic planning in a competitive environment [2], it becomes increasingly important to retrieve knowledge from these data repositories, especially while health care organizations are facing a major challenge on improving the quality of the service delivered, while maintaining the costs affordable [1].

The rise of genomics and the accumulation of heterogeneous amounts of biomedical data is inciting the development of new systems-based approaches to life sciences, creating a substantial need for flexible data modeling and analysis tools to help retain useful insights from the overwhelming size and dimension of the obtained data. The computing mechanism of distinguishing patterns in large data sets is denominated data mining and involves methods crossing statistics, machine learning and database systems [3]. The process encompasses the extraction of data patterns through the use of intelligent methods with the goal of distilling relevant information from a data set and transform it into an interpretable structure for further use [4]. The use of data mining techniques on clinical data has the potential to improve decision making in diagnosis, find ways of preventing some diseases, towards a better patients' care.

Personalized medical therapies hold the promise of a tailored service and treatment based on information that is patient-specific. Modeling complex pathologies like cancer and contributing to this therapy's optimization constitutes a great challenge in systems medicine, whose results are expected to have high social and economical impact. In this context, biomarkers information and indicators of disease progression can lead to the identification of co-variables related to the disease outcome, helping unravel relationships in the input data space to diminish the complexity of the task at hand and elucidate the intricate connections of different types of epigenomic abnormalities.

1.2 Objectives

Although some efforts of applying engineering methods to model multi-scale physiological systems in an integrative way have been made [5–9], a definition of the best mathematical strategy, its computational implementation and its effective application to real clinical problems, such as patient profiling and therapy optimization, remain elusive and constitute the main motivation and innovative nature of this topic of research.

This report reviews the problems in prognostic medicine where statistical and machine learning methods techniques that tackle high-dimensional clinical data may be of use. The conducted research opens room to the creation of novel modeling and decision techniques to be used in systems medicine conjugating the reduction of the variables space and the stratification of objects, and able to maximize the interpretability of the data and consequently improve analysis.

As the discovery of disease subtypes via the exploration of gene expression data using unsupervised clustering methodologies is one of the most important research areas in personalized medicine, one of the other goals settles on producing new frameworks capable of highlighting information hidden in inner layers of data. The otherwise obscured knowledge can then be used to, for example, extrapolate new behaviors through the analysis of partitions of different types of cancers of The Cancer Genome Atlas [1].

1.3 Original Contributions

An entropy-regularized fuzzy model for the clustering and disjoint principal component analysis method is developed and matched against other state of the art methods. The introduction of a relaxation in the object assignment phase permits a more just classification and more accurate interpretation. A regularization is performed following an entropy approach as it allows to ditch the cryptic fuzzyfication tuning parameter and adds missing meaningful physical properties to the simpler fuzzy k-means methodology. The new model is tested and compared to others introduced in Chapters 3 and 4.

A technique is developed to disclose otherwise obstructed information in the data following a nested, cyclical pipeline that digs deeper and deeper into the hidden layers of the data and continuously rearranges the linear combinations of the features to secure the optimal between cluster deviance of the object clusters and subclusters. The new technique is applied to real clinical data and compared to a similar tandem approach.

1.4 Document Outline

A brief overview of the most relevant state of the art methods can be read in Chapter 2. The qualities of traditional tandem approaches are disputed using synthetic and real data, and comparisons are made to two other solutions in Chapter 3. Chapter 4 introduces a classical cluster partition approach to an integrated dissection of diseases and individuals into homogeneous groups with similar survival responses,

[1] More about TCGA at https://cancergenome.nih.gov/

and sets the foundation for the proposed methodologies detailed in Chapter 5. Chapter 6 reveals the analysis of several data sets and the performance results of the new frameworks. The final conclusions are taken and further possible lines of developments are suggested in Chapter 7.

2

Literature Review

Contents

This chapter provides a brief overview of the most relevant methods to perform dimensionality reduction and clustering of objects.

To combat the curse of dimensionality prevalent in high-dimensional clinical data [10], where the number of attributes clearly outnumbers the number of samples, the original data space needs to be reconfigured to be able to be expressed in a lower-dimensional setup. As the data grows bigger in size the interpretability concern becomes ever more relevant and there is a need to cleverly generate the feature subset. Dimensionality reduction is the process of reducing the number of variables under consideration by obtaining a set of principal relevant features, and can be divided into feature selection and feature extraction methods [11].

2.1 Classification of Feature Selection Methods

Feature selection or variable subset selection methods can be used in data preprocessing to achieve efficient data reduction as they attempt to find a subset of non-redundant features to construct a model whose simplification makes them easier to analyze and interpret. The reduction of complexity allows for faster training times, less propension to overfitting, and minimizes the curse of dimensionality [12]. These algorithms are separated in three classes: filters, wrappers, embedded/hybrid methods [13, 14]. A brief overview of them follows.

2.1.1 Filter Methods

Filters incorporate an independent measure for evaluating subsets without involving any subsequent learning algorithm, discarding the least interesting variables according to said independent criterion. The developed measures for feature filtering can be classified into information, distance, consistency, similarity, and statistical measures [13].

Only after it finds the best features can the employed data modeling algorithm use them, Fig. 2.1. Despite this approach being computationally inexpensive, the method can miss features that are not useful by themselves but can be very useful when in conjunction with others, making this the worst performing class of methods [14].

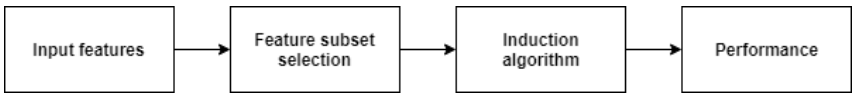


Figure 2.1: Filter method for feature selection.

Filters can operate in two ways: they can evaluate a single feature at a time - in which case the method is univariate - or they can evaluate the entire subset - multivariate methods [15]. Feature subset generation for multivariate filters depends on the search strategy and, although there are many search strategies, the four usual starting points for feature subset generation are: forward selection, backward elimination, bidirectional selection, and heuristic feature subset selection.

2.1.2 Wrapper Methods

In wrapper methods, a subset of features is generated and a model is trained according to that generation [16]. Based on inferences drawn from the previous model, features are added or removed from the previous subset, Fig. 2.2. This subset adequacy evaluation through a learning algorithm (cross-validation) reduces the problem to a search problem, and makes this class of feature selection methods the one with usually the best generalization ability [14].

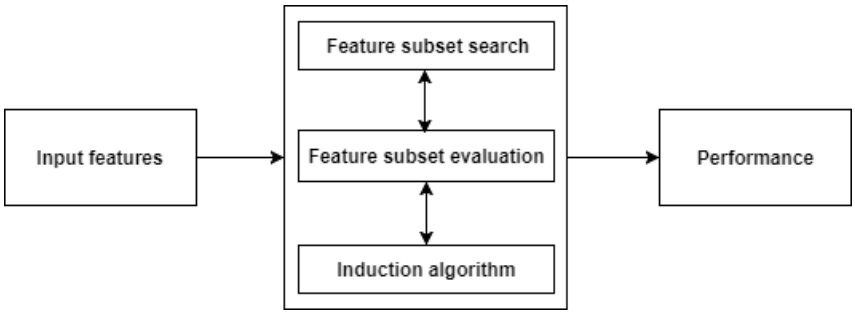


Figure 2.2: Wrapper method for feature selection.

Regarding high-dimensional data, using search-based or wrapper techniques, and many filter methods based on redundancy assessment, can be computationally prohibitive due to model training [14].

2.1.3 Embedded and Hybrid Methods

Embedded methods combine filter and wrapper methods to get the best of both worlds - obtain the low complexity associated with filters and the high accuracy of wrappers [17]. They consider not only relations between input and output feature, but also indulge in local searches for features that allow better local discrimination. The method does this by utilizing an independent criterion to decide the optimal subsets for a known cardinality, and then deciding on a final optimal subset through the evaluation of the learning algorithm among the optimal subsets across different cardinalities.

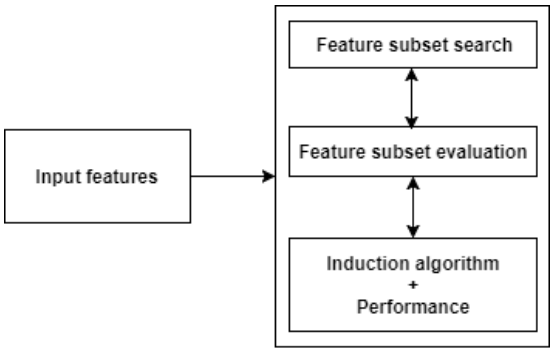


Figure 2.3: Embedded method for feature selection.

Note that the features are selected as part of a learning algorithm, meaning the subset of features is selected in conjunction with the classifier. A shortcoming of this approach is the fact that the selected subset becomes classifier dependent, meaning they may not work with any other classifier [15]. An extensive analysis of hybrid models for feature selection can be found in [18].

2.2 Feature Extraction Techniques

This variable reduction approach is a dimensionality reduction technique that works by performing a linear or non-linear transformation on the input space to obtain a new, lower-dimensional feature space. In this section some of these techniques are highlighted. A more extended listing of relevant methodologies can be found here [14].

2.2.1 Linear Discriminant Analysis

The statistical procedure referred to as LDA is a generalization of Fisher's linear discriminant, a supervised method used to find a linear combination of features that characterizes or separates two or more classes of objects. It operates on continuous-valued variables and frequently uses a multivariate Gaussian distribution for the class conditional density. Assuming examples within one class or the other are normally distributed (linearity) it analytically determines the best separating hyperplane between the classes from their means and variances. This method aims to find a projection A that maximizes the ratio of between-class variance S_b to the within-class variance S_w in any particular data set (Fisher's criterion),

$$\arg\max_A \frac{|AS_bA^T|}{|AS_wA^T|}.$$

In LDA, data sets can be transformed and test vectors can be classified in the transformed space by a different approach: class-independent transformation [19]. In this approach, each class is considered as a separate class against all others, and involves maximizing the ratio of the overall variance to within class variance. Because there is only one optimizing criterion used all data points irrespective of their class identity, will be transformed by the same operation.

Exploratory analysis is engaged by applying the initially described method to the *Iris* dataset. The linear discriminants' (LD) loadings are explicit in Table 2.2 and a confusion matrix fabricated after performing the most discriminative linear projection between classes is shown in Table 2.3. The tests were done using the *lda* function of the *MASS* package in R.

Table 2.1: *Iris* dataset group mean.

Species	Attributes [cm]			
	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.006	3.428	1.462	0.246
Versicolor	5.936	2.770	4.260	1.326
Virginica	6.588	2.974	5.552	2.026

The simplicity of the dataset chosen allows for a clearer interpretation of the results and facilitates the comparison with other analysis typologies. It is easily attested that the classification produced was almost ideal, but linear discriminant analysis holds some limitations not reproducible with the chosen dataset. Because it is limited by the number of classes on the generation of the number of discriminative axes, when facing a situation where the number of features outnumbers the number of class labels the

Table 2.2: Linear discriminants' coefficients.

Attributes	Linear Discriminants	
	Component 1	Component 2
Sepal Length	-0.829	0.024
Sepal Width	-1.534	2.165
Petal Length	2.201	-0.932
Petal Width	2.810	2.839

Table 2.3: Confusion matrix after LDA's projection.

Species	Predicted Species		
	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	1	49

loss of information becomes too noticeable, and the classification accuracy inherent to the use of this technique plunges, bearing this method unusable.

Due to computational limitations it is commonly believed that a direct LDA solution for high-dimensional data is infeasible. However, modifications of the simultaneous diagonalization procedure permit to discard the null space of S_b – which carries no discriminative information – and to keep the null space of S_w, which is very important for classification. The result is a unified LDA algorithm that gives an exact solution to Fisher's criterion whether or not S_w is singular (case when data matrices are degenerate) [20].

2.2.2 Principal Component Analysis

The method also mentioned to as PCA is an unsupervised feature extraction method and a statistical procedure that uses an orthogonal transformation to convert a set of values of possibly correlated variables into a set of observations of linearly uncorrelated variables called principal components. This transformation generates a new coordinate system whose axes can project the data into a lower dimensional space, defined in such a way that the first principal component has the largest possible variance, assuring minimum information loss. Each succeeding component, in turn, has the highest variance possible under the constraint that it is orthogonal to the preceding components [21].

PCA as an exploratory tool to develop predictive models can be achieved by eigendecomposition of a data correlation or covariance, or by singular value decomposition of the centered and normalized data matrix [22]. More precisely, letting p be the number of dimensions of each data sample, N the number of samples, and $\mathbf{X}$ the $N \times p$ data matrix, the symmetric covariance matrix $\mathbf{C_X}$ can be eigendecomposed to

$$\mathbf{C_X} = \mathbf{UVU}^T, \tag{2.1}$$

where the columns of $\mathbf{U}$ contain u_d eigenvectors and $\mathbf{V} = diag(\lambda_1, ..., \lambda_d)$ with λ_i the i-th eigenvalue.

Sorting the resulting matrix columns by decreasing eigenvalues results in the eigenvector or projection matrix E. The dimension reduction can then be obtained by projecting $\mathbf{X}$ on the recently calculated orthogonal eigenvectors, and selecting the biggest p' components to represent the data. This selection allows the data to be molded into a lower dimensional feature space since $p' < p$, and the new representation of the data can be achieved by a $N \times p'$ matrix $\mathbf{Y}$. An important property of this algebraic application is that the resulting principal components are obligatorily orthogonal and their variance are given by the corresponding eigenvalue.

After applying this methodology to the *Iris* dataset the principal components (PC) are generated and displayed in Table 2.4, and their respective variances is detailed in Table 2.5. This information was obtained applying the PCA method available from the *FactorMineR* package.

Table 2.4: Principal components' loadings.

Attributes	Component			
	PC1	PC2	PC3	PC4
Sepal Length	0.5211	-0.3774	0.7196	0.2613
Sepal Width	-0.2693	-0.9233	-0.2444	-0.1235
Petal Length	0.5804	-0.0245	-0.1421	-0.8014
Petal Width	0.5646	-0.0669	-0.6343	0.5236

Table 2.5: Principal components' variance.

Measure	Component			
	PC1	PC2	PC3	PC4
Standard Deviation	1.7084	0.9560	0.3831	0.1439
Proportion of Variance (%)	73.3	22.7	3.6	0.4
Cumulative Proportion (%)	73.3	96.0	99.6	100

The first two principal components are responsible for explaining approximately 95% of the variance. In a low dimensional space such as this one, this measure carries some significance and the importance of the first two principal components can be assumed to be high, and sufficient to explain closely enough the entirety of the original data. In Fig. 2.4, a visualization technique known as biplot is used to draw a low dimensional graph in order to find patterns hidden on data and to interpret relationships between samples and variables. The visualization was obtained using the *ggbiplot* package in R.

The summarization in the same plane of the classified data and the loadings' weights of Table 2.4 associated with a certain input variable allow by visual inspection to conclude that the loadings related to the Petal have a strong negative influence in the first component, but carry no relevance in the second. Thus not having almost any impact on the growth of the explained variance from the the previous component to the latter.

Comparing to LDA, the supervised learning mechanism holds final results closer to the truth due to the existence of *a priori* knowledge, as explained in Subsection 2.2.1. The main difference between the

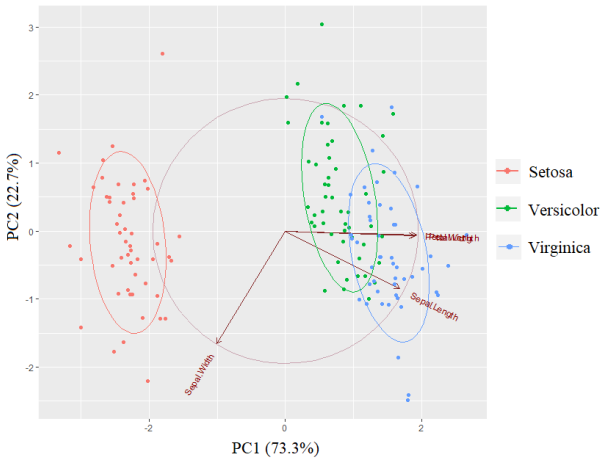
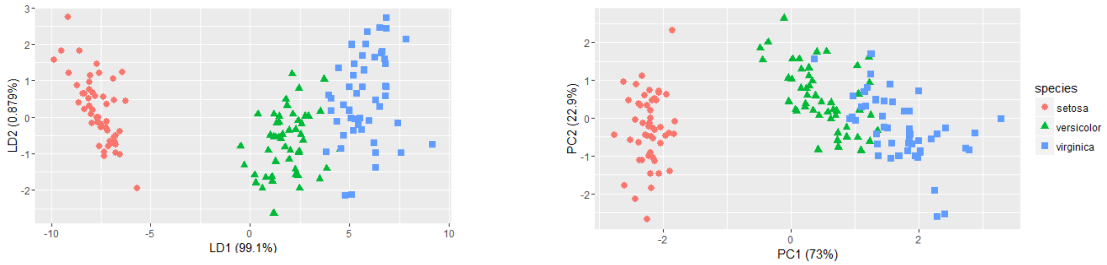


Figure 2.4: Biplot of the *Iris* data after performing PCA.



(a) **(b)**

Figure 2.5: *Iris'* classification after projecting data into a lower-dimension space. (a) Reduced-rank discriminant analysis; (b) Principal Component Analysis.

two methods is that PCA does more of feature classification, reshaping and relocating the original data sets when transformed to a different space, while LDA concentrates on data classification, not changing the location but simply trying to provide more class separability and draw a decision region between the given classes.

There is no unique way of determining the optimal number of principal components to keep, as the application of PCA is intensely dependent on the problem we face and the solution we aim at achieving. Despite all of that, statistically sane solutions to this issue have been presented, namely cross-validation relating to covariance structure modelling methodologies such as the one seen in [23] and [24].

One term often confused with the one analyzed in this section is Factor Analysis; more specifically the notion of Exploratory Factor Analysis (EFA) - the attempt to discover the nature of the constructs influencing a set of responses - creates some confusion. The two can be distinguished because they are based on different models. The first difference is that the direction of influence is reversed: EFA assumes that the measured responses are based on the underlying factors while in PCA the principal components are based on the measured responses. The second difference is that EFA assumes that the variance in the measured variables can be decomposed into that accounted for by common factors and that accounted for by unique factors, while principal components are simply defined as linear combinations of the measurements, containing both common and unique variance. For more information regarding this methodology refer to [25, 26].

2.2.3 Compendium of Related Feature Extraction Work

Since each principal component is a linear combination of all the original variables, carrying in a practical situation no nonzero loadings, the component interpretation task is exceptionally toughened. To overcome this shortcoming, various PCA-based methodologies have been proposed based on the application of new rotation techniques or by inducing the inclusion of zeros in the obtained components. Regarding the latter, a Simple Principal Component methodology [27] has been proposed, restricting the components' loadings to be equal to the constants -1, 0, or 1. In 2003, SCoTLASS [28] was introduced as a maximal variance approach that obtains components where a L_1 bound is introduced on the sum of the absolute values of the loadings, some becoming null. To avoid outlier corruption of the mathematically optimal method trying to be found, an alternative weighted generalization of PCA [29] was built to increase robustness by assigning different weights to data objects based on their estimated relevancy. In the same tone, Robust Principal Component Analysis (RPCA) is presented to work with grossly corrupted observations via a decomposition in low-rank and sparse matrices modification on the original statistical procedure [30]. Later, the notion of sparsity is reinforced in Sparse Principal Component Analysis [31]; idea already expanded in the form of Sparse Principal Component Analysis and Iterative Thresholding, a new iterative thresholding approach for estimating principal subspaces in the setting where the leading eigenvectors are sparse. It was proven under a spiked covariance model that the new approach recovers the principal subspace and leading eigenvectors consistently (and optimally), in a range of high-dimensional sparse settings [32]. Based on linear methods are other non-linear developments that can be broadly classified into two groups: those that just serve as some sort of visualization tool, and those that provide a mapping (from high-dimensional space to low-dimensional embedding or vice versa). Examples include the Kernel PCA [33] which transports the originally linear operations of PCA in a reproducing kernel Hilbert space, the locally-linear embedding [34], and others [35, 36].

2.3 Cluster Analysis

The process of grouping a set of objects into classes of similar objects is called clustering. In the field of machine learning, clustering is a case of unsupervised learning, not relying on predefined classes or class-labeled training examples, being a form of learning by observation rather than learning by examples. This adaptable solution that separates data guided by a certain criterion can be viewed as a data compression technique as it allows for the consideration of singular groups, disparate to the remaining assemblies under a defined metric. Many clustering algorithms exist in the literature - including the partitioning and hierarchical methodologies most often used in gene expression data study - and it becomes difficult to provide a precise categorization of them all because these categories may overlap and a method may have properties from several categories [37].

2.3.1 Partitioning Methods

Given a data set of n objects and k clusters to form, a partitioning algorithm organizes the objects into k clusters ($k \leq n$), where each partition represents a cluster formed to optimize an objective partitioning criterion, like a distance-based measurement, so that the objects within a cluster are "similar", whereas the objects of different clusters are "dissimilar" in respect to data set attributes. The most well-known and commonly applied partitioning methods are k-means, k-medoids, and their variations [4].

Centroid-based Technique The k-means algorithm partitions a set of n objects into a specified-number k of clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. In this particular implementation the objective partitioning criterion is based on the mean value of the objects included in a given cluster, which serves as the base to determine the cluster's centroid.

Firstly, it randomly selects k of the objects, each of which initially representing a cluster center. For each of the remaining objects, an object is assigned based on the distance between the object and the cluster mean to the cluster to which it is the most similar. It then computes the new mean for each cluster. This iterative process continues until the criterion function converges. Frequently, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2, \tag{2.2}$$

where E is the sum of the square error for all objects in the data set, p is the point in space representing a given object, and m_i is the centroid of cluster C_i (p and m_i are multidimensional). This is to say that for each object in each cluster the distance between the object and its cluster center are squared and summed. This function tries to make the resulting k clusters as compact and as separate as possible.

There are a few variants of the k-means method, differing in the selection process of the initial k means, the strategy for the computation of dissimilarity, and the strategies for calculating the cluster centers. One alternative is to define at the beginning the number of clusters and the initial clustering through an hierarchical agglomeration algorithm [38], and then use iterative relocation to improve the clustering.

Representative Object-based Technique One of the faults of the k-means algorithm is its sensitivity to outliers; an extremely large valued object may substantially distort the distribution of data. This effect is exacerbated by the use of the square-error function referred earlier in 2.2.

The basis of the k-medoids method for grouping n objects into k clusters starts by selecting, using one representative object per cluster, actual objects to represent the clusters instead of taking the mean value of the objects in a cluster as a reference point. Each remaining object is then clustered with the representative object to which it is the most similar. The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding

reference point. The absolute-error criterion is redefined as

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|^2, \tag{2.3}$$

where E is now the sum of the absolute error for all objects in the data set, p is the point in space representing a given object in cluster C_j, and o_j is the representative object of C_j. The algorithm iterates until each representative object is actually the medoid - the most centrally located object - of its cluster.

Because the medoid is less influenced by extreme values than a mean the method is more robust than k-means in the presence of noise and outliers. Notwithstanding, its processing is more costly [39], and note that both methods require the specification of the number of clusters k. In order to combat the shortfall of the scalability of the Partition Around Medoids (PAM) methodology with large data sets a sampling-based method called Clustering LARge Applications (CLARA) has been proposed, reducing the complexity of attribution of medoids significantly [40].

Aside from using the mean or the medoid, other alternative measures of cluster center are also commonly used in partitioning methods. An example is the k-modes [41] method, which by replacing the centroids with modes extends the k-means paradigm to cluster categorical data, uses new dissimilarity measures to deal with categorical objects, and a frequency-based method to update modes of clusters.

2.3.2 Hierarchical Methods

A hierarchy concept [42] can be formulated following two methods: agglomerative, also referred to as bottom-up approach, where each object starts from the bottom as a single cluster, being repeatedly grouped with neighboring clusters to form higher-level concept until all objects fall into a single group; and divisible, also referred to as a top-down approach, where as opposed to the previous, one single cluster containing all objects is decomposed into several subclusters until each group consists of only one object, forming a lower level hierarchy.

The graphical display of the entire process is called the dendrogram and is often viewed as a graphical summary of the data. Usually, we are looking for huge distance vertical jumps that indicate that something merged at this level should not have been merged, allowing us to deduce that the groups being combined probably do not belong to the same cluster. Hierarchical clustering has successfully been employed and able to identify clinically relevant tumor subtypes in several studies [43–45]. The agglomerative clustering approach is explicit in Algorithm 2.1.

The hierarchical clustering methods vary with respect to the (difficult) choice of the distance metric and cluster merging known as linkage, which can be application dependent. One commonly used "distance" between two clusters is the average linkage method - in this method the distance between two Clusters 1 and 2 is defined as the average of all distances between each member in Cluster 1 and each member in Cluster 2. Other alternatives are the single linkage (or nearest neighbor) and the complete linkage (or furthest neighbor). In single linkage method, the distance between Cluster 1 and Cluster 2 is the shortest distance from any element of Cluster 1 to any element of Cluster 2, while in contrast, the

Algorithm 2.1. Agglomerative Hierarchical Clustering.

1: Start with a collection C of n single clusters c_i containing one object x_i

2: Repeat until only one object is left:

(a) use a distance matrix $\mathbf{D}$ to find a pair of clusters that are closest to each other, $\min_{i,j} D(c_i, c_j)$

(b) merge cluster c_i with c_j to form a new cluster c_{i+j}

(c) remove c_i and c_j from the collection C and add c_{i+j} to C

complete linkage method defines the maximum distance from any element of Cluster 1 to any element of Cluster 2. One important note is that in the presence of outliers, using single linkage the unwanted values are accounted at last, while using complete linkage the outliers are considered at first. Thus, to avoid this sensitivity to outliers the average linkage is engaged and the resulting clusters are based on the average density.

Most of the approaches touched on so far in this section (with the notable exception of single-linkage hierarchical clustering) are biased towards clusters with convex, hyper-spherical shape. A detailed review of these clustering algorithms is provided in [46] and more recently in [47].

2.3.3 Other Relevant Clustering Methods

Remaining in the most classical approaches, model based clustering [48] is introduced as a formation of methods that assume that the data follow a mixture of underlying known probability distributions and pursuits to optimize the fit between some mathematical model and the data. Here the problems of selecting a good number of clusters and class-label individuals are realized as model selection problems in the probability framework [49]. Despite the fact that this approach has a statistically solid foundation for estimating and selecting a model the distributional assumptions are largely difficult to justify for high dimensional datasets [50]. Although some transformations can be enforced to mold the data into a known distribution, a legitimate transformation function is not easily identified. As a result, the methods may not optimize the likelihood converging to a local minima and end up having spurious clusters due to the lack of satisfying distributional assumptions; notwithstanding, interesting results have been achieved when dealing with clinical data [51–53].

Organizing and searching data often relies on the detection of areas where objects form groups with certain similarities but in high dimensional scenarios all objects appear to be sparse and dissimilar in many ways (the distance measurement between pairs of points becomes meaningless and the average density of points anywhere in the data is likely to be low) which prevents common data organization strategies from being efficient [54]. In data mining, moving forward with more specialized alternatives, efforts have been made on finding methods for effective cluster analysis in large databases, focusing on scalability for clustering complex types of data and shapes, and on high-dimensional clustering techniques. Two ideologies have been brought up closely trailing the hierarchy concepts presented in the previous

subsection. In dimension-growth subspace clustering, one takes advantage of the downward closure property of density to reduce the search space using an *a priori* style approach - if there are dense units in k dimensions there are also dense units in all $(k-1)$ dimensional projections. The nature of the bottom-up approach leads to overlapping clusters, where one instance can be in either none or more clusters, an advantageous condition in subspace clustering since the clusters often exist in different subspaces and thus symbolize disparate relationships. These methods determine locality by creating bins for each dimension and using those bins to form a multidimensional grid, obtaining meaningful results dependent on the proper tuning of the grid size and the density threshold parameters. These can be particularly difficult to set, especially since they are used across all of the dimensions in the dataset. A popular adaptation of this strategy provides data driven, adaptive grid generation to stabilize the results across a range of density thresholds by determining the cut-points for the bins on each dimension. An example is found in CLIQUE [55], a density and grid-based subspace clustering methodology that adopts a fixed grid size. Once the dense subspaces are found they are sorted by the fraction of the dataset covered by the dense units in the subspace. The subspaces with the greatest coverage are kept and the rest are pruned. The algorithm then finds adjacent dense grid units in each of the selected subspaces using a depth first search. Clusters are formed by combining these units using a greedy growth scheme - the algorithm starts with an arbitrary dense unit and greedily grows a maximal region in each dimension until the union of all the regions covers the entire cluster. Redundant regions are removed by a repeated procedure where the smallest redundant regions are discarded until no further maximal region can be removed [56]. Unlike many applications, CLIQUE can find any number of clusters in any number of dimensions, not depending on any additional information introduced by a parameter. Other insights can be grasped utilizing dimension-reduction subspace clustering that confluences multiple iterations of expensive algorithms in the full set of dimensions, and usually requires strategic implementations of some sampling technique to improve performance. Top-down algorithms create clusters that are partitions of the dataset, meaning each instance is assigned to only one cluster. The most critical parameters for top-down algorithms is the number of clusters and the size of the subspaces, which are often very difficult to determine ahead of time. Moreover, since the subspace size is a specified parameter, these algorithms tend to find clusters in the same or in similarly sized subspaces. For techniques that use sampling, the size of the sample is another critical parameter and can play a large role in the quality of the final results. PROCLUS [57] is an example of dimension-reduction subspace clustering, working in a manner similar to the already referenced k-medoids method. It works by sampling the data, selecting a set of k medoids and iteratively improving the clustering in a three phased approach consisting of initialization, iteration, and cluster refinement.

As with any clustering techniques, finding meaningful and useful results depends on the selection of appropriate techniques and tuning of the algorithm's parameters. In order to do this, there needs to be an understanding of the domain specific context to be able to best evaluate the results from various approaches. One must also understand the various advantages/disadvantages, and biases of the potential clustering algorithms.

2.4 Validation Indicators

Although not a primary point of development of this thesis, validation indicators are ever present in assuring, or attempting to assure the ideal clustering performed in a given dataset. Historically, a host of performance metrics have been brought up for the evaluation of clustering results based on a singular realization of a point-set process [58, 59]. Indisputably, the accuracy measurement of a cluster operator is not feasible when based on a singular application. It would not be logical to think that a classifier could be validated through the analysis of a single point without knowledge of the true classification of that point. Analogously, a cluster operator should not be evaluated facing a nonexistent ground-truth partition. The spatial structure of the grouping outputted by the cluster operator can be organized based on its various aspects, for example compactness.

One can loosely divide validity measures into three categories. The first is supported on the calculation of properties of the resulting clusters, for instance, separation, compactness, and roundness. The approach is named internal due to the absence of additional information regarding the data. The second is supported on comparisons of partitions originated from runs of the same algorithm minus a change in subsets of the data or in the input parameters. This one is named relative and shares the lack of additional information present in the first approach. The final one, denominated external, is some sort direct or indirect error measurement.

In this section, several validation indices will be covered, and their motivations and advantages/disadvantages discussed. A more thorough review can be retrieved reading the work of Halkidi [60].

2.4.1 Internal Validation Indicators

The simplest way to assess a clustering algorithm is by guidance of internal indicators as the focus here is solely on the spatial distribution of the points and on the computed cluster labels. This family of methodologies is based on the assumption that the algorithms should look for groups far from members of other clusters and whose members are close to each other.

Dunn's Indices Following a simple concept - the further apart are the groups, relative to their size, the larger the index and the better the clustering - this ratio addresses the disparities of the minimum distance between two clusters and the size of the largest cluster [61], i.e.

$$V(C) = \frac{\min\limits_{i \neq j} d_C(C_i, C_j)}{\max\limits_{j} \Delta(C_j)},$$

where d_C is the distance between two clusters and Δ the cluster sizing function. Due to the high number of possible options for the computation of the numerator and the denominator the ability to mix and match distance and cluster-sizing measures can be staggering [62].

Silhouette Index The silhouette of a cluster is defined as the average silhouette width of its points, where the width of a certain point defines its proximity to its own cluster relative to its proximity to

other clusters, that is

$$S(x) = \frac{min\_avg(x) - avg(x)}{\max\left(min\_avg(x), avg(x)\right)},$$

where x is the data point, avg the function that computes the average distance between x and all other points in its own cluster, and min_avg the minimum of the average distances between x and the points in the other clusters. For each x, the silhouette width ranges from -1 to 1: if the value is closer to -1, the point is on average closer to another cluster than the one to which it belongs; and the contrary also holds true, meaning the more compact and separated the clusters the higher the index will be.

2.4.2 Relative Validation Indicators

Neither are internal indices sensible to the stability of the algorithm against variations in the data, nor do they sense the consistency of the results in the case of redundancy. A family of more complex indices arises and attempts to exploit the repetition of information assumed to be prevalent in the data available.

Figure of Merit The figure of merit (FOM) of a variable is calculated by grouping the samples after removing the given feature and measuring the average distance between all samples and their cluster's prototypes concretely on the inspected feature. If this average distance is small, the algorithm is considered consistent because it partitioned the samples in compact clusters even with the feature removed. Speaking in terms of the heuristic, a clustering method that produces consistent clusters should be able to predict removed features, and thus have a low FOM index. However, the FOM performance based on simulated data is usually below that of simpler validation indices, including many internal ones. Do note however that the necessity of repeating clustering many times, in conjunction with the poorer performances against simpler techniques in simulated data, makes this measure one of the last choices for selecting a validation index [63].

Stability The ability of a grouped data set to predict the clustering of another data set sampled from the same source is measured by the instability index [64]. It functions by dividing the set of points to be categorized in two parts: initially, on the first collection of points, the clustering algorithm under scrutiny is applied and obtained labels over these points are used to train a classifier that segregates the entire space. Both the original clustering algorithm and this new classifier are applied to the second part, generating two distinct sets of labels. The disagreement between these labels, averaged over repeated random partitions of the points, defines the instability of the clustering algorithm.

The dependence on the number of clusters dictates the need for the instability index to be normalized when used for model selection. Other preoccupations occur in the classification rule selection, which can strongly influence the output [64]. Simulation inquiries show underperformance against other internal and relative indicators, despite being one of the most time consuming ones since it involves the reoccurring application of a clustering algorithm and training a classifier [65].

2.4.3 External Validation Indicators

The last family compares properties of an algorithm's fabricated clusters against that of known ground-truth clusters.

Hubert's Correlation This statistic computes, for the achieved partition after applying a clustering algorithm and the expected partition, the correlation between the respective co-occurrence matrices. Not relying on label permutations since co-occurrence matrices do not depend on the classifications used to define the groupings, the index exploits the notion that similar partitions have similar co-occurrence matrices, and consequently high correlation [66].

(Adjusted) Rand Statistics and Jaccard coefficient Working with the same co-occurrence matrices of the previous statistic, the measure of disagreement constrained to $[0, 1]$ between the clusters is quantified by the number of pairs of points i and j that reside in each of four categories:

1. i and j fall in the same cluster in the computed and expected cluster;
2. i and j fall in the same cluster in the computed partition, but in different clusters in the expected partition;
3. i and j fall in different clusters in the computed partition, but in the same cluster in the expected partition;
4. i and j fall in different clusters in the computed and expected cluster.

The Rand statistic measures the proportion of pairs of vectors that agree by belonging either to the same cluster (1.) or to different clusters (4.) in both partitions and assumes the form of

$$R = \frac{(1.) + (4.)}{M},$$

where $M = n \cdot \frac{n-1}{2}$ is the number of pairs composed of different points. The bias present in this indicator is shown as two random partitions do not produce a constant R, problem corrected by the adjusted Rand index (ARI) that assumes a generalized hyper-geometric distribution as the model of randomness. The ARI has the maximum value of 1, and its expected value is 0 in the case of random clusters - a larger ARI means a higher agreement between two partitions. The ARI is recommended for measuring agreement even when the partitions compared have different numbers of clusters. The Jaccard coefficient touches on the same metrics by measuring instead the proportion of pairs that belong to the same cluster in both partitions, relative to all the pairs belonging to the same cluster in at least one of the two partitions, and is given by

$$J = \frac{(1.)}{(1.) + (2.) + (3.)}.$$

3

Beyond Tandem Analysis

Contents

This chapter addresses the problem of dimensionality reduction and its effect on data structure masking. Strategies to overcome this difficulty are presented and analyzed with both synthetic and real data.

When it is thought that some of the features studied do not contribute much to identify the clustering structure, or when the number of features is large, researchers promote the application of discrete and continuous models in a sequential manner to detect non-observable dimensions that summarize the information available in the data set. This operation frequently consists in performing PCA before applying a clustering algorithm on the scores of the objects on the first components [67]. This kind of approach was firstly named "tandem analysis" by Arabie and Hubert [68] and was already disputed by De Sarbo et al. [69] because the dimensions identified by feature extraction/selection techniques may not necessarily help us to understand the group structure of the data, possibly obscuring and masking the taxonomic information in the process.

With a mind set aiming at packing the vital information available, De Soete [70] proposed Reduced K-Means (RKM), a procedure that represents each cluster by its centroid in a low-dimension space. In a second step, the low-dimensional representation of the clusters is obtained after projecting the objects to the centroid space. In this methodology, a subspace of the full data space is found such that the actual data points present in the full space have the smallest sum of squared distances to the centroids lying in a subspace. The method may fail to find relevant clusters residing in a subspace when the data has too much variance in the orthogonal direction to the one enclosing the relevant cluster, as this variance may significantly impact the computation of the sum of squared distances between the data points and the prototypes. Focusing on reaching a subspace representation of the data, a method called Factorial K-Means [67], attempting to find the subspace where the projected observations have the smallest sum of squared distances to the central points of each cluster in the same subspace, is brought up. Other processes combining cluster analysis and the search for a new dimensional representation exist, based, for example, on unfold analysis or multidimensional scaling applications.

The tandem approach is conjectured to perform worse when variables are added that are unrelated to the cluster structure, and the study in this chapter confirms this conjecture. A theoretical comparison between FKM and RKM is developed and it is confirmed that they complement each other: for RKM, the subspace recovery is undermined with increasing relative sizes of complement residuals compared to the subspace residuals; while the reverse holds true for FKM. As both methods can suffer from subspace and membership recovery problems, it is essential to take into account the basis of the content and context of a given problem and critically evaluate each of the methods' solutions.

3.1 Tandem Analysis

An example where the inclusion of irrelevant features potentiates the masking of the groups' structure is showcased because it may help to clarify some of the problems brought to light here. In Fig. 3.1, and similarly to [71], 42 objects are plotted on two variables spread out as well-defined irregular hexagonal structures symbolizing three different classes. In this example, the objects were also described by other four noise random variables generated by a normal distribution with 0 mean and variance 6. The new 42×6 matrix is partitioned using the k-means clustering algorithm and the results who identify the group membership of all observations are illustrated by their classification from 1 to 3, in Fig. 3.1.

In the figure, it is visible that the two-dimensional data set has been masked, if not completely hidden, by the six-dimensional data set. Namely, we can affirm the missclassification of 26 of the 42 objects. The use of a variable reduction technique to extract the most relevant information may be deliberated to reduce the masking effect in this scenario, but this line of thought is not viable as can be seen in Figs. 3.2a, 3.2b, 3.2c, and 3.2d, where PCA is applied on the six-dimensional space and k-means is sequentially performed on the increasing dimensions of the components' scores. Note that in the last three of these computations, in the scores of the three, four and five principal components resulting from the k-means algorithm, the location of the objects in Subfigures 3.2b, 3.2c, and 3.2d is projected on the two original variables that define the three well-separated classes, while the class membership in these subfigures correspond to the results of the respective tandem analyses. The percentage of total variance explained by the different solutions is showcased in Table 3.1. In Fig. 3.2a, the first two components are not the original ones that represent the well-defined and separable three classes, which coincides in objects not being close to those of the same class and being misslabeled in return. Once we increase the number of principal components to be considered when applying the clustering algorithm the situation becomes slightly better, and the number of missclassifications brought down from 26 to 10, in Fig. 3.2d.

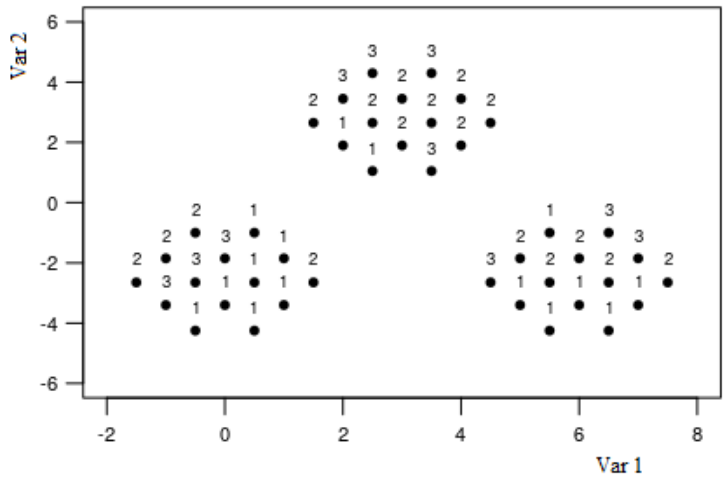
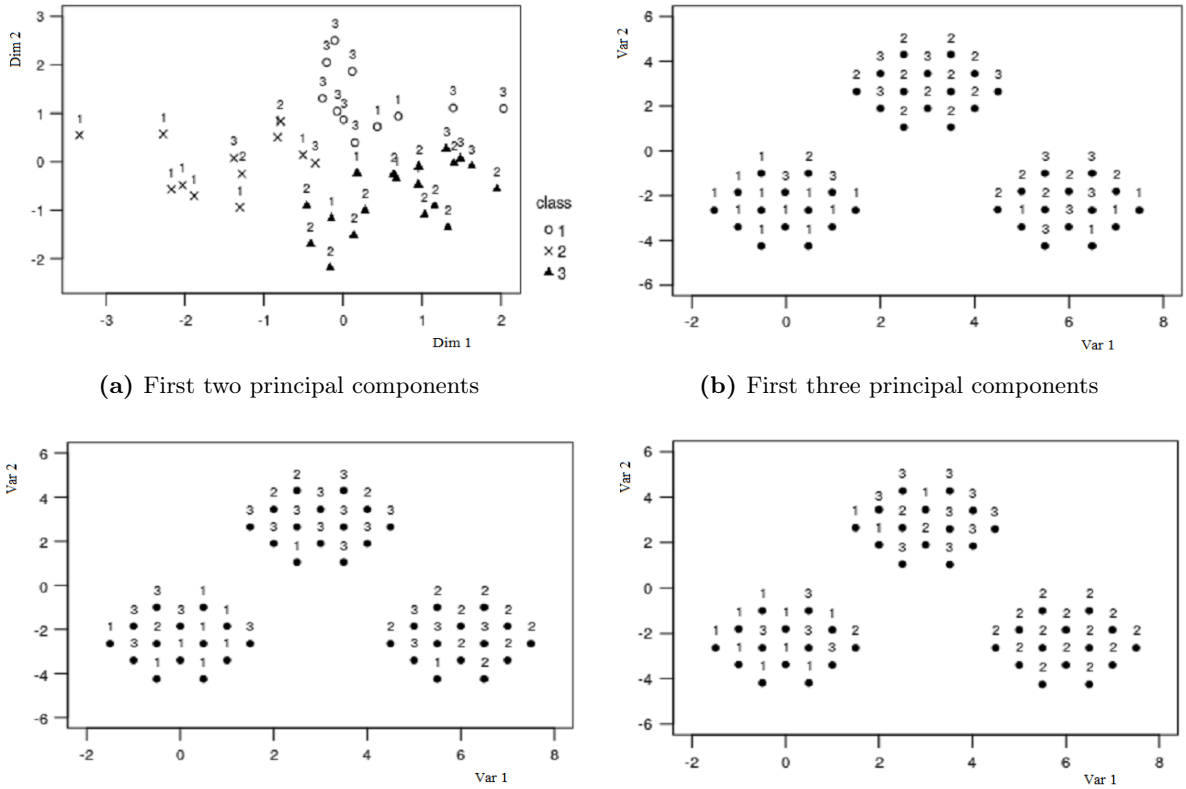


Figure 3.1: K-means classification of 42 objects described by six variables, two of which determine the location of the points in the plot (three classes) and the other four variables are randomly generated by normal distribution.

The tandem analysis carried out in this section is the most popular, and usually performed. To



(a) First two principal components

(b) First three principal components

(c) First four principal components

(d) First five principal components

Figure 3.2: Tandem analysis. K-means clustering computed on: (a) the first two components' scores; (b) three components' scores; (c) four components' scores; (d) five components' scores. Note: Classifications (b), (c) and (d) are represented on the two variables that defined the three well-separated classes in Fig. 3.1.

avoid the case of distorted mutual distances between the objects one should substitute the standardized scores by those multiplied by the square roots of the eigenvalues, providing a much better conservation of the distances. In our example this action did not yield any improvements and proved to be inefficient, having produced highly similar outcomes to what was previously obtained in the case study using 2 to 5 components. We make the conclusion that the poor performance of the tandem approach is not sourced in the bad space representation, but instead it originates from the simple fact that the principal components are not upbringing the best achievable agglomerations.

Table 3.1: Explained total variance and cumulative variance.

Components	Eigenvalue	% variance	% cumulative
1	1.23	25.33	25.33
2	1.03	17.52	42.85
3	1.00	16.78	59.63
4	0.98	16.15	75.78
5	0.89	13.21	88.99
6	0.81	11.01	100.00

3.2 Two alternatives: Factorial and Reduced K-means

Before advancing with any comparison to a sequential tandem approach, two similar to each other but easily distinguishable integrated methods are introduced and analyzed.

3.2.1 Optimization of the Loss Functions

Both integrated approaches aim at identifying the best partition of objects described by the best orthogonal linear combination of variables following the least squares criterion. A dual objective is attempted to be achieved: optimal data synthesis of objects and attributes occur simultaneous to variable selection in cluster analysis, where the features that contribute the most to select a label to all data points are pinpointed. The methods deconstruct the data matrix $\mathbf{X}$ into an object membership assignment matrix $\mathbf{U}$, an orthonormal components' score matrix $\mathbf{A}$ that expresses the loadings of the variables, and a cluster centroid score matrix $\bar{\mathbf{Y}}$. An artificial example of a possible layout of the matrices, with $I = 3$, $J = 6$, $P = 2$ and $Q = 2$, follows:

$$\mathbf{U} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0.4 & 0.0 \\ 0.3 & 0.0 \\ 0.0 & 0.7 \\ 0.0 & 0.8 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}, \qquad \bar{\mathbf{Y}} = \begin{bmatrix} 2.4 & 2.9 \\ -0.6 & -5.1 \end{bmatrix}.$$

Matrix $\mathbf{U}$ indicates that Object 1 belongs to group one, and the remaining objects to the other group. Loading matrix $\mathbf{A}$ shows Variables 1 and 2 associated with the first subset of attributes, Variables 3 and 4 associated with a second dimension, and the remaining not belonging to any subset, and thus irrelevant to information representation (masking features that do not help represent the structure of the data). The centroid matrix contains the centroids location of the two clusters.

In respect to clustering, these initiatives are categorically identified as selection and weighting approaches, being the fundamental difference to other approaches in the same category that in RKM and FKM the selection, weighting, and clustering are done simultaneously, instead of being two very distinct phases of the process. The distinction between these modified k-means methods lies on the objective functions considered by their models. The RKM loss function to minimize is written as

$$F_{RKM}(\mathbf{U}, \mathbf{A}, \bar{\mathbf{Y}}) = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}||^2, \tag{3.1}$$

whereas the FKM minimizes the loss function

$$F_{FKM}(\mathbf{U}, \mathbf{A}, \bar{\mathbf{Y}}) = ||\mathbf{X}\mathbf{A}\mathbf{A}^{\mathrm{T}} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}||^2 = ||\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{Y}}||^2. \tag{3.2}$$

When the prototypes are located in the full space, i.e. when the number of variable clusters are equal to the number of variables in the data, the methods trace back to the original k-means algorithm. The mathematical representation introduced provides interesting properties to the methods: the clusters have permutational indeterminacy - the value of the loss function can remain intact by replacing $\mathbf{U}$ by $\mathbf{UR}$, with a permutation matrix $\mathbf{R}$, provided that the permutation is compensated for in the centroid matrix as $\mathbf{R}^{\mathrm{T}}\bar{\mathbf{Y}}$ - and both the components' score and centroid matrix have rotational indeterminacy - the value of the loss function can remain intact by replacing $\mathbf{A}$ by $\mathbf{A}\boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a orthonormal rotation matrix, provided that the rotation is compensated for in the centroid matrix as $\bar{\mathbf{Y}}\boldsymbol{\Psi}$.

As is observable from Eq. (3.1), RKM minimizes the sum of the squared distances between the observed and the centroids located in a subspace of the data which is spanned by the columns of $\mathbf{A}$. From (3.2) it is taken that FKM minimizes instead the within-clusters deviance in the reduced space, i.e. the sum of the squared distances between the centroids in the projected space and the observed data points that are projected onto the subspace in which the centers of the clusters reside.

3.2.2 Model Analysis

Let's now analyze the model fit for each case. The RKM model fitted by (3.1) is

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}_R, \tag{3.3}$$

where $\mathbf{E}_R$ is an $(I \times J)$ residual matrix. Given the fact that the optimal $\bar{\mathbf{Y}}$ is equal to $(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{XA}$ then (3.3) is rewritten as

$$\mathbf{X} = \mathbf{P}_{\mathrm{U}}\mathbf{XAA}^{\mathrm{T}} + \mathbf{E}_R,$$

where $\mathbf{P}_{\mathrm{U}} = \mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}$ is a projection matrix on the space spanned by the columns of $\mathbf{U}$ that denotes the weights of the objects to compute the cluster center. On the other hand, the FKM fitted by (3.2) is

$$\mathbf{XAA}^{\mathrm{T}} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}_F, \tag{3.4}$$

where $\mathbf{E}_F$ is an $(I \times J)$ residual matrix. Taking the optimal $\bar{\mathbf{Y}}$ we can once again rewrite a model as

$$\mathbf{XAA}^{\mathrm{T}} = \mathbf{P}_{\mathrm{U}}\mathbf{XAA}^{\mathrm{T}} + \mathbf{E}_F.$$

The equations reinforce the data reconstruction process achieved by both methods: RKM only has the centroids lying in the reduced space for the reconstruction, while FKM assumes that the objects also lie in the lower-dimensional space.

Note that the residuals of the FKM model can be expressed as $\mathbf{E}_F = \mathbf{XAA}^{\mathrm{T}} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}$, where the matrix $\mathbf{E}_F$ belongs in the row space of $\mathbf{A}^{\mathrm{T}}$: Understanding that $\mathbf{E}$ is in the same column space as $\mathbf{E}_F$ permits for $\mathbf{E}_F = \mathbf{EA}^{\mathrm{T}}$ to hold. This way, when the full space of $\mathbf{X}$ is represented, $\mathbf{E}_F$ includes the

subspace residuals of the projected data. With

$$\mathbf{XAA}^{\mathrm{T}} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}_F = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{EA}^{\mathrm{T}} = (\mathbf{U}\bar{\mathbf{Y}} + \mathbf{E})\mathbf{A}^{\mathrm{T}},$$

and recalling the rotation property of the projected configuration of objects, Equation (3.4) is rewritten as $\mathbf{XA} = \mathbf{U}\bar{\mathbf{Y}} + \mathbf{E}$. The FKM model as it stands models the orthogonal projection onto a subspace of the data. Another residual term describing the residuals in the orthocomplement subspace of $\mathbf{A}$ is added to model $\mathbf{X}$ instead of $\mathbf{XAA}^{\mathrm{T}}$; the so called complement residuals, $\mathbf{X} - \mathbf{XAA}^{\mathrm{T}}$, are denoted by $\mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}$, where $\mathbf{A}^{\perp}$ is a columnwise orthonormal matrix constrained by $\mathbf{A}^{\mathrm{T}}\mathbf{A}^{\perp} = 0$. The full model of $\mathbf{X}$ is formulated and serves as the backbone of the discussion over the performances both approaches,

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{EA}^{\mathrm{T}} + \mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}. \tag{3.5}$$

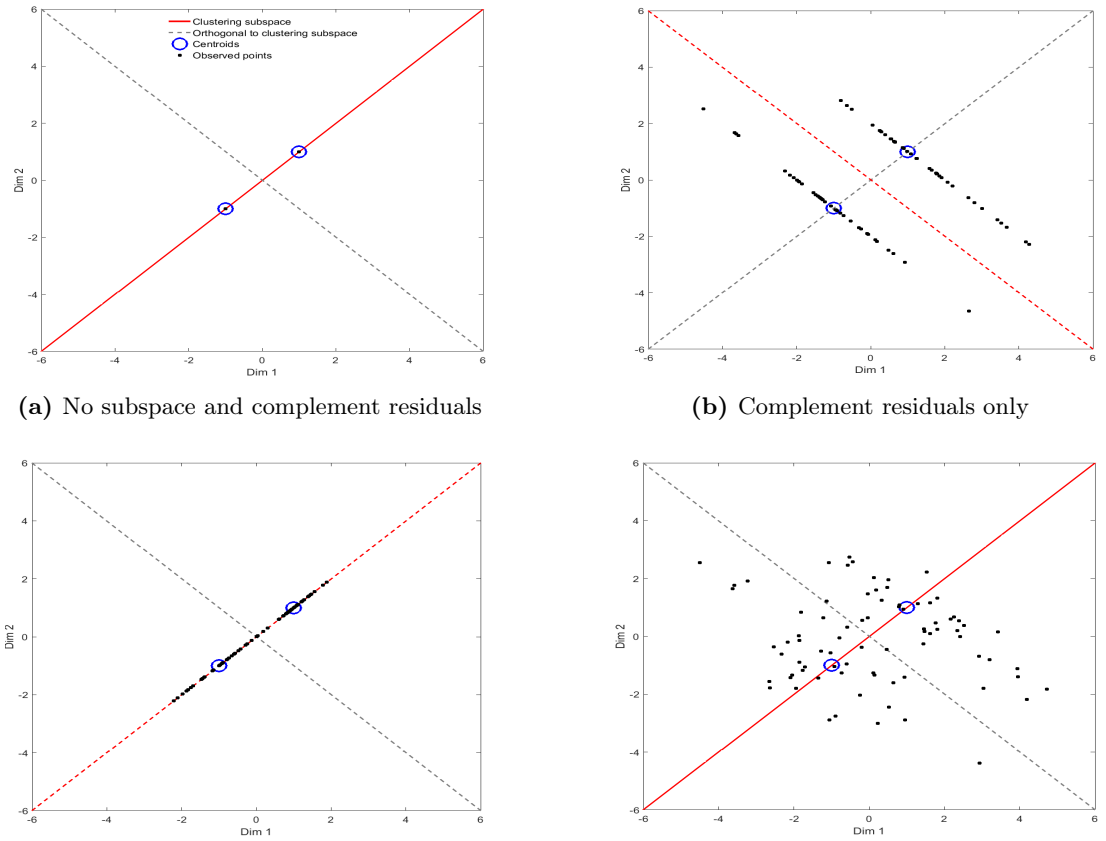
Model (3.5) is broken down into two types of residuals and a structural part. The residuals are distinguished as subspace residuals $\mathbf{EA}^{\mathrm{T}}$ located within the subspace where the prototypes (RKM) or the prototypes and data points (FKM) lie, and as complement residuals $\mathbf{E}^{\perp}\mathbf{A}^{\perp}$ located within the complement of the subspace. Looking at (3.5) one can idealize four types of data composed of different configurations: data that contain both subspace and complement residuals ($\mathbf{E} \neq 0$ and $\mathbf{E}^{\perp} \neq 0$), data that holds subspace residuals only ($\mathbf{E} = 0$), data that holds subspace residuals only ($\mathbf{E}^{\perp} = 0$), and lastly data that contain neither subspace nor complement residuals ($\mathbf{E} = 0$ and $\mathbf{E}^{\perp} = 0$). In Fig. 3.3, fixing $J = 2$, $C = 2$, and $Q = 1$, examples of the idealized types are showcased.

To retrieve further insight into the masking of variables it is worth it to analyze the data deconstruction. Because these concealed features are not affiliated in any way with the clustering structure the assumption of them belonging entirely to the residual portion of (3.5) is made. As they do not contain to any extent the clustering structure in the object assignment matrix, it means that they do not have any weights in the underlying variable attribution matrix. Consequently, the nature of masking variables must be fully present in the complement residual.

3.2.3 Finding the Ideal Data

From (3.2) it follows that the FKM's loss function is null if and only if $\mathbf{XA} = \mathbf{U}\bar{\mathbf{Y}}$. Considering the $(I \times J)$ data matrix $\mathbf{X}$ can be expressed in terms of the $(J \times Q)$ loadings matrix $\mathbf{A}$ and $\mathbf{A}^{\perp}$, with $\mathbf{B}$ a $(I \times Q)$ matrix and $\mathbf{C}$ a $(I \times (J - Q))$ matrix, as

$$\mathbf{X} = \mathbf{BA}^{\mathrm{T}} + \mathbf{CA}^{\perp\mathrm{T}}, \tag{3.6}$$



(a) No subspace and complement residuals

(b) Complement residuals only

(c) Subspace residuals only

(d) Both subspace and complement residuals

Figure 3.3: Examples of data configurations. (a) Neither subspace nor complement residuals; (b) Complement residuals only; (c) Subspace residuals only; (d) Both subspace and complement residuals. Note the choice of subspace (one dimension subspace represented by a diagonal line from bottom left to top right).

replacing (3.6) in $\mathbf{XA} = \mathbf{U}\bar{\mathbf{Y}}$ we find $\mathbf{XA} = \mathbf{BA}^{\mathrm{T}}\mathbf{A} + \mathbf{CA}^{\perp\mathrm{T}}\mathbf{A} = \mathbf{B} = \mathbf{U}\bar{\mathbf{Y}}$. Returning to the ideal FKM data we have $\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{CA}^{\perp\mathrm{T}}$, and stating the general matrix $\mathbf{C}$ by $\mathbf{E}^{\perp}$, the full class is

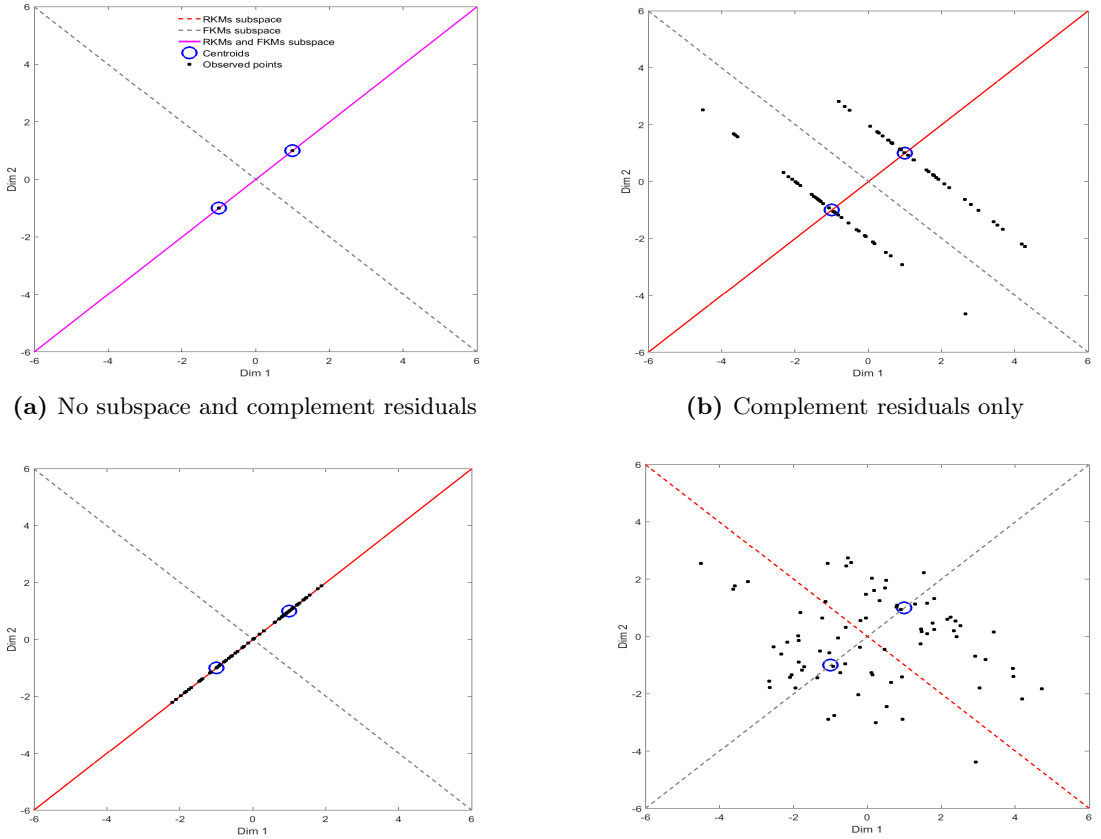
$$\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}. \tag{3.7}$$

Equation (3.4) reiterates the characteristics of the ideal FKM data as being the one with null subspace residuals, and having no particular restriction on the complement residuals, Fig. 3.3a and 3.3b.

From (3.1) it follows that the RKM's loss function is null if and only if $\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}$. Equation (3.3) illustrates the ideal RKM data, showcased in Fig. 3.3a, as being the one with null subspace and complement residuals, thus encompassing the domain of the ideal FKM data (the reverse does not hold true, ideal FKM data does not guarantee ideal RKM data).

3.3 Performance Comparison

The varying success in clustering structure recovery within the different layouts of the data displayed in Fig. 3.3 is examined and possible solutions are displayed in Fig. 3.4.



(a) No subspace and complement residuals

(b) Complement residuals only

(c) Subspace residuals only

(d) Both subspace and complement residuals

Figure 3.4: Examples of possible FKM and RKM solutions for the four idealize types of data. (a) Neither subspace nor complement residuals; (b) Complement residuals only; (c) Subspace residuals only; (d) Both subspace and complement residuals. The lines in the subfigures indicate the subspace(s) where RKM and/or FKM may end up.

3.3.1 Absence of Subspace and Complement Residuals

We start by considering data without subspace and complement residuals, an instance that allows RKM's and FKM's model to reach a perfect compliance with the available information.

Contrarily to RKM that yields a perfect solution, FKM may arbitrarily determine solutions with the nullification of the loss function irrespective of the projected subspace having the clustering structure or not. Assume, for example, that the complement space has more dimensions than the subspace with the wanted structure to be identified. A projection on an orthogonal subspace to the one of interest implies that the loading matrix is estimated to be $\hat{\mathbf{A}} = \mathbf{A}^{\perp}\mathbf{R}$, with $\mathbf{R}$ an arbitrary orthonormal matrix. Consequently, the coordinates of the lower-dimensional data point representation are equal to $\mathbf{X}\hat{\mathbf{A}} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\mathbf{A}^{\perp}\mathbf{R} = \mathbf{U}\bar{\mathbf{Y}}0 = 0$, the loss function being zero with an arbitrary $\hat{\mathbf{U}}$ and $\bar{\mathbf{Y}} = 0$. The objective function is obviously also zero if $\hat{\mathbf{A}}$ is taken within the desired subspace itself, as $\mathbf{AR}$. Whether or not the estimated feature attribution matrix is in the column space of $\mathbf{A}$ and/or in its orthogonal complement $\mathbf{A}^{\perp}$ the minimum value of F_{FKM} is zero. The possible arbitrary response provided by FKM is easily detectable as the total variance of $\mathbf{X}\hat{\mathbf{A}}$ is zero when the solution resides simply in the orthocomplement subspace. Figure 3.4a provides a visual cue on the possible two FKM solutions, as well as the single RKM solution.

3.3.2 Complement Residuals Only

In the second case, assuming that the correct numbers of clusters of objects and variables are being estimated and that the complement residuals have full rank $J-Q$, it can be proven that FKM estimates $\hat{\mathbf{A}} = \mathbf{A}^{\perp}\mathbf{R}$, $\hat{\mathbf{U}} = \mathbf{U}\boldsymbol{\Psi}$ and $\hat{\bar{\mathbf{Y}}} = \boldsymbol{\Psi}^{\mathrm{T}}\bar{\mathbf{Y}}\mathbf{R}$, with $\boldsymbol{\Psi}$ some permutation matrix (for more precise clarifications see [72]). In this situation the RKM objective function is

$$\begin{aligned} F(\mathbf{U}, \bar{\mathbf{Y}}, \mathbf{A}) &= ||(\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}) - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}\hat{\mathbf{A}}^{\mathrm{T}}||^2 \\ &= ||\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}\hat{\mathbf{A}}^{\mathrm{T}}||^2 + ||\mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}\hat{\mathbf{A}}^{\mathrm{T}}||^2 \\ &= ||\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}\hat{\mathbf{A}}^{\mathrm{T}}||^2 + ||\mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}||^2 - 2\mathrm{tr}(\hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{A}^{\perp}\mathbf{E}^{\perp\mathrm{T}}) + 2\mathrm{tr}(\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\mathbf{A}^{\perp}\mathbf{E}^{\perp\mathrm{T}}) \\ &= ||\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}\hat{\mathbf{A}}^{\mathrm{T}}||^2 + ||\mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}||^2 - 2\mathrm{tr}(\hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{A}^{\perp}\mathbf{E}^{\perp\mathrm{T}}). \end{aligned}$$

Even with a perfect estimation of $\mathbf{U}, \bar{\mathbf{Y}}$ and $\mathbf{A}$ the term $||\mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}||^2$ is not absent of value and RKM may end up with a response that may partly occupy the $\mathbf{A}^{\perp}$ subspace if this solution has a lower F, Fig. 3.4b.

3.3.3 Subspace Residuals Only

In the third case the observed data follow $\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}\mathbf{A}^{\mathrm{T}}$, and the RKM loss function value is guaranteed to be larger than zero. Due to the fact that any interception of the solution with $\mathbf{A}^{\perp}$ adds on to F the subspace solution will always be the correct one, despite the invariable expected degradation of the structure's subspace recovery with an increase of subspace residual variances.

With this configuration, when FKM estimates $\hat{\mathbf{A}} = \mathbf{A}^{\perp}\mathbf{R}$, then $\mathbf{X}\mathbf{A}^{\perp}\mathbf{R} = (\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}\mathbf{A}^{\mathrm{T}})\mathbf{A}^{\mathrm{T}}\mathbf{R} = 0$, i.e. the projected data are zero, with $\bar{\mathbf{Y}} = 0$ and $\mathbf{U}$ is arbitrarily estimated. When the estimated loading matrix $\hat{\mathbf{A}}$ partly or fully resides in the subspace spanned by $\mathbf{A}$ the loss value is larger than zero and, therefore, solutions with $\hat{\mathbf{A}} = \mathbf{A}^{\perp}\mathbf{R}$ are favoured. The performances of both methodologies are illustrated in Fig. 3.4c.

3.3.4 Subspace and Complement Residuals

In the last case, the observed data follow $\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}\mathbf{A}^{\mathrm{T}} + \mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}$, and the presence of both subspace and complement residuals makes it difficult to intuitively proclaim any statement with respect to the RKM's and FKM's solutions. Based on the behaviors with only one of the residual components activated, the conjecture that, in the face of significant subspace residuals when compared to complement ones, RKM will perform better is made. The reverse holds true in the case of only complement residuals. In the work of Timmerman et al. [72] the conjecture is confirmed via an extensive simulation study; RKM and FKM are dictated to complement each other, although RKM is set up as having a wider applicability, especially when the majority of the variables reflect the clustering structure.

3.4 Integrated Approach Application

Following the discussion, performance is assessed by applying the methodology to a first simulated data already analyzed with tandem analysis, and a second data set analyzing data of psychiatric patients.

3.4.1 Simulation Study

Because we know that in the present situation the agglomerations lie in a subspace of the full data space, RKM and FKM are applied to the data, specified the centroid to be located in a two-dimensional subspace. The result exposed in Fig. 3.5 reveals structure recovery and data classification failure of RKM.

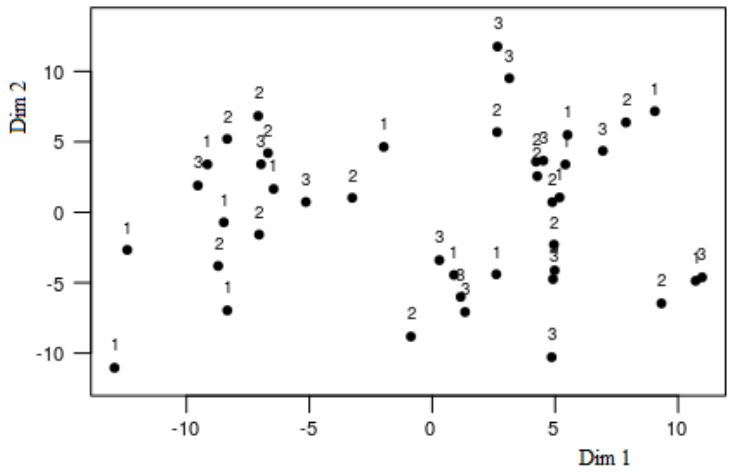


Figure 3.5: Classification of 42 objects in a low-dimensional space represented by the first two dimensions of the reduced k-means analysis.

Turning to the FKM procedure, Table 3.2 introduces the correlations between the model defined factors and the six original features. Figure 3.6 represents the 42 objects laid on the first two factors discovered and illustrates the impact of the random noise generated variables is drastically attenuated to the point of the well-separated structure making its appearance again.

Table 3.2: Correlation between the first two dimensions of the factorial k-means analysis and the six variables.

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
Dim 1	**0.9987**	0.0432	-0.0235	0.0085	0.0101	0.0034
Dim 2	0.0440	**-0.9958**	0.0253	-0.0645	0.0112	-0.0381

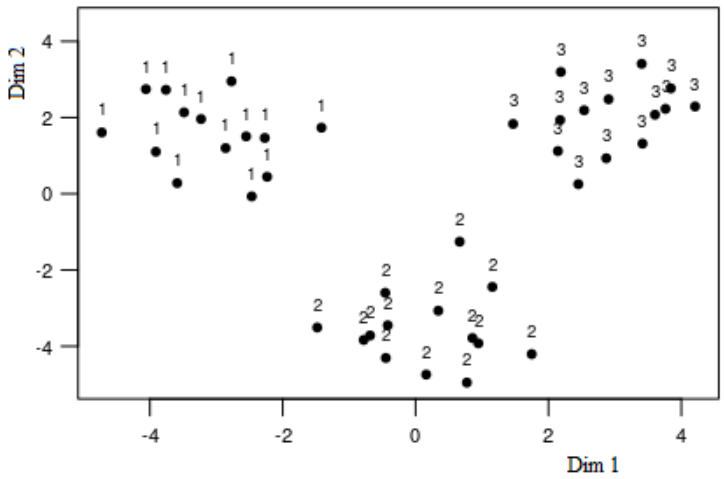


Figure 3.6: Classification of 42 objects in a low-dimensional space represented on the first two dimensions of the factorial k-means analysis.

3.4.2 Archetypal Psychiatric Patient Data

In the study of Mezzich and Solomon in [73], eleven psychiatrists were asked to describe the psychiatric patients' type according to four diagnostic categories associated with the DSM-II nomenclature of mental disorders of the American Psychiatric Association [74]. These categories are paranoid schizophrenic (PS), manic-depressive manic (MDM), manic-depressive depressed (MDD), and simple schizophrenic (SS). The patients were rated on the basis of 17 symptoms on a seven point scale from the Brief Psychiatric Rating Scale (BPRS): 7 meaning extremely severe and the opposite extreme, 1, meaning not present. These clinical annotations result in a 44×17 data matrix.

The data matrix was inspected with RKM and FKM over 500 random starts with four clusters of objects and a flexible number of features subsets ranging from 1 to 17. The evaluation metric chosen for the quality of recovery of the patients' types was the ARI between the expected and the estimated memberships. For each number of components the variance of the observed data var($\mathbf{X}$) was calculated, as well as variance of the projected data, var($\mathbf{XA}$). Several questions arise from the work done: Can the archetypal patients be obtained from the data? Can the symptoms be distilled to fewer relevant dimensions? And which features are important in the definition of the diagnostic characterizations under study?

Selection of Solutions

In Table 3.3 the relation between the increasing number of variables subsets and the non-decreasing fit of the RKM model is immediately clear. A series of analyses focused on predicting the optimal number of components Q is available. The model optimizing the balance between complexity/fit is chosen supported by the ARI values, exposing the fact that the addition of more components would not improve the clustering (a solution with 3 components was considered optimal).

Because the FKM model fit of the lower-dimensional data representation more often than not decreases with increasing numbers of components - as $\mathbf{XA}$ varies with differing Q's - the conclusions on FKM taken

over the var($\mathbf{X}$) values are rather unreliable as a low variance value has dubious interpretations: it is either indicative of high success in neglecting masking features, or indicative of failure in modelling the agglomeration structure.

The Adjusted Rand Index values of most of the FKM solutions bring to light the poor retrieval of the clustering structure. In fact, only the use of all or almost all components ($Q = 16$) appear to be at a satisfactory level of recovery. As these solutions are more complex, and a clear approach to the selection of the number of components in FKM analysis is lacking, in the remaining data set study only the optimal RKM solution is used.

Table 3.3: Values of var($\mathbf{X}$), var($\mathbf{XA}$) and Adjusted Rand Indices for each FKM and RKM solution.

Q	RKM			FKM		
	var($\mathbf{X}$)	var($\mathbf{XA}$)	ARI	var($\mathbf{X}$)	var($\mathbf{XA}$)	ARI
1	0.41	0.97	0.45	0.00	0.98	-0.01
2	0.53	0.87	0.69	0.01	0.92	0.10
3	0.64	0.85	0.88	0.02	0.82	0.12
4	0.64	0.79	0.88	0.02	0.75	-0.01
5	0.64	0.83	0.88	0.03	0.60	-0.01
6	0.64	0.80	0.88	0.03	0.55	0.01
7	0.64	0.78	0.88	0.02	0.60	0.17
8	0.64	0.78	0.88	0.03	0.50	0.10
9	0.64	0.75	0.88	0.03	0.46	0.09
10	0.64	0.73	0.88	0.05	0.50	0.26
11	0.64	0.71	0.88	0.08	0.55	0.37
12	0.64	0.69	0.88	0.06	0.41	0.15
13	0.64	0.68	0.88	0.10	0.47	0.12
14	0.64	0.66	0.88	0.42	0.74	0.37
15	0.64	0.65	0.88	0.46	0.72	0.51
16	0.64	0.65	0.88	0.55	0.69	0.78
17	0.64	0.64	0.88	0.64	0.64	0.88

Interpretation of the Model

In an attempt to facilitate interpretation and to reach the most simplified version of the solution, the clusters were characterized by the linear combination of the fewer components possible. In Table 3.4 one can easily relate Cluster 1 to Component 2, Cluster 2 to Component 1, Cluster 3 to an almost similarly weighted contribution of both Components 1 and 2, and Cluster 4 to Component 3.

Table 3.4: Rotated centroid scores of the clusters ($C = 4$) on the components ($Q = 3$) of the RKM solution; MDD is manic-depressive depressed,MDM is manic-depressive manic, SS is simple schizophrenic and PS is paranoid schizophrenic.

Centroids	Components		
	I (MDM)	II (PS)	III (SS)
1 (PS)	-0.04	**0.34**	0.09
2 (MDM)	**0.45**	-0.04	0.12
3 (MDD)	**-0.27**	**-0.25**	-0.02
4 (SS)	-0.09	-0.10	**0.44**

As the variables are taken from the BPRS, specifically developed to help diagnose psychiatric patients, it is not surprising that every single feature appears to have a strong input in at least one component, and is therefore considered relevant in the characterization of the archetypal patient. Inspecting the loadings in Table 3.5 divulges that Component I shows high positive loads on symptoms present in typical MDM patients and low loadings elsewhere; therefore, Component I is labeled as MDM. Following a similar vein, Components II and III are labeled PS and SS, respectively.

Having dealt with the labeling of the components one can now assign the clusters accordingly as Clusters 1 to 4 are catalogued as PS, MDM, MDD and SS, respectively. As PS, MDM and SS clusters are predominantly related to a unique component their symptoms are attributed respecting the concerned component. An exception on the MDD cluster is evidenced in the roughly equally negatively weighted combination of PS and MDM symptoms - the negative signs implying the absence of MDM or PS symptoms in MDD (the reverse analogy holding true as well).

Table 3.5: Rotated loadings of 17 variables on the components of the RKM solution ($Q = 3$); MDM is manic-depressive manic, SS is simple schizophrenic and PS is paranoid schizophrenic. Loadings relevant to categorize each component are indicated in bold.

Variables	Components		
	I (MDM)	II (PS)	III (SS)
Excitement	**-0.33**	0.11	**-0.22**
Grandiosity	**-0.37**	0.00	**-0.14**
Somatic concern	**0.21**	-0.36	-0.13
Anxiety	**-0.44**	0.06	-0.27
Emotional withdrawal	**0.26**	-0.25	**0.29**
Motor retardation	**0.34**	**-0.19**	-0.06
Depressive mood	**0.30**	**-0.31**	**-0.31**
Guilt feelings	0.28	**-0.22**	**-0.34**
Mannerisms and posturing	-0.04	**-0.18**	**0.40**
Hostility	-0.16	**-0.18**	**-0.31**
Suspiciousness	-0.15	**-0.43**	0.06
Hallucinatory behavior	-0.17	**-0.37**	0.15
Uncooperativeness	-0.26	**-0.23**	-0.06
Unusual tought content	-0.12	**-0.35**	0.06
Conceptual disorganization	-0.06	**-0.23**	0.12
Blunted effect	0.21	0.01	**0.44**
Tension	-0.22	0.01	**-0.25**

In Fig. 3.7 the observed scores are projected in **XA** and the cluster prototypes are visible. The paranoid schizophrenic agglomeration appears to be the one with smaller variability with reference to the annotated psychiatrists' grading. The two endured missclassifications have origins in different psychiatrists, denoting perhaps a subjective entity introduced in the classification, such as academic background or the clinicals' experience.

In conclusion, and providing a definite answer to the questions brought up, the RKM model allows, from the known descriptions, a good recovery of the four types of patients. The relationships between the symptoms and patients were cleared as PS, MDM and SS patients seemed to have specific patterns of symptoms; on the other hand, MDD patients contrasted with this behavior and showed symptoms

absent in PS and MDM patients, and seemed to not have symptoms present in PS and MDM.

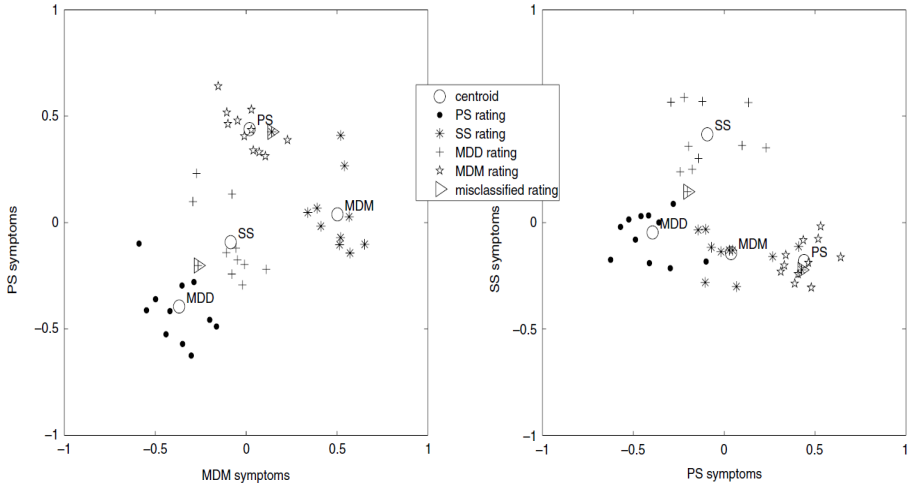


Figure 3.7: Observed scores projected to the RKM subspace, for (left) the MDM versus PS dimension, and (right) the PS versus SS dimension (inspired by [72]).

4

Clustering and Disjoint Principal Component Analysis

Contents

This chapter provides a description of the Clustering and Disjoint Principal Component Analysis algorithm [75], expressed as a solution to an optimization problem. The numerical model of this iterative process is also detailed. The algorithm follows the methods alluded in Chapter 3 and introduces new constraints in the allocation of variables in subgroups.

When dealing with real data sets, there may be the need to reduce not solely the dimension of the feature space, but also to unveil some patterns among the objects. The addressed methodology obtains the desirable scenario for data interpretation and visualization by attaining non overlapping clusters of objects and disjoint and sparse classes of variables. It is heavily linked to RKM, distinguishing itself due to constraints imposed on the variable allocation matrix $\mathbf{A}$. Here in CDPCA [75], because there is a particular interest in defining factors of maximal variance to specify the classification of the features, the preferred approach is the maximization of the between-class deviance in the reduced space, as performed by reduced k-means.

Following the discussion of the clustering and disjoint PCA model and the least-squares estimation of the model, performance is assessed by applying the methodology firstly to a data set describing short-term macroeconomic scenario of OECD countries and a second data set analyzing the small round blue cell tumors.

4.1 Model Definition

The CDPCA model is the result of applying the k-means algorithm on a data matrix to be able to represent the objects by their centroid, and simultaneously performing PCA on the transformed data matrix [75]. The main goal is to find a cluster of objects along a set of centroids and at the same time partition the variables into a set of disjoint components, while maximizing the between cluster deviance in the reduced space of the components. The model can be expressed by

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{X}} + \mathbf{E}_1 \qquad \text{(K-means step on } \mathbf{X}\text{)} \tag{4.1a}$$

$$= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}_1 + \mathbf{E}_2 \qquad \text{(PCA step on } \mathbf{U}\bar{\mathbf{X}}\text{)} \tag{4.1b}$$

$$= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E} \tag{4.1c}$$

where $\mathbf{E}$, $\mathbf{E}_1$ and $\mathbf{E}_2$ are $I \times J$ error matrices and $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$. From Eq. (4.1c) one can write the CDPCA model in order of $\mathbf{E}$, $\mathbf{E} = \mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}$, and rewrite the CDPCA problem into a minimization of the error matrix,

$$\min_{\mathbf{U},\bar{\mathbf{Y}},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2, \tag{4.2}$$

where $\mathbf{U}$ is a binary and row stochastic matrix, $\bar{\mathbf{Y}}$ is an object centroid matrix in the reduced space and $\mathbf{A}$ is a column-wise orthonormal matrix where each row contributes to only one column. The transformation of the problem reflected in (4.2) to the equivalent maximization of the between cluster deviance in the

reduced space follows as

$$\begin{aligned}\|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2 &= \|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\|^2 \\ &= \mathrm{tr}\{[\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}][\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}]^{\mathrm{T}}\} \\ &= \mathrm{tr}\{[\mathbf{U}\bar{\mathbf{X}}\mathbf{A}][\mathbf{U}\bar{\mathbf{X}}\mathbf{A}]^{\mathrm{T}}\} \\ &= \|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\|^2,\end{aligned}$$

and since $\|\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2 = \|\mathbf{X}\|^2$ - $\|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2$ is applicable,

$$\begin{aligned}\|\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2 + \|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2 &= \mathrm{tr}\{[\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}][\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}]^{\mathrm{T}}\} + \mathrm{tr}\{[\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}][\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}]^{\mathrm{T}}\} \\ &= \mathrm{tr}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\} - 2\mathrm{tr}\{\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\} + 2\mathrm{tr}\{\mathbf{U}^{\mathrm{T}}\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\bar{\mathbf{X}}^{\mathrm{T}}\} \\ &= \mathrm{tr}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\} - 2\mathrm{tr}\{\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\} + 2\mathrm{tr}\{\mathbf{U}^{\mathrm{T}}\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\bar{\mathbf{X}}^{\mathrm{T}}\} \\ &= \mathrm{tr}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\},\end{aligned}$$

the problem can be rewritten to

$$\max_{\mathbf{U},\bar{\mathbf{X}},\mathbf{A}} \|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\|^2.$$

The inclusion of the auxiliary matrix $\mathbf{V}$, whose nonzero entries identify nonzero elements of $\mathbf{A}$, and knowing $\bar{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{A}$, the CDPCA problem can be tackled as a quadratic mixed continuous and integer problem [76] given by

$$\max F = \|\mathbf{U}\bar{\mathbf{Y}}\|^2,$$

and constrained to the allocation of I objects and P clusters (4.3a), to the allocation of J variables into Q disjoint components (4.3b) and constrained to the PCA implementation (4.3c).

$$u_{ip} \in 0,1,\ i = 1,...,I;\ p = 1,...,P \quad \text{and} \quad \sum_{p=1}^{P} u_{ip} = 1,\ \ i = 1,...,I;\ p = 1,...,P, \tag{4.3a}$$

$$v_{jq} \in 0,1,\ j = 1,...,J;\ q = 1,...,Q \quad \text{and} \quad \sum_{q=1}^{Q} v_{jq} = 1,\ \ j = 1,...,J;\ q = 1,...,Q, \tag{4.3b}$$

$$\sum_{j=1}^{J} a_{jq}^2 = 1,\ q = 1,...,Q \quad \text{and} \quad \sum_{j=1}^{J} a_{jq}a_{jr} = 1,\ \ q = 1,...,Q-1;\ r = q+1,...,Q. \tag{4.3c}$$

The maximum dissimilarity of centroids is represented by the maximization of the objective function. The total variance of the data in the reduced space can be expressed as

$$F = \mathrm{tr}(\mathbf{U}\bar{\mathbf{Y}}(\mathbf{U}\bar{\mathbf{Y}})^{\mathrm{T}}) = \mathrm{tr}((\mathbf{U}\bar{\mathbf{Y}})^{\mathrm{T}}\mathbf{U}\bar{\mathbf{Y}}). \tag{4.4}$$

To solve (4.4) an iterative algorithm called alternating least-squares algorithm (ALS) is proposed in [75].

Algorithm 4.1. CDPCA.

1: Concerning the objects:

- allocate the I objects into P clusters. ▷ matrix $\mathbf{U}$
- calculate the centroids in the space of the observed variables. ▷ matrix $\bar{\mathbf{X}}$
- identify objects by the cluster centroids in the observed variables space. ▷ matrix $\mathbf{Z}$

2: Concerning the variables:

- allocate the J variables into Q subsets. ▷ matrix $\mathbf{V}$
- obtain the loadings of the Q components. ▷ matrix $\mathbf{A}$
- calculate the centroids in the reduced space of the CDPCA components. ▷ matrix $\bar{\mathbf{Y}}$
- identify the objects in the reduced space of the CDPCA components. ▷ matrix $\mathbf{Y}$

Figure 4.1 illustrates the algorithm's progression. In step 1, and after standardizing the data composed of I objects and J variables, the objects are assigned to P clusters following matrix $\mathbf{U}$. Afterwards, matrix $\mathbf{Z}$ is created by assigning each row of the data matrix with the correspondent centroid. In step 2, matrix $\mathbf{V}$ specifies the allocation of the J variables into Q disjoint components, and matrix $\mathbf{A}$ the CDPCA loadings. To obtain these two matrices an iterative algorithm covers row-by-row, column-by-column, matrices $\mathbf{V}$ and $\mathbf{A}$ in order to maximize the objective function F.

At the end of one iteration the component score matrix and the object centroid matrix in the reduced space are found, and thus the I objects of the data matrix are allocated into P clusters, displayed in a lower dimensional space of Q disjoint components. In the coming iteration the input matrix makes its appearance in the form of $\mathbf{Y}$. The algorithm is interrupted after assessing the solutions and checking if the difference between the two consecutive objective function values F_k and F_{k+1} is smaller than a specified tolerance threshold.

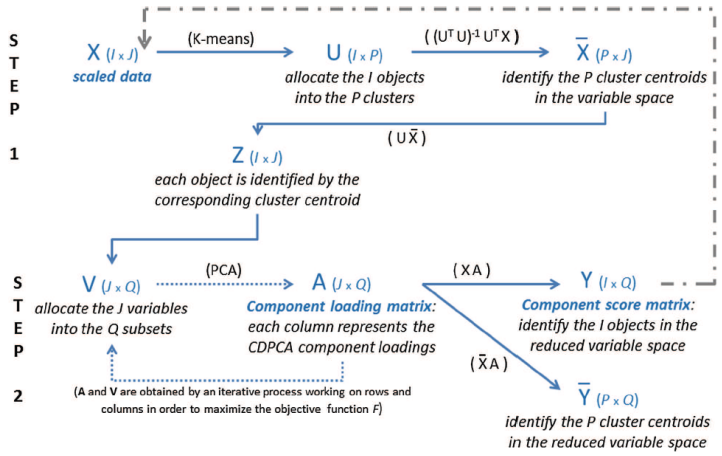


Figure 4.1: The two basic steps of one iteration of the Alternating Least-Squares algorithm for performing CDPCA (extracted from [76]).

Since F is bounded above the output of each iteration will converge to a stationary point - at least a local maximum of the problem [75]; the algorithm can then be considered an heuristic and thus, to

achieve a global maximum, several distinct initializations of the allocations' matrices $\mathbf{U}$ and $\mathbf{V}$ should be considered.

4.2 Algorithms

In this section the algebraic features of the algorithm are further elaborated.

4.2.1 Initialization

At the beginning, the data matrix $\mathbf{X}$ is standardized and the parameters to perform CDPCA are initialized following Algorithm 4.2.

Algorithm 4.2. CDPCA initialization.

1: Parameters associated to the objects:
- Randomly generate the initial object assignment matrix $\mathbf{U}_0$ so that there is only a nonzero element equal to 1 per row.
- Compute mean of each variable into each object cluster and the object centroid matrix $\bar{\mathbf{X}}_0$.
- All the objects are identified by its cluster centroids.

2: Parameters associated to the components:
- Randomly generate the initial variable assignment matrix $\mathbf{V}_0$ so that there is only a nonzero element equal to 1 per row.
- Construct the CDPCA component loading matrix A_0 column-by-column, solving Q independent PCA subproblems, one by each column.

The original variables belonging to the q-th CDPCA component are identified by the nonzero elements of the q-th column of V_0. These elements will be considered in the PCA subproblem to obtain the nonzero elements of the q-th column of $\mathbf{A}_0$ that correspond to the first principal component obtained from PCA applied on the submatrix $\mathbf{W}_0^{(q)}$. This submatrix is extracted from the centroid-based data matrix where each object is identified by the corresponding centroid, $\mathbf{Z}_0 = \mathbf{U}_0\bar{\mathbf{X}}_0$, and restricted to the original variables assigned into the q-th column of $\mathbf{V}_0$. Thus, the q-th column of $\mathbf{A}_0$ provides the direction vector with maximum variability amongst the centroids in the subspace defined by the original variables assigned to the q-th column of $\mathbf{V}_0$.

4.2.2 General Iteration

After performing the initialization steps, at the beginning of the $(k+1)$-th iteration of the algorithm the matrices $\bar{\mathbf{X}}_k$, $\mathbf{V}_k$, $\mathbf{A}_k$ and $\bar{\mathbf{Y}}_k$ are all known. Making $\mathbf{X}_{k+1}$ the result given by one run of the K-means algorithm on the score matrix $\mathbf{Y}_k = \mathbf{X}\mathbf{A}_k$ starting from the object centroid matrix $\bar{\mathbf{Y}}_k$ in the reduced space, the CDPCA general iteration is defined in Algorithm 4.3.

Algorithm 4.3. CDPCA general iteration.

1: Parameters associated to the objects:

- P new clusters are obtained updating the centroid matrix by $\bar{\mathbf{X}}_{k+1} = (\mathbf{U}_{k+1}^T \mathbf{U}_{k+1})^{-1} \mathbf{U}_{k+1}^T \mathbf{X}$ and the object centroid-based matrix by $\mathbf{Z}_{k+1} = \mathbf{U}_k \bar{\mathbf{X}}_k$.

2: Parameters associated to the components:

- $\mathbf{V}_{k+1}$ and $\mathbf{A}_{k+1}$ are sequentially updated row-by-row, and in each row, the process is consecutively performed column-by-column, in an interdependent relationship with the maximization of the objective function F.

These steps are repeated while there are empty clusters. At the end of the first step of the general iteration every single cluster should be assigned with at least one object. If any cluster becomes empty then a selection process takes place where half of the objects on the biggest cluster are assigned into one of the empty clusters.

Regarding the second step of the algorithm, to update matrix $\mathbf{V}_k$, that specifies a partition of the original variables into Q disjoint components, each original variable is evaluated to find the component that maximizes the objective function F. To begin with, the first row of $\mathbf{V}_k$ is updated by detecting for which column the allocation of nonzero elements achieves better results in maximizing F. For the first variable in $\mathbf{V}_{k+1}$ the best component is selected by solving Q PCA subproblems associated with $\mathbf{W}_{k+1}^{(q)}$. In the q-th PCA subproblem the first principal component is calculated determining the update of the q-th column of $\mathbf{A}_{k+1}$, and the centroid matrix in a reduced space, and the objective function value can be calculated by $\bar{\mathbf{Y}}_{k+1} = \bar{\mathbf{X}}_k \mathbf{A}_{k+1}$ and $F_{k+1} = \text{tr}((\mathbf{U}_{k+1} \bar{\mathbf{Y}}_{k+1})^T \mathbf{U}_{k+1} \bar{\mathbf{Y}}_{k+1})$.

The same rationale is repeated for the following rows of $\mathbf{V}_k$ making $\mathbf{V}_{k+1}$ update row-by-row. Taking into account the J original variables to obtain $\mathbf{V}_{k+1}$ and $\mathbf{A}_{k+1}$, $J \times Q$ subproblems are solved. In each subproblem, a subspace of variables is considered and the best direction with maximum explained variability is obtained performing a PCA step, leading to a maximization of the between cluster deviance given by $F_{k+1} \ / \ \|\mathbf{Y}_{k+1}\|^2$. These attributed components aren't sorted in a traditional, decreasingly way, and an additional step of rearranging the output into a classical form of representation is done.

4.3 Empirical Examples

The clustering and disjoint PCA has been applied to two real data sets to show the performances of the methodology. The first data set describes the short-term scenario of the OECD countries analyzed also with tandem analysis and factorial k-means. The second data set takes into account microarray experiments of Small Round Blue Cell Tumors (SRBCT) of a childhood cancer study by Khan et al. [77].

4.3.1 Latest Short-term Macroeconomic Scenario

To test the ability of CDPCA in identifying classes of similar economies and helping understand the relationships within the set of financial indicators, a short-term scenario from December 2017 on the

macroeconomic behavior of the national economies of twenty countries, members of the Organization for Economic Co-operation and Development (OECD), is analyzed. The six main economic indicators whose interactions reflect the economies are: leading indicator (LI), trade balance (TB), unemployment rate (UR), interest rate (IR), gross domestic product (GDP), net national savings (NNS), shown in more detail at Table 4.1.

Table 4.1: Latest short-term indicators and economic performance indicators (December 2017).

Country	Gross Domestic Product (GDP)	Leading Indicator (LI)	Unemployment rate (UR)	Interest rate (IR)	Trade balance (TB)	Net national savings (NNS)	Class membership obtained by the CDPCA model
Australia (A-lia)	5.8	21.9	5.5	2.6	0.7	2.7	1
Austria (A-tria)	4.6	12.4	4.9	0.6	-1.8	7.9	1
Belgium (Bel)	3.4	9.3	6.3	0.7	4.9	3.8	1
Canada (Can)	5.4	17.9	5.8	1.8	-0.7	2.8	1
Denmark (Den)	3.8	10.9	5.2	0.5	2.9	12.2	1
Germany (Ger)	3.8	15.0	3.4	0.3	7.7	10.1	1
Japan (Jap)	1.5	8.5	2.5	0.1	0.5	4.2	1
Mexico (Mex)	8.3	21.9	3.4	7.3	-0.9	5.9	1
Netherlands (Net)	4.4	10.7	3.9	0.5	9.3	11.9	1
Norway (Nor)	5.8	14.1	3.9	1.6	5.5	17.7	1
Sweden (Swe)	4.4	18.5	6.1	0.7	-0.2	12.5	1
Switzerland (Swi)	1.4	13.3	3.3	-0.1	5.3	13.7	1
Finland (Fin)	3.7	6.1	8.1	0.6	-0.9	1.0	2
France (Fra)	2.8	9.5	9.2	0.8	-3.5	2.6	2
Italia (Ita)	2.1	-0.1	11.2	2.1	2.7	0.7	2
Portugal (Por)	4.1	-0.4	7.4	3.1	-7.2	-2.3	2
Spain (Spa)	4.0	5.4	15.9	1.6	-2.3	3.9	2
United Kingdom (UK)	3.8	15.5	4.2	1.2	-6.8	0.0	2
United States (USA)	4.1	16.9	3.8	2.3	-4.1	3.7	2
Greece (Gre)	2.0	-18.5	20.1	6.0	-12.2	-7.6	3

[a] Source: OECD, Paris (2017).
[b] Definitions and notes: GDP – Percentage change from previous year; seasonally adjusted; LI – A composite indicator based on other indicators of economic activity (qualitative opinions on production or employment, housing permits, financial or monetary series, etc.), which signals cyclical movements in industrial production from 6 to 9 months in advance; UR – Percentage of civilian labor force; seasonally adjusted; IR – long-term; TB – (goods) % of GDP at current price 2017; NNS – % of GDP at current prices 2017.

After standardizing the variables no significant correlation is observed between the six economic factors and the analysis was carried out computing the first two principal components, classifying countries on the basis of those initial object scores. The results are shown in Table 4.2 and Fig. 4.2. The k-means algorithm was run on the first two PCA starting from random partitions. It was necessary to run k-means for a large number of initial random starts to dissipate the influence of local optima convergence (running 1 000 times k-means found the present optimal solution after 537 runs). The first two components explain 54% and 27% of the total variance. The first dimension is characterized almost equivalently by net national savings, unemployment rate, leading indicator (7%) and trade balance (6%); the other two variables explain 3%, while 69% is due to the interrelations (sum of the covariances) among the six variables. The second PCA dimension is characterized mainly by gross domestic product (30%), interest rate (22%) and only slightly by leading indicator (6%)(the remaining three variables explain 4%), while 38% is due to interrelations among the six variables. Note how leading indicator characterizes

both dimensions. Countries clusters, in Fig. 4.2, are highlighted by dotted ellipses. The between cluster deviance of the optimal solution was equal to 51.9% of the total deviance.

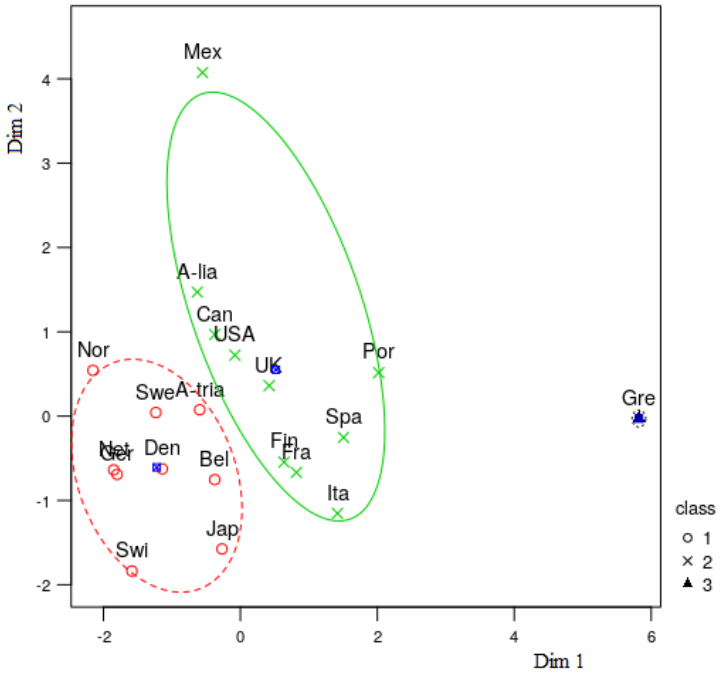


Figure 4.2: Tandem analysis results on OECD dataset. K-means clustering (k = 3) computed on the first two principal components (1 000 random start for k-means). Points without labels represent the centroids of clusters. Clusters of countries are highlighted by ellipses.

The classification of the groups can provide a depiction of the economic progression in the past few years, reflecting the divergent behaviors materializing in the set of countries, particularly throughout the European region. Demonstrating instances of faster and slower growth patterns, sizable differences are shown between the first and second, and third classes. Prosperity in countries such as Italy, Portugal and Spain translate skeletal convergence, where less developed economies in Europe close in upon the more enriched ones. There are also disparities around core countries of the monetary Union created in the last century, with Germany facing opposite sides of the spectrum when compared to some other euro area countries (the highlight being France). The long-standing stability witnessed in the northern Europe region helps explain the strong position taken by its forming countries in the examined scenario. In summary, a growth below potential rates is observed for the United Kingdom and the United States. Mexico, in the second group, is characterized by a high inflation (only comparable to Greece) and a high GDP that propels the country to the group's extremity; while Greece, with a very negative trade balance and leading indicator, and record-high employment rate, appears as a clear outcast.

The results of the CDPCA are reported in Table 4.2 and Fig. 4.3. The three clusters of countries are also highlighted by three dotted ellipses. The CDPCA was run 750 times to increase the chance to find the global optimal solution and that result was found 6 times in that time. In the best execution the algorithm converged after 5 iterations (with convergence tolerance value equal to 10^{-5}). The two components of the clustering and disjoint PCA explain almost the same part of the variance explained by the PCA (39% and 38%, respectively).

Table 4.2: Component loadings for PCA and CDPCA.

Variables	PCA loadings		CDPCA loadings	
	Component 1	Component 2	Component 1	Component 2
GDP	-0.192	**0.700**	0	**-0.353**
LI	**-0.464**	**0.305**	0	**-0.683**
UR	**0.478**	-0.091	**0.639**	0
IR	0.294	**0.592**	0	**-0.532**
TB	**-0.449**	-0.223	**0.584**	0
NNS	**-0.482**	-0.094	**0.613**	0

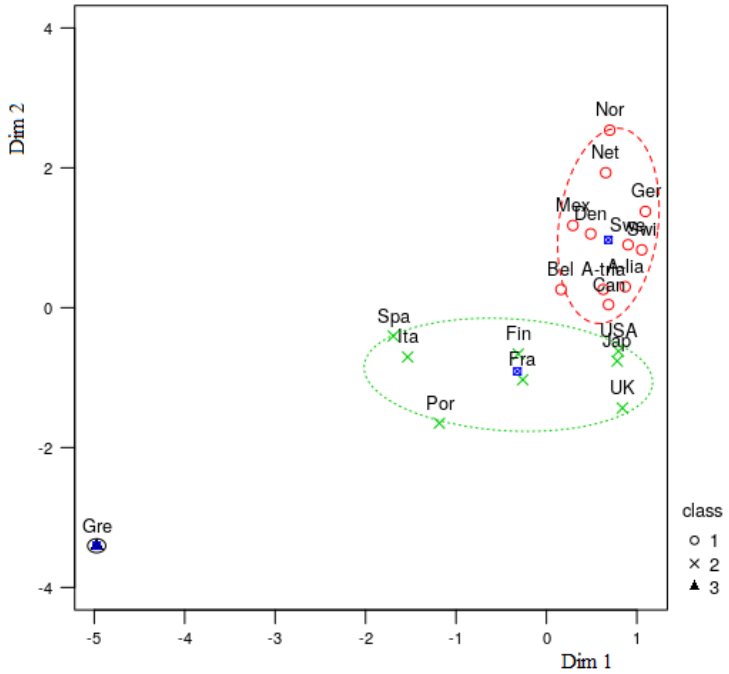


Figure 4.3: Clustering and disjoint PCA results on OECD dataset. Points without labels represent centroids of clusters. Clusters of countries are highlighted by ellipses.

The first component of the CDPCA is characterized mainly by unemployment rate (18%), net national savings (16%) and trade balance (15%).Therefore, the components of the PCA and disjoint PCA are almost the same both in terms of variance explained and of variables most contributing to specify the components. However, the disjoint PCA clearly shows more homogeneous clusters (between cluster deviance of CDPCA is $||\mathbf{U}\bar{\mathbf{Y}}||^2/||\mathbf{Y}||^2 = 87\%$ of total deviance, while between cluster deviance of the tandem analysis is only about 52% of the total deviance). In CDPCA there is a 51% contribution due to interrelations among UR, NNS, and TB. Note that only leading indicator appears differently from PCA. The second CDPCA component is characterized by leading indicator (21%), interest rate (13%) and only slightly by gross domestic product (6%); while 60% contribution is due to interrelations among LI, IR and GDP.

The classification on the two dimensions defined by the CDPCA model is similar to the one of the tandem analysis with Australia, Canada and Mexico moving to class 1, and Japan doing the converse.

Touching on the results of factorial k-means, in Table 4.3, the first component is specified by the main contribution of TB and NNS, while the second component is mainly characterized by GDP, LI

Table 4.3: Correlation between variables and the two factors specified by the FKM analysis .

	GDP	LI	UR	IR	TB	NNS
Dim 1	-0.177	-0.066	**-0.321**	0.189	**-0.637**	**0.583**
Dim 2	**0.587**	**-0.755**	-0.289	**-0.419**	0.039	-0.033

[a] Contributions with absolute values larger or equal to 0.3 are highlighted.

and IR. Building a comparison with CDPCA it can be observed that the only rearrangement is in the way UR contributes somewhat similarly to both dimensions, as it happens with LI for PCA (in tandem analysis). This ambiguity in the explanation of components causes difficulties in their understanding and interpretation due to the overlap of the same variables in explaining different factors. Naturally, the differences of components produce different classifications, Fig. 4.4.

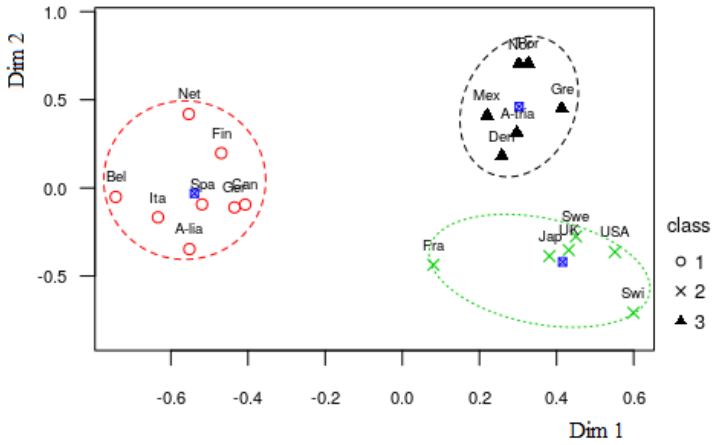


Figure 4.4: Factorial k-means results on OECD dataset. Points without labels represent centroids of clusters. Clusters of countries are highlighted by ellipses.

The number of clusters for objects can be chosen guided by the indices introduced in the literature. When speaking of variables the number of partitions should be smaller or equal to the maximum number of eigenvalues since we are expecting more clusters than latent features. By examining this example it can be observed that there isn't a instance of the two groups of variables that possesses more than a single eigenvalue greater than one. Accordingly, CDPCA has found two clusters of variables, each one explained by a subset of variables only expressing a component variable. Contrarily, if further separations are enforced and more clusters of features are necessary then the best outcome goes on to divide the second group, leaving variable IR alone in the case of a feature division into 3 clusters.

Please note that the orthogonality of the produced components is not enforced or necessarily but that when the number of variable partitions is accurately selected the correlation between components is regularly small (this example showcases a correlation of 0.11 between both components).

4.3.2 Small Round Blue Cell Tumors Data

The gene expression data[2] originally described the genomic information of 88 individuals. From those 88 examples five were cases of non-SRBCT occurrences and were removed from the testing samples as we which to only study subtypes of the same disease. The data included an evaluation of 2308 genes and encompassed 29 cases of Ewing sarcoma (EWS), categorized as 1, 11 cases of Burkitt lymphoma (BL), categorized as 2, 18 cases of neuroblastoma (NB), categorized as 3, 25 cases of rhabdomyosarcoma (RMS), categorized as 4.

Once again, prior to performing tandem analysis the variables were standardized on the columns representing the 2308 genes. The analysis was carried out computing the first principal components, classifying patients on the basis of first 21 component scores that explained 71.0% of the total variance. The results are shown in Fig. 4.5. The k-means algorithm was run on top of the PCA starting from random partitions and it was necessary to run it for a large number of initial random starts to mitigate the impact of the presence of several local optima (running 250 times k-means found the present best solution after 103 runs). The first two components explain a mere 10% and 8% of the total variance. Both the first and second dimensions of PCA are characterized by a rather homogeneous contribution of a series of variables, and present a negligible percentage of interrelations contribution. The between cluster deviance of the optimal solution was equal to 28.8% of the total deviance and provided 23 correct predictions.

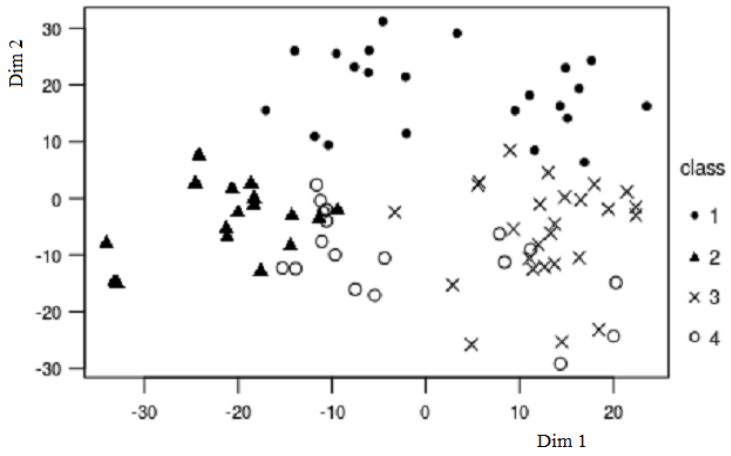


Figure 4.5: Tandem analysis results on the SRBCT dataset. Overlapping clusters make interpretation more difficult or impossible. Labels refer to k-means classification.

The classification into four groups is not at all intuitive due to the presence of overlapping points. With the exception of the first cluster whose boundaries could be easily delimited without clashing against the other partitions, the remaining three clusters all seem to intercept each other.

The CDPCA solution with P = I clusters (corresponding to the case of partitioning only variables) explains a relevant percentage of the total variance with the same number of components as PCA, but in such cases the interpretability of the clusters becomes a more serious task. In the new study 4 components were chosen to facilitate graphical analysis. The algorithm was run 250 times to increase the chance to

[2]See `http://research.nhgri.nih.gov/microarray/Supplement/`.

find the global optimal solution and the optimal solution was found 2 times in those runs, in Fig. 4.6. In the best solution, the algorithm converged after 9 iterations (with convergence tolerance value equal to 10^{-3}) after 347.78 seconds. The first two components of the clustering and disjoint PCA explain slightly less than the variance explained by the PCA, with 9% and 7%, respectively.

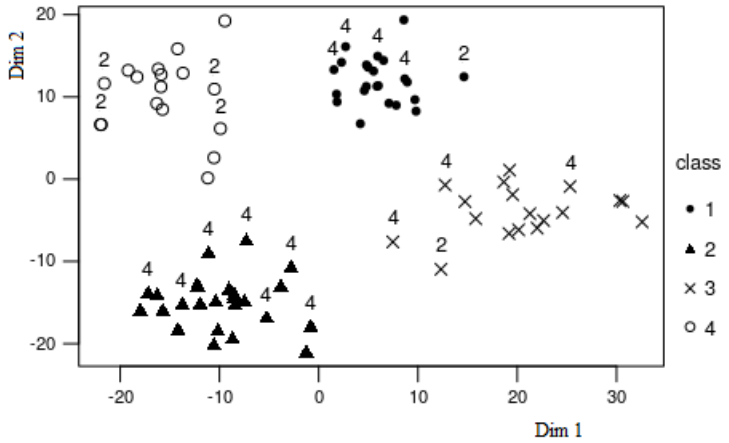


Figure 4.6: Clustering and disjoint PCA results on the SRBCT dataset. Numbers above points showcase the missclassifications of even classes, the ones that casted the most doubts in determining the corresponding cluster.

Despite the fact that the components of CDPCA possess rather consistent low scores like what was witnessed in the PCA components, the disjoint PCA clearly shows more homogeneous clusters (between cluster deviance of CDPCA is $||\mathbf{U}\bar{\mathbf{Y}}||^2/||\mathbf{Y}||^2 = 83\%$ of total deviance, while between cluster deviance of the tandem analysis is only about 30% of the total deviance).

In this difficult dataset the classification on the dimensions defined by the CDPCA model is slightly better than the one of the tandem analysis with 36 correct predictions. The 43.4% accuracy translates a less than optimal subspace recovery and could be explained by the elliptical cluster shape the algorithm tends to attribute not corresponding to the real pattern present in the data, and due to the possible presence of outliers that disturb the classification process.

5

Proposed Methodologies

Contents

This chapter presents two new methods aiming at increasing the interpretability of high-dimensional data analysis results.

After insurging against the use of tandem techniques, and reviewing the properties of the algorithm to be extended, two new methods are proposed. The first developed method is the Relaxed Clustering and Disjoint Principal Component Analysis (RCDPCA), resulting stemming from the inclusion of a fuzzy model in the object assignment phase of CDPCA. The second method is the Nested Clustering and Disjoint Principal Component Analysis (NCDPCA), a new pipeline whose purpose is to bridge the gap between integrated approaches and uncovering information hidden in sublayers of data by maximizing the between cluster deviance of the cluster and subclusters instances.

5.1 Relaxed CDPCA

The RCDPCA algorithm addresses the possibility of an object having characteristics associated with one or more clusters by instilling a fuzzified object stratification at each step of the iterative and converging process. The algorithm remains interested in maximizing the between cluster deviance by virtue of a greedy search and retains the properties of the work introduced in Chapter 4.

In the genome and DNA data studied in the development of these methods the focus was not on assuming or wishing to test a theoretical model of latent factors causing the observed variables and thus factor analysis was not considered. Instead, and with the interest of simply reducing the correlated observed variables to a smaller set of important independent composite variables, the use of PCA as the solution for the dimensionality reduction problem is retained, as it is an adequate approach to the task at hand.

The k-means implementation discussed in the Chapter 2 comes with its hiccups. Considering, for example, a symmetric dataset with two clusters and a point in the middle equidistant to the prototype of both of them, hard k-means assigns a not very intuitive crisp label to the point in question. Other properties also need consideration: hard k-means tends to get stuck in local minimum and has the necessity of conducting several runs with different initializations [78] or needing sophisticated initialization methods like the Latin hyper cube sampling [79], and the crisp memberships prohibit ambiguous assignments - when clusters are badly delineated or overlapping, there is a need to relax requirement $u_{ip} \in \{0, 1\}$.

Understanding Soft Clustering

Soft or fuzzy clustering allows gradual memberships of data points to clusters in $[0, 1]$, providing the flexibility to express data point that can belong to more than one cluster. These membership degrees offer a finer degree of detail of the data model by expressing how ambiguously/definitely a certain point $\mathbf{x}_i$ should belong to a certain cluster C_p. The solution spaces gain the form of fuzzy partitions of $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_I\}$. A probabilistic cluster partition is formally defined in Def. 5.1.

Definition 5.1. Probabilistic Cluster Partition.

Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_I\}$ be the set of given objects and let P be the number of clusters $(1 < p < I)$ represented by the fuzzy sets Γ_{C_p}, with $p = 1, ..., P$. Then we call $\mathbf{U} = (u_{ip}) = (\Gamma_{C_p}(\mathbf{x}_i))$ a probabilistic cluster partition of $\mathbf{X}$ if

$$\sum_{i=1}^{I} u_{ip} > 0, \qquad \forall\, p \in \{1, ..., P\}, \qquad \text{and} \tag{5.1a}$$

$$\sum_{p=1}^{c} u_{ip} = 1, \qquad \forall\, i \in \{1, ..., I\} \tag{5.1b}$$

hold. The $u_{ip} \in [0, 1]$ are interpreted as the membership degree of datum $\mathbf{x}_i$ to cluster C_p relative to all other clusters.

Constraint (5.1a) guarantees that no cluster is empty - a requirement in classical cluster analysis. The second constraint states that the sum of membership degrees must be one for each $\mathbf{x}_i$: each datum receives the same weight in comparison to all other data, making the groups exhaustive. The combination of these two conditions imply that no cluster can contain full membership of all data points and that membership degrees for a given datum resemble probabilities of being member of its corresponding cluster.

Let us develop a clear analytical understanding of the fuzzyfied version of the hard clustering method used in CDPCA, which describes a partition of $\mathbf{X}$ as a probability oriented problem. The fuzzy clustering criterion we discuss generalizes the within groups sum of square errors function J_f, initially reported by Dunn in [80] as an algorithm akin to hard c-means. Shortly thereafter, Dunn's function became a special case of the first infinite family of fuzzy clustering algorithms based on a least-squared errors criterion [81].

Definition 5.2. Fuzzy C-Means Functional.

Let $J_f : M_{fc} \times \mathbb{R}^{cp} \to \mathbb{R}^+$ be

$$J_f(\mathbf{U}, \bar{\mathbf{y}}) = \sum_{i=1}^{I} \sum_{p=1}^{c} (u_{ip})^m (d_{ip})^2 \tag{5.2}$$

where $\mathbf{U}$ is a fuzzy c-partition of $\mathbf{X}$ and u_{ip} denotes the grade of membership of the i-th data pairs in the p-th cluster;

$$\bar{\mathbf{y}} = (\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, ..., \bar{\mathbf{y}}_c) \in \mathbb{R}^{cp} \quad \text{with } \bar{\mathbf{y}}_p \in \mathbb{R}^P$$

is the cluster center or prototype of $\mathbf{u_p}$, $1 \leq p \leq c$;

$$(d_{ip})^2 = ||\mathbf{x}_i - \bar{\mathbf{y}}_p||^2 \quad \text{and } ||\cdot||$$

is any inner product induced norm on $\mathbb{R}^p$; and

weighting exponent $m \in [1, \infty[$.

The original Dunn's functional is obtained by setting $m = 2$ in Eq. (5.2). Inspecting J_f reveals that the distance between all data points $\mathbf{x}_i$ and a fuzzy prototype $\bar{\mathbf{y}}_p$ is considered as the measure of dissimilarity, $d_{ip} = ||\mathbf{x}_i - \bar{\mathbf{y}}_p||$; the squared distance is then weighted by $(\mathbf{u}_{ip})^m = (\mathbf{u}_p(\mathbf{x}_i))^m$, the m-th power of $\mathbf{x}_i$'s membership in the fuzzy cluster $\mathbf{u}_p$. Because each term of J_f is proportional to $(d_{ip})^2$, J_f is a squared error clustering criterion, and its least-squared error stationary points are solutions of

$$\min_{M_{fc} \times \mathbb{R}^{cp}} \{J_f(\mathbf{U}, \bar{\mathbf{y}})\}. \tag{5.3}$$

An infinite family of fuzzy clustering algorithms - one for each $m \in [1, \infty[$ - is attained via necessary conditions for solutions of (5.3). The basic theorem follows.

Theorem 5.1.

Assume $||\cdot||$ to be inner product, $m \in [1, \infty[$, let $\mathbf{X}$ have at least $P < I$ distinct points, and define $\forall\ i$ the sets

$$I_i = \{p | 1 \leq p \leq P; d_{ip} = ||\mathbf{x}_i - \bar{\mathbf{y}}_p|| = 0\},$$
$$\bar{I}_i = \{1, 2, ..., P\} - I_i,$$

then $(\mathbf{U}, \bar{\mathbf{y}}) \in M_{fc} \times \mathbb{R}^{cp}$ may be globally minimal for J_f only if

$$I_i = \emptyset \Rightarrow u_{ip} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ip}}{d_{ij}}\right)^{2/(m-1)}} \tag{5.4a}$$

or

$$I_i \neq \emptyset \Rightarrow u_{ip} = 0 \quad \forall\ p \in \bar{I}_i \qquad \text{and} \qquad \sum_{p \in I_i} u_{ip} = 1, \tag{5.4b}$$

and

$$\bar{\mathbf{y}}_p = \frac{\sum_{i=1}^{I} (u_{ip})^m \mathbf{x}_i}{\sum_{i=1}^{I} (u_{ip})^m} \quad \forall\ p. \tag{5.4c}$$

One obtains (5.4a) by fixing $\bar{\mathbf{y}} \in \mathbb{R}^{cp}$ and applying Lagrange multipliers to the variables $\mathbf{u}_{ip}$, Appendix A. This application requires one technical maneuver: we relax $\mathbf{U}$ so that the columns of $\mathbf{U}$ can be uncoupled and minimized term by term. Equation (5.4b) is the necessary alternate form for memberships of $\mathbf{x}_i$ when $\exists\ i$ that dictates $d_{ip} = 0$. In this particular case, and whenever this "singularity" occurs, it is imposed that $\mathbf{x}_i$ must have no membership in any cluster $\mathbf{u}_p$ where $d_{ip} > 0$. The requisites for the resolution of the algorithmic singularity in the second task of step 1 of Alg. 5.1 found below are now stated. We make note that the referred singularity $\mathbf{x}_i = \bar{\mathbf{y}}_p$, $\exists\ i, p$ hardly ever occurs in practice, since computerized round offs usually preclude this eventuality. Straightforwardly, (5.4c) is the necessary condition derived via unconstrained optimization of J_f with fixed $\mathbf{U}$ and having $\bar{\mathbf{y}}_p$ as variables.

The fuzzy c-means clustering algorithm is just a set of iterations through the necessary conditions (5.4),

constraints that hold for any inner-product norm metric, specifically a positive-definite matrix $\mathbf{A}$ that induces a norm via the weighted inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y} = \sum_{i=1}^{z} \sum_{j=1}^{z} \mathbf{x}_i a_{ij} \mathbf{y}_j.$$

Emphasizing the dependence of J_f on the norm defining matrix $\mathbf{A}$ yields

$$J_f(\mathbf{U}, \bar{\mathbf{y}}, \mathbf{A}) = \sum_{i=1}^{I} \sum_{p=1}^{c} (u_{ip})^m (\mathbf{x}_i - \bar{\mathbf{y}}_p)^T \mathbf{A} (\mathbf{x}_i - \bar{\mathbf{y}}_p). \tag{5.5}$$

Algorithm 5.1. Fuzzy C-Means.

1: Fix number of clusters P, $2 \leq p < I$, and fix m, $1 \leq m < \infty$. Initialize $\mathbf{U}^{(0)}$ and at iteration $l = 0, 1, 2, \ldots,$:
- calculate c fuzzy cluster centers $\bar{\mathbf{y}}_p^{(l)}$ with (5.4c) and $\mathbf{U}^{(l)}$.
- update $\mathbf{U}^{(l)}$ using (5.4a, 5.4b).
- compare $\mathbf{U}^{(l)}$ to $\mathbf{U}^{(l+1)}$: if $||\mathbf{U}^{(l+1)} - \mathbf{U}^{(l)}|| < \epsilon_L$ stop; otherwise return to first step of 1.

A more careful examination of Alg. 5.1 reveals that the non-singular case of $m = 1$, with the Euclidean norm $\mathbb{R}^p$, transforms this procedure in the hard c-means algorithm [82].

Traditional fuzzy c-means requires a set of input parameters including p, m, $\mathbf{U}^{(0)}$, $||\cdot||_{\mathbf{A}}$, and ϵ_L. There is an infinite family parametrized by m for each $\mathbf{A}$ selection - as $m \to 1$ fuzzy c-means theoretically converges to the hard c-means solution, where J_f has the added variable $\mathbf{A}$. Transversely, as $m \to \infty$, one sees u_{ip} in (5.4a, 5.4b) $\to (1/P)$ $\forall$ p, i, so $\bar{\mathbf{y}}_p \to \mu$ the centroid of $\mathbf{X}$ $\forall$ p. Thus when m approaches its maximal value the only optimal pair for J_f is $(\bar{\mathbf{U}}, \mu)$ = (centroid of M_{fc}, centroid of $\mathbf{X}$), and $J_f \to 0$. Above all, the larger m is the "fuzzier" are the membership assignments; and contrariwise, as $m \to 1$, the fuzzy c-means solutions become hard. The choice of m necessary to implement Alg. 5.1 controls the extent of membership sharing between fuzzy clusters in $\mathbf{X}$ in the form of a weighting exponent; its optimal choice is however not supported by any theoretical basis.

Addressing the Artificial Parameter

It is needless to say that the FCM method casts itself as a wondrous method mainly due to the introduction of the magical number m, but certain questions arise:

1. What is the physical meaning of the parameter m?
2. And what valid criterion can be used to decide the membership assignments?

To shed light upon this questions the notion of Maximum-Entropy Inference (MEI) is introduced: an unbiased inference method provided by information theory, more strictly, Shannon's concept of "amount of information" [83] for ill-defined problems on the basis of the given information. The MEI problem's

structure is one of finding a probability assignment or membership function u_{ip} which avoids bias while agreeing with whatever information is given. The problem can be formally presented as

$$\max \left\{ -\sum_{i=1}^{I} \sum_{p=1}^{c} u_{ip} \ln u_{ip} \right\}, \tag{5.6}$$

subject to $c_1, c_2, ..., c_m$, the m constraints of the given information. Defining the local loss function as the within-group sum of squared error

$$L = \sum_{i=1}^{I} \sum_{p=1}^{c} u_{ip} (d_{ip})^2, \tag{5.7}$$

the algorithm is set to minimize the objective function

$$J_f(\mathbf{U}, \bar{\mathbf{y}}) = \sum_{i=1}^{I} \sum_{p=1}^{c} u_{ip} (d_{ip})^2 + \gamma \sum_{i=1}^{I} \sum_{p=1}^{c} u_{ip} \ln u_{ip}, \tag{5.8}$$

where u_{ip} satisfies the conditions

$$0 \leq u_{ip} \leq 1 \qquad \forall\ i, p, \tag{5.9a}$$

$$0 < \sum_{i=1}^{I} u_{ip} < I \qquad \forall\ p, \tag{5.9b}$$

$$\sum_{p=1}^{c} u_{ip} = 1 \qquad \forall\ i. \tag{5.9c}$$

The functional (5.8) can simultaneously minimize the within cluster dispersion as it forces u_{ip} to minimize the weighted sum of squared distances, and maximize the negative weight entropy to determine clusters to contribute to the association of objects. The first term represents the cost function of the standard k-means algorithm, and is complemented by a second term that forces the maximization of the entropies of the distributions over the clusters described by u_{ip}. This way u_{ip} naturally distances itself from a crisp assignment, which is the minimum entropy setup.

The fuzzy clustering problem with maximum-entropy inference becomes one of finding a set of prototypes which minimizes (5.8) and a membership assignment which satisfies constraints (5.9c). To solve Eq. (5.6) subject to constraints (5.9) the alternating minimization procedure between membership matrix $\mathbf{U}$ and cluster center matrix $\bar{\mathbf{y}}$ can be applied to (5.8), resulting in the solutions

$$u_{ip} = \frac{e^{-\frac{d_{ip}^2}{2\sigma^2}}}{\sum_{j=1}^{c} e^{-\frac{d_{ij}^2}{2\sigma^2}}} \qquad \forall\ i, p,$$

$$\bar{\mathbf{y}}_p = \frac{\sum_{i=1}^{I} u_{ip}\mathbf{x}_i}{\sum_{i=1}^{I} u_{ip}} \qquad \forall\ i,$$

where σ is the Lagrangian multiplier from (5.7), Appendix B.

The entropy regularization allows us to avoid using the artificial fuzziness parameter m, replaced by the degree of fuzzy entropy γ, related to the concept of temperature in statistical physics, $2\sigma^2$. An interesting property and advantage of a membership regularization approach is that the prototypes are obtained as weighted means with weights equal to the membership degrees (rather than to the membership degrees at the power of m as is for the fuzzy c-means).

Input

Besides the input dataset $\mathbf{X}$, the optional specification of a ground truth classification vector, and the selection of the attributes to be considered through parameter $fixAtt$ (each position of the vector of length J contains the selected q variable cluster) this method possesses several input parameters: P, the number of clusters of objects; Q, the number of clusters of variables; $ent = \gamma$, the (positive) relaxation indicator, tol, the convergence tolerance; $maxit$, the maximum number of iterations; and r, the number of runs of the algorithm to achieve the final solution. The number of clusters chosen is unequivocally problem dependent.

Output

The RCDPCA is a soft classifier and the output consists of not only the partitions of objects and variables, but also relevant additional information such as the loop for computing the best solution, the component loading and score matrix, the centroids in the reduced space, the objective function's value, the between cluster deviance (and deviance over the total variability), a table containing the final classification and the error norm of the obtained RCDPCA model.

Algorithm

In Alg. 5.2 we have the sequence of operations needed to compute each object's relaxed assignment into clusters. First we compute the baseline object partition u_{ip} according to a distance metric. Then for each individual feature, we test non-null entries into each variable subset and extrapolate the arrangement that maximizes the established functional and guarantees the most interpretable subspace layout. The convergence to an objective function value lower than a defined threshold culminates in the final dimensionality reduction.

Algorithm 5.2. Relaxed CDPCA.

Require: X: input dataset, P: number of object clusters, Q: number of variable clusters,...
Ensure: data classification.

for *loop* in $1 : r$ **do**
  $Xs \longleftarrow$ Scale(X) ▷ Standardize columns and center data
  $iter \longleftarrow 0$ ▷ Iteration number
  $U \longleftarrow$ RandMat(I, P) ▷ Random initialization of object assignment matrix
  $V \longleftarrow$ RandMat(J, Q) ▷ Random initialization of variable assignment matrix
  if $fixAtt = \emptyset$ **then**
    $indexSelectFree \longleftarrow 1 : J$ ▷ Unassigned variables
  else
    ApplyFixAtt$(V, fixAtt)$ ▷ Variables are fixed to intended subset
    $indexSelectFree \longleftarrow$ which$(fixAtt == 0)$
  end if
  $X.bar \longleftarrow (U^T U)^{-1} U^T Xs$ ▷ Identify centroids in variable space
  $X.group \longleftarrow UX.bar$ ▷ Object identified by its centroid
  for j in $1 : Q$ **do**
    $A[, j] \longleftarrow$ PMEigen$(X.group, V)$ ▷ Power Method to calculate the largest eigenvalue
  end for
  $Y.bar \longleftarrow X.barA$ ▷ Centroids in reduced variable space
  $F_{max} \longleftarrow (UY.bar)^T (UY.bar)$
  $conv \longleftarrow 2 \cdot tol$
  while $conv > tol$ **do**
    $iter \longleftarrow iter + 1$
    $Y \longleftarrow XsA$ ▷ Component score matrix
    $U \longleftarrow$ FuzzyEnt(Y, P, ent) ▷ Reset and update U through the regularized fuzzy method
    Repeat steps 12, 13 ▷ Update structure X
    for j in $indexSelectFree$ **do** ▷ Iterate in the search for the best V and A
      $V[j,] \longleftarrow 0$
      for g in $1 : Q$ **do**
        $A \longleftarrow$ PMEigen$(X.group, V, g)$ ▷ Recalculate the largest eigenvalue
        if $(UX.barA)^T (UX.barA) > F_{max}$ **then**
          $F_{max} \longleftarrow (UY.bar)^T (UY.bar)$
          $posMax \longleftarrow g$
        end if
      end for
      $V[j, posMax] \longleftarrow 1$ ▷ Best subset for feature j
    end for
    Repeat steps 17, 22 ▷ Update structure Y
    $conv \longleftarrow (UY.bar)^T (UY.bar) - F_{max}$
    if $conv > tol$ **then** ▷ Check end loop conditions
      $F_{max} \longleftarrow (UY.bar)^T (UY.bar)$
      if $iter == maxit$ **then** break **end if**
    else
      break
    end if
  end while
end for
return $Obj_{RCDPCA} \longleftarrow$ BestLoop() ▷ Construct object with variables of the best run

Comments

This method extends the ideas of the Clustering and disjoint PCA. Being a fuzzy procedure its output is more flexible and intuitive in terms of analysis as it allows the definition and quantification of class membership adjusted through a relaxation parameter. The entropy regularization approach alleviates the weight computation to determine a prototype, but consequently transforms the assignment matrix entries to being based directly on distances instead of the more easily understandable distance ratios used in the original fuzzy construction.

5.2 Nested CDPCA

Taking inspiration in the automatization of the classification of objects, NCDPCA allows to unearth knowledge hidden in a secondary layer of data. The method projects the integrated approach of CDPCA in a cyclical manner, digging deeper and deeper into the behaviors and sub-behaviors that the data points characterize. Figure 5.1 provides a visual display of the nested process focusing on two layers determined by the same functional.

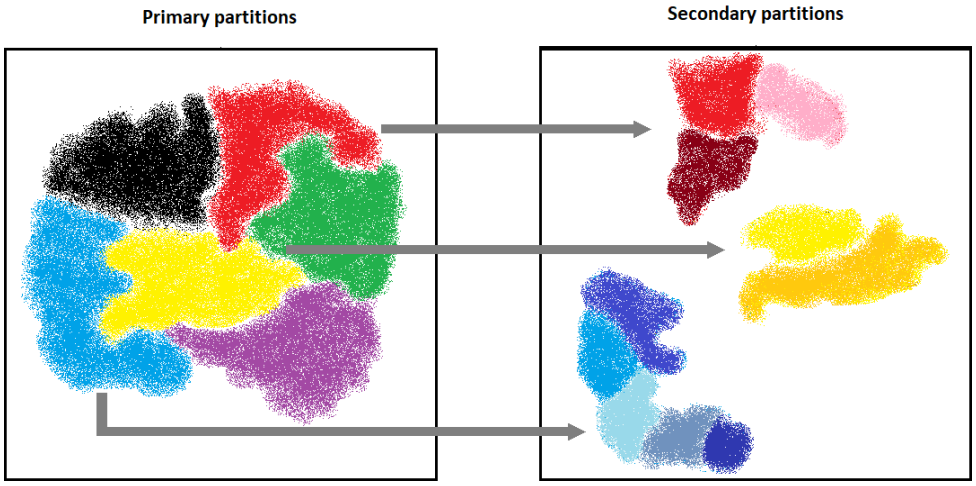


Figure 5.1: Illustrative graphical display of the NCDPCA algorithm. In this particular case, the first partition reveals six clusters represented by different colors. In the second phase of the nested process three of the clusters are reevaluated and further divisions can be retrieved for analysis.

Input

Similar to the RCDPCA method, it takes one additional input parameter in the form of a vector K that specifies the number of clusters for each partition obtained in the first layer (initial CDPCA run).

Output

As a nested process it returns a list of CDPCA objects containing the information already described in the first proposed methodology. Each object on the list arises from each run of the base algorithm.

Algorithm

The method can be implemented following Alg. 5.3. Firstly, we compute the first layer partitions and sequentially advance and examine each individual cluster originated in the first phase of the program.

Algorithm 5.3. Nested CDPCA.

Require: X: input dataset, P: number of object clusters, K: number of subclusters,...

Ensure: data classification.

1: $secondLayer \longleftarrow$ vector("$list$", $P+1$) ▷ Carries subclusters information
2: **for** $group$ in $0:P$ **do**
3: **if** $group == 0$ **then**
4: $firstLayer \longleftarrow$ CDpca($data, class, fixAtt, Q, P, ...$)
5: **else**
6: $data \longleftarrow data$[which($firstLayer\$U[, group] == 1$)
7: $class \longleftarrow class$[which($firstLayer\$U[, group] == 1$)]
8: $P \longleftarrow K[group]$
9: $secondLayer[[group]] \longleftarrow$ CDpca($data, class, fixAtt, Q, P, ...$)
10: **end if**
11: **end for**
12: $subClass \longleftarrow$ numeric(nrow($data$)) ▷ Build second layerś classification vector
13: **for** $group$ in $1:P$ **do**
14: **if** $K[group] == 1$ **then**
15: $subClass$[which($firstLayer\$U[, group] == 1$)] $\longleftarrow max(subClass) + 1$
16: **else**
17: **for** $subCluster$ in $1:K[group]$ **do**
18: $firstData \longleftarrow$ which($firstLayer\$U[, group] == 1$)
19: $secondData \longleftarrow$ which($secondLayer[[group]]\$U[, subCluster] == 1$)
20: $subClass[firstData[secondData]] \longleftarrow max(subClass) + 1$
21: **end for**
22: **end if**
23: **end for**
24: $secondLayer[[P+1]] \longleftarrow subClass$ ▷ Attach new full classification to list
25: **return** list($firstLayer, secondLayer$)

6

Results

Contents

In this chapter the main results obtained experimentally are presented and compared with the results of other approaches.

Cancer classification is intimately connected to cancer treatment and enhancements in this area contribute significantly to advances in patient recovery processes. Today, and always, the main challenges of cancer treatment are summed as the creation of target therapies to pathogenetically distinguish tumor types, and to maximize the treatment's efficacy and minimize its toxicity. Classically the focus has been on the study of the morphological appearance of a tumor but the fact that this analysis links similar histopathological appearances to different clinical courses with different responses to therapy is extremely limiting. In a few cases, such has the SRBC tumors studied in Chapter 4, this clinical heterogeneity has been "corrected" by dividing morphologically akin tumors into subtypes with distinct pathogeneses (in the mentioned case, the tumors are currently subclassified on a molecular level by Ewing sarcoma, Burkitt lymphoma, neuroblastoma, rhabdomyosarcoma, and other types). For many more tumors, relevant subclasses are likely to exist but have yet to be properly defined by molecular markers.

The disease's classification has been hard to accomplish partly because it has, on a historical level, relied on specific biological insights, rather than unbiased approaches for recognizing disorder subtypes. In this chapter we rely on a systematic approach based on global gene expression analysis through simultaneous expression monitoring of hundreds of genes using DNA microarrays, escaping the traditional descriptive rather than analytical microarray studies, and focusing on cell culture rather than primary patient material, in which genetic noise can obfuscate underlying reproducible expression patterns.

6.1 Leukemia Data

This clinical dataset[3] contains gene expression data from the leukemia microarray study of Golub et al. [84] and can be easily accessed from the R package *multtest*. The data contained 38 cases of human acute leukemias, 27 of which were acute myeloid leukemia (AML), categorized as 1, and the remaining 11 acute lymphoblastic leukemia (ALL), categorized as 2. These two classes are particularly relevant as their identification is critical for successful remission; chemotherapy regimens for ALL therapy differ significantly from AML (and vice versa), and although recovery can be accomplished cure rates are markedly diminished, and unwanted toxic effects are realized.

After scaling and centering 3051 genes the benchmark analysis was carried out computing the first principal components and classifying patients on the basis of the first 22 component scores explaining approximately 82% of the total variance. The results are shown in Fig. 6.1. The k-means algorithm was performed on the first two PCA starting from random partitions 500 times and found the present optimal solution with 13.6% between cluster deviance after 409 runs. The first two components explain 15.6% and 9.4% of the total variance, low scores anticipated due to the scale of the matrix handled.

[3]Find more at `http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?paper_id=43`.

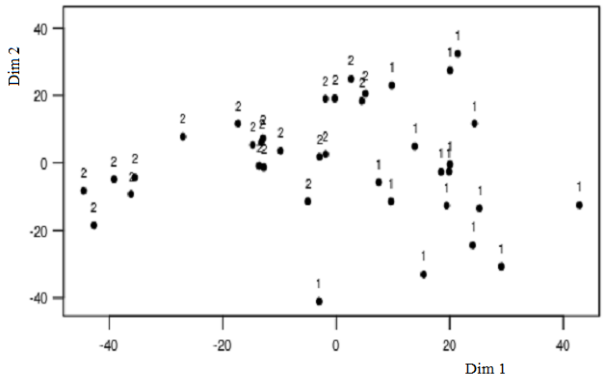


Figure 6.1: Tandem analysis. Classification of leukemia patients represented on the first two principal components.

The witnessed separation of the two groups is forced. No overlap occurs by definition but the plane responsible for the stratification appears to have a short margin to the flexible classification boundaries - another plausible explanation for the 8.4% of tumors correctly attributed in this test (accounting for 32 wrong labels). Although with such low percentages of correct classification estimation one wonders the validity of the classifier.

Without knowledge on the conditions of the data, and the disposition of the residuals, it is wise to attempt a multitude of angles and methods to extract the widest range of pertinent information. In Fig. 6.2 and Fig. 6.3 the integrated approaches studied in depth in Chapter 3 are revisited in this experiment with a distinct parameter Q of 10 for RKM and FKM to attempt to maintain a total variance explanation similar to the tandem analysis one and retain the validity of the comparison. The between cluster deviance of the first method is 78.4% and its output's sensitivity increase led to a jump to 31.6% of proper labeling (accounting for 26 wrong decisions). In the case of FKM a subspace was found such that the data points were aligned and were perfectly stratified into two clusters (the error function reached its lower bounded limit and 100% within cluster deviance was observed in the projected space) product of an

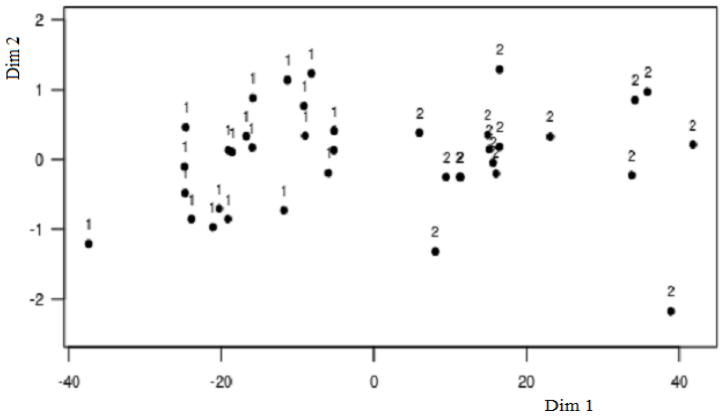


Figure 6.2: Classification of leukemia patients represented on the first two dimensions of the reduced k-means analysis.

arbitrary response. The presence of all the variables in the making of the new feature space culminated in the imposition of null subspace residuals and resulted in the data points being tidied up into a speckle. The end result is a 68.4% accuracy in classification, but a model not very flexible to further object

introductions into the system.

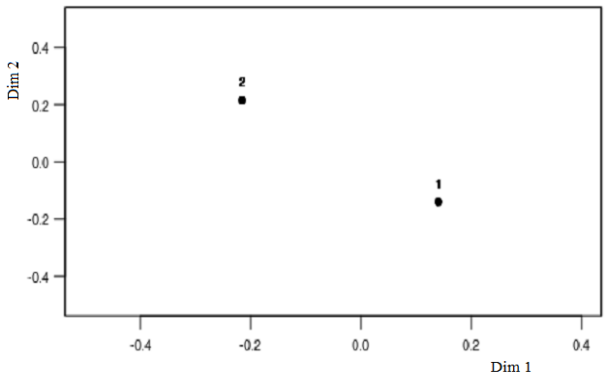


Figure 6.3: FKM clustering of leukemia patients in low dimensional space.

The results of the Relaxed CDPCA are reported in Fig. 6.4 and follow the convergence observed in regular CDPCA (not shown to avoid redundancy). The algorithm was run 150 times to increase the chance to find the global optimal solution. This optimal response was found 1 time in the 5th run, with the algorithm converging after 5 iterations, taking 310.5 seconds (against 179.3 seconds of an equivalent CDPCA execution). The two components of the relaxed clustering and disjoint PCA explain a low variance of approximately 5.2% and 5.0% respectively, and demonstrate high correlation, incentivizing the correction of the number of clusters of objects and variables, and also the occurrence of an initial test to assess the impact of similarly correlated variables in the definition of the components. Despite the

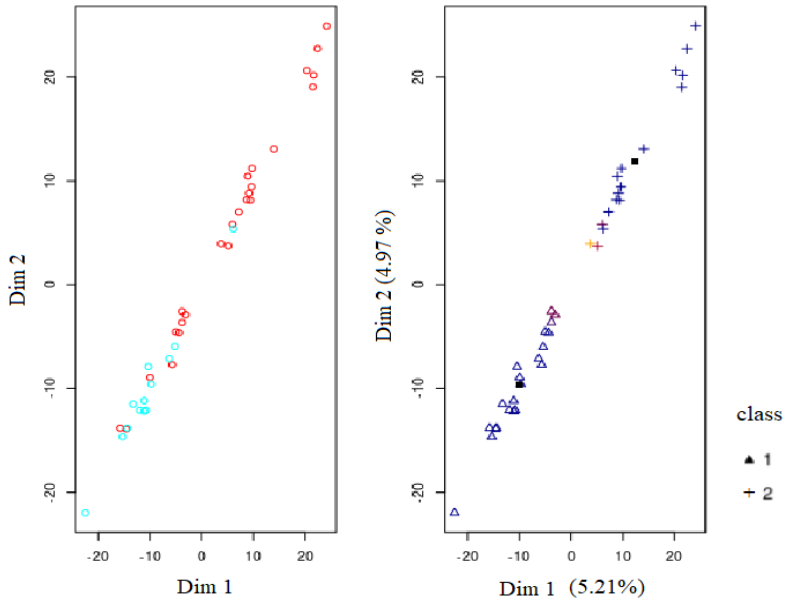


Figure 6.4: Fuzzy model incorporated in CDPCA applied to leukemia data. On the left the real classification with blue symbolizing the first class and red the second; On the right RCDPCA classification with brighter colors corresponding to less certain affirmations of class memberships.

fact that the components of RCDPCA denote similar variance explanation to the one seen in the other processes, the disjunction of the feature components clearly shows more distinctive clusters and presents a between cluster deviance of $||\mathbf{U}\bar{\mathbf{Y}}||^2/||\mathbf{Y}||^2 = 79\%$ of total deviance. In this dataset the classification on the dimensions defined by the RCDPCA model is a step above to the previous ones with 28 correct

predictions, being limited however by the 27% total variance explained. The 73.7% accuracy translates a less than optimal subspace recovery and can be explained by the deficient definition of Cluster 1 and due to the proximity of label 2's with the first partition.

These results demonstrate the feasibility of cancer classification based solely on the monitoring of gene expression and suggest a strategy for uncovering and predicting other types of cancer divisions independently of previous biological knowledge.

6.2 SRBCT Data Reevaluation

The objective of this test was to consolidate the theoretical advances expressed in the previous chapter. For that, we reevaluated the Small Round Blue Cell Tumors dataset for further analysis, Figs. 6.5 and 6.6.

RCDPCA took approximately 7 iterations to converge and brought in a between cluster deviance of 83%, explaining 8.0% and 6.6% of the total variance in the first and second component respectively, while taking 334.3 seconds. Assessing the color scheme we verify the confusion resides mainly in the even number classes, 2 and 4, the classes that produced the biggest analysis divergence in Chapter 4. The incorporation of a palette facilitates further analysis with the identification of a "danger zone" in the subspace and additional patients will be evaluated with more considerations. Space is opened for a specific reevaluation of this area using the Nested framework. The second application consisted of a

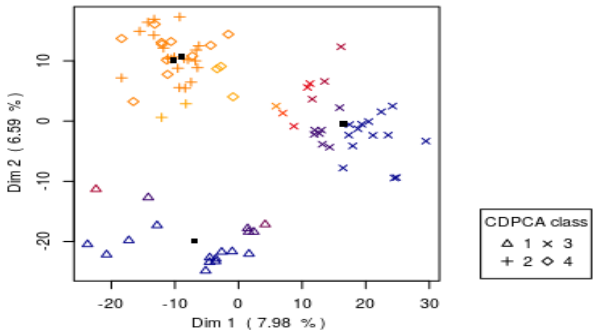


Figure 6.5: Relaxed clustering and disjoint PCA results on the SRBCT dataset.

tandem approach of PCA + fuzzy c-means and translates all the issues discussed in the previous chapters. Even with adjusted "fuzziness" parameters the doubts cast by this test are too much to bare and the system evaluation is dimmed unusable.

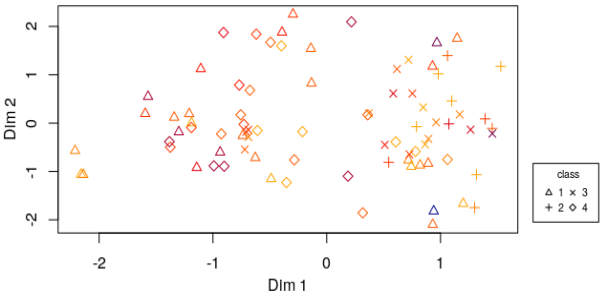


Figure 6.6: Results on the SRBCT dataset using the traditional tandem approach.

6.3 Hormonal Associated Cancer Discrimination

To evaluate the capacities of the nested methodology a special data set was assembled, combining entrances of several instances of breast and prostate cancer (hormonal), and melanomas, all belonging to the TCGA data repository. The three initial sets of collected cancer data were composed of two-dimensional arrays of 1204 positive cases of breast cancer patients over 19660 genes, 547 cases of prostate cancer over 19660 genes, and finally 84 instances of melanoma examples over 52746 genomes. To obtain a valid genomic matrix the genes were scaled, intercepted, and we were left with 1835 patients studied across 19435 genes.

This configuration allowed for the direct appraisal of whether this two tumorous types were identifiable through genetic makeup analysis, and permitted the appraisal of the clarity of the partition between three tumor subtypes perhaps carrying similar global pathogenetically characteristics.

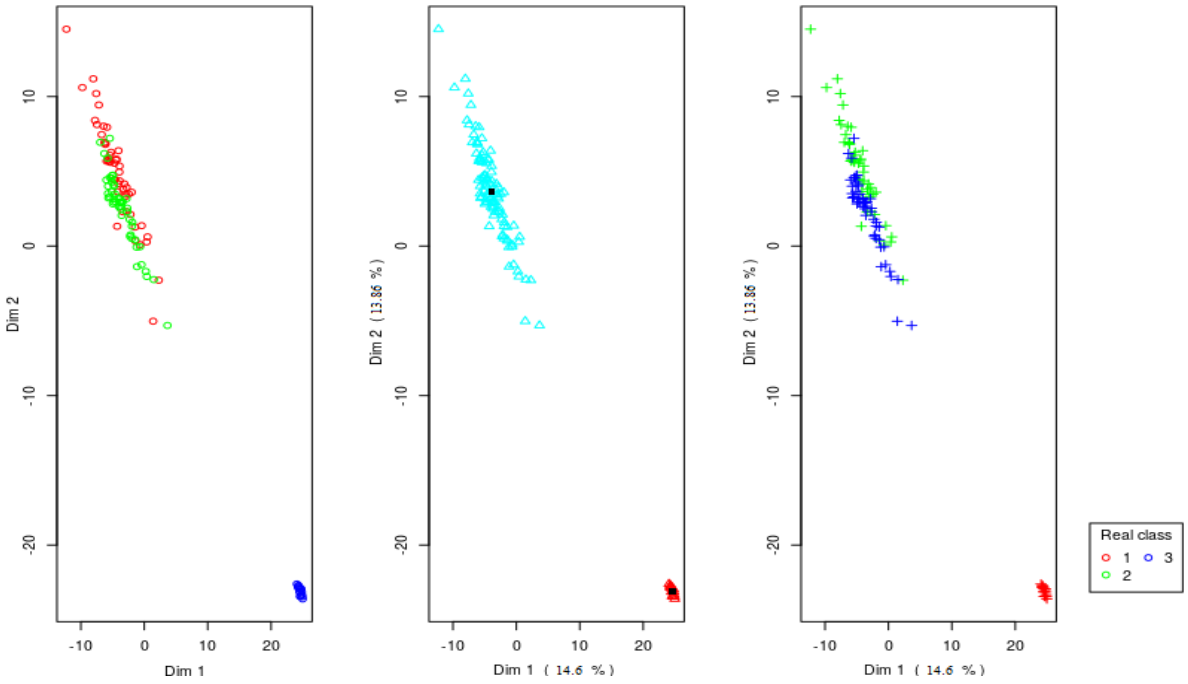


Figure 6.7: Nested clustering and disjoint PCA results on the TCGA custom cancer dataset. Real classification on the left; first layer partition at the middle; and second layer partition represented in the computed feature space on the right. Note that not all objects are drawn (removed to facilitate the perception of the clusters' overlap).

Visible in Fig. 6.7, the proposed method achieved a between cluster deviance of 92.1% in an average of 4 iterations for the first stratification of data and 86.0% between cluster deviance after 6 iterations in the following sublayer analysis, explaining practically 29% of the total variance with 2 components. The colors of the middle and right subfigures do not have any particular meaning and serve only as a visual guide to better distinguish the separation of the second sublayer displayed in the feature space determined by the initial layer analysis. Once again the components appear to have high correlation. We can see now that there may be merit in discarding variables thought to be measuring the same underlying (but "latent") aspect of a collection of variables, because including the nearly-redundant variables can cause the PCA to overemphasize their contribution. It should be abundantly clear that setting aside variables

known to be strongly correlated with others can have a substantial effect on the PCA results.

These results are considered extremely positive when reflecting on the amount of genomic information condensed in this figure. We see a proper evaluation of the first layer as the non-hormonal melanoma cases are well-condensed and far from the remaining points. In the second layer the reevaluation of the selected points in a new subspace, Fig. 6.8, allows for the reinterpretation of the amalgam of points that is shown above. The linear recombination of the variables to best suit the inspection of the sublayer allowed the labeling clarification of objects fairly overlapped with one another.

All the melanoma cases were precisely estimated while, considering 1 as the positive label, the second grouping witnessed a 96.5% of recall (the fraction of relevant instances that have been retrieved over the total amount of relevant instances) and 91.3% precision (the fraction of relevant instances among the retrieved instances), summarized in a 87.5% of correct categorizations.

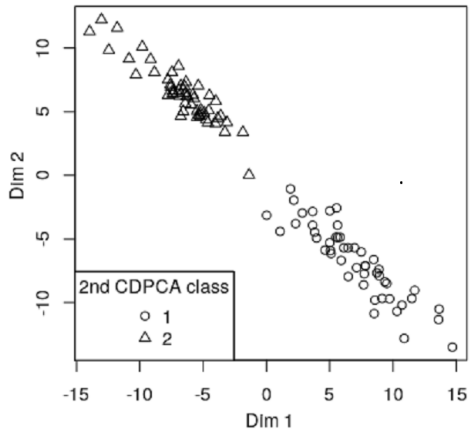


Figure 6.8: Detailed view of the second partition decision boundaries and subspace rearrangement.

Further assessments were made under the same conditions but following a cyclical tandem analysis path. The assessment continuously and stressfully collided against an inability to separate even the first obvious segregation. In the end, a between cluster deviance of 33.8% occurred, with the first two components explaining approximately 29% and 10% of the data variance.

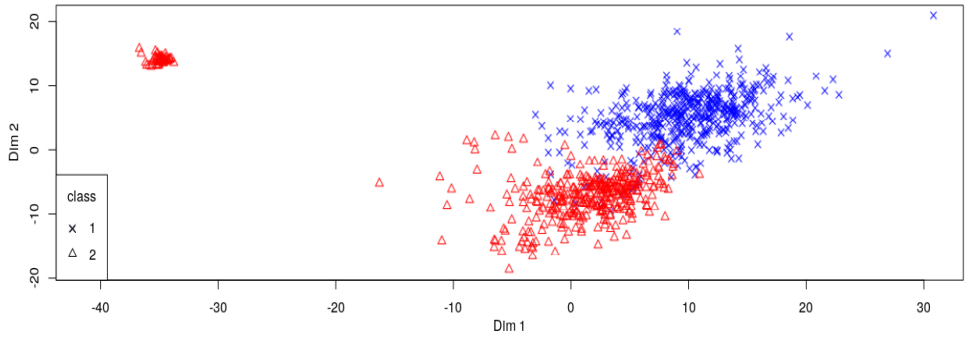


Figure 6.9: Nested tandem analysis fails to separate hormonal from non-hormonal tumors.

7

Conclusions

Contents

Results show that an integrated solution is an effective mechanism to calculate a new linear subspace arrangement of the feature space and to categorize data in the reduced pace. The properties of the control function that models the system and clustering generated can be altered by an appropriate choice of the objective function weights. Results show that the methods can be applied successfully to a wide range of systems with differing characteristics. In higher dimensional data there is space for a judgment call based on the analytical objectives and knowledge of the data to determine possible pre-processing of data prior to running the referred integrated approaches.

Results also show that the relaxed classification of the points is an effective way of combating ambiguity as it retains the variance explained as well as the rigid CDPCA, although it may require a trade-off with the speed of computation. Contrary to what one could expect, the results presented showed that the time consumption related to the heavier computations of a relaxed object attribution matrix was not severe (taking up to 3 more times than the CDPCA execution in our simulations) and does not hamper the validity of the application of the new methodologies.

7.1 Achievements

A fuzzy model is developed and introduced in the remaining CDPCA computations, capturing the objects' dynamics and considering the classification nuances required for a fairer diagnosis of the situation. The experimental data are in agreement with the initial considerations. The accuracy of the model allows for the definition of well stratified groups and a more hassle-free perception of the behavior in the test environment. The maximum-entropy approach of the fuzzy model retains the characteristics of the previously available solution adding flexibility and much needed mathematical features that support the analysis, at the cost of increased algorithmic running time. It offers improved stability over other methods such as FKM which failed to recover the intended subspace in several instances.

A technique is proposed and used to find groupings in a sublayer of the data. The equilibrium achieved with this system has improved variance over the obtained with the tandem solution. The general framework can be extended to benefit any methodology and may be used in the future in a wide set of applications.

7.2 Future Work

During the development of this work, several divergent lines of research or possible analysis were identified. However, given the time limitations for this thesis development, they have not been addressed:

- The dimension-reduction step can be reevaluated and factor analysis can be incorporated if the desired analysis warrants it, i.e. when we wish to assume or to test a theoretical model of latent factors causing the observed variables;
- Other fuzzy techniques can be used to model the clustering structure, which can be particularly useful when the structure appears to be non-spherical, for example through density-based solutions;
- Promote outlier desensitization in the methodology - analyze the introduction of an additional

cluster to combat noise and outlier influence. The cluster is not to be explicitly associated to any prototype, and its center to be at a constant distance delta to all data points;

- CDPCA maximizes the between cluster deviance of the reduced space unconditionally to the within variance. It would be alluring to find the optimal solution by imposing a certain within cluster deviance or by attempting a convex combination of both the between and within cluster deviances when constructing the cost function. Following a similar line of thought, a convex combination of hard and fuzzy c-means may be added to contribute to the objective function;
- A systematic study of the effects of reducing the initial feature pool through the use of neighborhood or correlation based methods;
- A study of the limits of performance, when missing data is present.

Bibliography

[1] Y. Atilgan and F. Dogan. Data mining on distributed medical databases: Recent trends and future directions. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, volume 11, pages 216–224, 2009. doi: 10.1007/978-3-642-03978-2_19.

[2] O. Liu Sheng and H. Garcia. Information Management in Hospitals: An Integrating Approach. In *9th IEEE International Phoenix Conference on Computers and Communications*, pages 296–303, Scottsdale, AZ, USA, 1990.

[3] S. Chakrabarti, M. Ester, U. Fayyad, and J. Gehrke. Data mining curriculum: a proposal. In *ACM SIGKDD*, pages 1–10, 2006.

[4] H. Jiawei et al. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, 2012. ISBN 978-0-12-381479-1. doi: 10.1016/B978-0-12-381479-1.00001-0.

[5] M. K. Obenshain. Application of data mining techniques to healthcare data. *Infection control and hospital epidemiology*, 25(8):690–5, 2004. ISSN 0899-823X. doi: 10.1086/502460.

[6] S. Palaniappan and R. Awang. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS International Conference on Computer Systems and Applications*, pages 108–115, 2008. ISBN 978-1-4244-1967-8. doi: 10.1109/AICCSA.2008.4493524.

[7] R. Shen, A. B. Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England)*, 25(22):2906–12, 2009. doi: 10.1093/bioinformatics/btp543.

[8] S. Zhang, C. C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391, 2012. doi: 10.1093/nar/gks725.

[9] M. Kormaksson, J. G. Booth, M. E. Figueroa, and A. Melnick. Integrative model-based clustering of microarray methylation and expression data. *Annals of Applied Statistics*, 6(3):1327–1347, 2012. doi: 10.1214/11-AOAS533.

[10] E. Keogh and A. Mueen. Curse of Dimensionality. In *Encyclopedia of Machine Learning*, pages 257–258. Springer US, Boston, MA, 2010. doi: 10.1007/978-0-387-30164-8.

[11] P. Pudil and J. Novovičová. Novel methods for subset selection with respect to problem knowledge. *IEEE Intelligent Systems and Their Applications*, 13(2):66–73, 1998. doi: 10.1109/5254.671094.

[12] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, 5, 2015. doi: 10.1038/srep10312.

[13] V. Kumar and S. Minz. Feature Selection : A literature Review. *Smart Computing Review*, 4(3):211–229, 2014. ISSN 22344624. doi: 10.6029/smartcr.2014.03.007.

[14] A. J. Ferreira. *Feature Selection and Discretization for High-Dimensional Data*. PhD dissertation, Instituto Superior Técnico, 2014.

[15] Z. M. Hira and D. F. Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015, 2015. doi: 10.1155/2015/198363.

[16] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324,

1997. doi: http://dx.doi.org/10.1016/S0004-3702(97)00043-X.

[17] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. *Engineering*, pages 74–81, 2001.

[18] S. Subramanian. *Hybrid Feature Selection Methods are the proven methods for Large Scale Feature Selection.* Lap Lambert Academic Publishing, 2011. ISBN 978-3-8443-9924-0.

[19] S. Balakrishnama and a. Ganapathiraju. Linear Discriminant Analysis - a Brief Tutorial. *Compute*, 11:1–9, 1998. doi: http://www.isip.piconepress.com/publications/reports/1998/isip/lda/.

[20] H. Y. Jie, H. Yu, and J. Yang. A direct LDA algorithm for high-dimensional data-with application to face recognition. *Pattern recognition*, 34(10):2067–2070, 2001. ISSN 00313203. doi: doi:10.1016/S0031-3203(00)00162-X.

[21] L. B. Almeida. An introduction to principal components analysis, December 2015.

[22] M. Wall, A. Rechtsteiner, and L. Rocha. Singular value decomposition and principal component analysis. *A Practical Approach to Microarray Data Analysis*, pages 91–109, 2003. ISSN 09240136. doi: 10.1007/0-306-47815-3_5.

[23] W. J. Krzanowski and P. Kline. Cross-Validation for Choosing the Number of Important Components in Principal Component Analysis. *Multivariate Behavioral Research*, 30(2):149–165, 1995. ISSN 15327906. doi: 10.1207/s15327906mbr3002_2.

[24] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014. ISSN 00189448. doi: 10.1109/TIT.2014.2323359.

[25] J. Decoster and G. P. Hall. Overview of Factor Analysis. *In Practice*, 37(2):141, 1998. ISSN 00031305. doi: 10.2307/2685875.

[26] D. Cramer and D. Howitt. Varimax rotation. In *The SAGE Dictionary of statistics*, page 188. SAGE Publications Ltd, 2004. doi: 10.4135/9780857020123.

[27] S. K. Vines. Simple principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(4):441–451, 2000. ISSN 0035-9254. doi: 10.1111/1467-9876.00204.

[28] N. T. Trendafilov and I. T. Jolliffe. Projected gradient approach to the numerical solution of the SCoTLASS. *Computational Statistics and Data Analysis*, 50(1 SPEC. ISS.):242–253, 2006. ISSN 01679473. doi: 10.1016/j.csda.2004.07.017.

[29] H. P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. A general framework for increasing the robustness of PCA-based correlation clustering algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5069 LNCS, pages 418–435, 2008. ISBN 3540694765. doi: 10.1007/978-3-540-69497-7_27.

[30] M. Hubert and S. Engelen. Robust PCA and classification in biosciences. *Bioinformatics*, 20(11):1728–1736, 2004. ISSN 13674803. doi: 10.1093/bioinformatics/bth158.

[31] H. Zou, T. Hastie, and R. Tibshirani. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006. doi: 10.1198/106186006X113430.

[32] Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1097.

[33] H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007. ISSN 00313203. doi: 10.1016/j.patcog.2006.07.009.

[34] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science (New*

York, N.Y.), 290(5500):2323–6, 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2323.

[35] M. Linting, J. J. Meulman, P. J. F. Groenen, and A. J. van der Koojj. Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, 12(3):336–358, 2007. ISSN 1939-1463. doi: 10.1037/1082-989X.12.3.336.

[36] L. J. P. Van Der Maaten, E. O. Postma, J. van den Herik, and H. J. Van Den Herik. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research*, 10(January):1–41, 2009. ISSN 0169328X. doi: 10.1080/13506280444000102.

[37] G. W. Milligan and M. C. Cooper. Methodology Review: Clustering Methods. *Applied Psychological Measurement*, 11(4):329–354, 1987. doi: 10.1177/014662168701100401.

[38] M. Greenacre. Hierarchical cluster analysis. *Correspondence Analysis and Related Methods*, 2:11, 2008.

[39] D. Xu and Y. Tian. A Comprehensive Survey of Clustering Algorithms. *The Annals of Data Science*, 2(2): 165–193, 2015. ISSN 2198-5804. doi: 10.1007/s40745-015-0040-1.

[40] Finding Groups in Data: An Introduction to Cluster Analysis, 2005.

[41] A. Chaturvedi, P. E. Green, and J. D. Caroll. K-modes Clustering. *Journal of Classification*, 18(1):35–55, 2001. ISSN 0176-4268. doi: 10.1007/s00357-001-0004-3.

[42] J. Hastie, Trevor, Tibshirani, Robert, Friedman. *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition*. Springer, New York, NY, 2009. doi: 10.1007/978-0-387-84858-7.

[43] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, and Others. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403 (6769):503–511, 2000. doi: 10.1038/35000501.

[44] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13790–13795, 2001. ISSN 0027-8424. doi: 10.1073/pnas.191502998.

[45] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, 8(8):816–824, 2002.

[46] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, Sept. 1999. ISSN 0360-0300. doi: 10.1145/331499.331504. URL `http://doi.acm.org/10.1145/331499.331504`.

[47] C. C. Aggarwal and C. K. Reddy. *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC, 1st edition, 2013. ISBN 1466558210, 9781466558212.

[48] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002. doi: https://doi.org/10.1198/016214502760047131.

[49] K. Y. Yeung, C. Fraley, A. Murua, a. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics (Oxford, England)*, 17(10):977–87, 2001. ISSN 1367-4803. doi: 11673243.

[50] C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Compu-*

tational Statistics & Data Analysis, 71:52 – 78, 2014. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2012.12.008. URL http://www.sciencedirect.com/science/article/pii/S0167947312004422.

[51] M. Hayes, Y. S. Pyon, and J. Li. A model-based clustering method for genomic structural variant prediction and genotyping using paired-end sequencing data. *Public Library of Science*, 7(12):1–13, 12 2012. doi: 10.1371/journal.pone.0052881. URL https://doi.org/10.1371/journal.pone.0052881.

[52] L. Xuan, F. Yuejiao, W. Xiaogang, and et al. Detecting differentially variable micrornas via model-based clustering. *International Journal of Genomics*, 2018:1–9, 2018. doi: https://doi.org/10.1155/2018/6591634.

[53] R. M. McCloskey and A. Poon. A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. *bioRxiv*, 2017. doi: 10.1101/165357. URL https://www.biorxiv.org/content/early/2017/07/19/165357.

[54] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 3(1):1–58, 2009. ISSN 15564681. doi: 10.1145/1497577.1497578.

[55] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 11(1):5–33, 2005. ISSN 13845810. doi: 10.1007/s10618-005-1396-1.

[56] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004. ISSN 19310145. doi: 10.1145/1007730.1007731.

[57] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park. Fast algorithms for projected clustering. *ACM SIGMOD Record*, 28(2):61–72, 1999. ISSN 01635808. doi: 10.1145/304181.304188.

[58] J. W. Van Ness. Admissible clustering procedures. *Biometrika*, 60:422—-424, 1973.

[59] S. Guenter, H. J. J. M. Bunke, W. Kropatsch, and M. Vento. Validation indices for graph clustering. In *Proc. 3rd IAPR-TC-15 Workshop on Graph-based Representations in Pattern Recognit*, pages 229—-238, 2001.

[60] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, Dec 2001. ISSN 1573-7675. doi: 10.1023/A:1012801612483. URL https://doi.org/10.1023/A:1012801612483.

[61] J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104, 1974. ISSN 00220280. doi: 10.1080/01969727408546059.

[62] N. Bolshakova and F. Azuaje. Cluster validation techniques for genome expression data. *Signal Processing.*, 83:825—-833, 2003.

[63] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. Dougherty. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40:807—-824, 2006.

[64] V. Roth, M. Braun, T. Lange, and J. Buhmann. Stability-based model order selection in clustering with applications to gene expression data. *Lecture Notes in Computer Science*, 2415:607–612, 2002.

[65] L. Dalton, V. Ballarin, and M. Brun. Clustering algorithms: On learning, validation, performance, and applications to genomics. *Current Genomics*, 10:430–445, 2009. doi: {10.2174/138920209789177601}.

[66] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Admissible clustering procedures. *J. Intell. Inf. Syst.*, 17:107—-145, 2001.

[67] M. Vichi and H. A. L. Kiers. Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 2001. doi: 10.1016/S0167-9473(00)00064-5.

[68] P. Arabie and L. Hubert. Cluster analysis in marketing research. In *Handbook of marketing research*, 1994.

[69] W. Desarbo, K. Jedidi, K. Cool, and D. Schendel. Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters*, 1991. doi: 10.1007/BF00436033.

[70] G. De Soete and J. D. Carroll. k-means clustering in a low-dimensional Euclidean space. In *New Approaches in Classification and Data Analysis*, Springer, Heidelberg, pp. 212–219. 1994.

[71] A. D. Gordon. Classification. *Monographs on statistics and applied probability*, 1999.

[72] M. E. Timmerman, E. Ceulemans, H. A. Kiers, and M. Vichi. Factorial and reduced K-means reconsidered. *Computational Statistics and Data Analysis*, 2010. doi: 10.1016/j.csda.2010.02.009.

[73] J. Mezzich and H. Solomon. Taxonomy and behavioral science. In *Quantitative studies in social relations*, Academic Press, London. 1980. ISBN 9780124933408.

[74] A. P. Association. Diagnostic and statistical manual of mental disorders, Washington, DC. 1968.

[75] M. Vichi and G. Saporta. Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, 53(8):3194–3208, 2009. ISSN 01679473. doi: 10.1016/j.csda.2008.05.028.

[76] E. Macedo and A. Freitas. The alternating least-squares algorithm for CDPCA. In *Communications in Computer and Information Science*, volume 499, pages 173–191, 2015. ISBN 9783319203515. doi: 10.1007/978-3-319-20352-2_12.

[77] J. Khan et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 2001. doi: 10.1038/89044.

[78] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification. *Pattern Analysis and Applications*, 1998. doi: 10.1007/BF01237942.

[79] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 2000. doi: 10.1080/00401706.2000.10485979.

[80] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 1973. doi: 10.1080/01969727308546046.

[81] J. C. Bezdek. *Fuzzy Mathematics in Pattern Classification*. PhD dissertation, Applied Math Center, Ithaca, USA, 1973.

[82] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 2000. ISSN 09603174. doi: 10.1023/A:1008940618127.

[83] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948. doi: 10.1145/584091.584093.

[84] T. R. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 1999. doi: 10.1126/science.286.5439.531.

A

Fuzzy C-Means Functional Optimization

This appendix provides the proof of the Lagrangian multiplier solution to the fuzzy c-means objective function.

Proof. Fix $\bar{\mathbf{y}} \in \mathbb{R}^{cp}$ and define $g_m(\mathbf{U}) = J_f(\mathbf{U}, \bar{\mathbf{y}})$ for any $\mathbf{U} \in M_{fco}$. Since $\mathbf{U}$ is degenerate - its columns are independent - and therefore

$$\begin{aligned}\min_{\mathbf{U}\in M_{fco}} &= \min_{\mathbf{U}\in M_{fco}} \left\{ \sum_{i=1}^{I}\sum_{p=1}^{c}(u_{ip})^m(d_{ip})^2 \right\} \\ &= \sum_{i=1}^{I}\left[\min_{\mathbf{u_i}\in conv(B_c)} \left\{ \sum_{p=1}^{c}(u_{ip})^m(d_{ip})^2 \right\}\right] \end{aligned} \tag{A.1}$$

where

$$conv(B_c) = \left\{ \mathbf{u_i} \in \mathbb{R}^c \middle| \sum_{p=1}^{c} u_{ip} = 1; u_{ip} \geq 0 \right\}.$$

The solution of (A.1) is achieved with Lagrange multipliers. For each term, let

$$g(\mathbf{u_i}) = \sum_{p=1}^{c}(u_{ip})^m(d_{ip})^2$$

and let its Lagrangian be

$$F_i(\lambda, \mathbf{u_i}) = \sum_{p=1}^{c}(u_{ip})^m(d_{ip})^2 - \lambda\left(\sum_{p=1}^{c} u_{ip} - 1 \right),$$

where $(\lambda, \mathbf{u_i})$ is stationary for F_i only if $\nabla_{\lambda,\mathbf{u_i})}F_i(\lambda, \mathbf{u_i}) = (0, \theta \in \mathbb{R}^c)$. Setting this gradient equal to zero leads to

$$\frac{\partial F_i}{\partial \lambda}(\lambda, \mathbf{u_i}) = \sum_{p=1}^{c} u_{ip} - 1 = 0, \tag{A.2a}$$

$$\frac{\partial F_i}{\partial u_{ts}}(\lambda, \mathbf{u_i}) = \left[m(u_{ts})^{m-1}(d_{ts})^2 - \lambda \right] = 0. \tag{A.2b}$$

From this,

$$u_{ts} = \left[\frac{\lambda}{m(d_{ts})^2} \right]^{1/(m-1)}. \tag{A.3}$$

Using (A.2a),

$$\begin{aligned}\sum_{l=1}^{c} u_{tl} &= \sum_{l=1}^{c} \left(\frac{\lambda}{m} \right)^{1/(m-1)} \left[\frac{1}{(d_{tl})^2} \right]^{1/(m-1)} \\ &= \left(\frac{\lambda}{m} \right)^{1/(m-1)} \left\{ \sum_{l=1}^{c} \left[\frac{1}{(d_{tl})^2} \right]^{1/(m-1)} \right\} = 1. \end{aligned}$$

Thus,

$$\left(\frac{\lambda}{m}\right)^{1/(m-1)} = \frac{1}{\sum_{l=1}^{c}\left(\frac{1}{(d_{tl})^2}\right)^{1/(m-1)}}.$$

Returning to (A.3),

$$\begin{aligned} u_{ts} &= \frac{1}{\sum_{l=1}^{c}\left[\frac{1}{(d_{tl})^2}\right]^{1/(m-1)}}\left[\frac{1}{(d_{ts})^2}\right]^{1/(m-1)} \\ &= \frac{1}{\sum_{l=1}^{c}\left(\frac{d_{ts}}{d_{tl}}\right)^{2/(m-1)}}. \end{aligned}$$

B

Regularization Approach to Fuzzy C-Means Functional

This appendix provides the proof of the Lagrangian multiplier solution to the fuzzy c-means objective function with entropy regularization.

We derive the update rule for the membership degrees in the usual way (cf. the preceding appendix) by incorporating them with Lagrange multipliers into the objective function. The resulting Lagrange function is

$$F_i(\lambda, \mathbf{u_i}) = \sum_{p=1}^{c} u_{ip}(d_{ip})^2 + \gamma \sum_{p=1}^{c} f(u_{ip}) - \lambda\Bigg(\sum_{p=1}^{c} u_{ip} - 1\Bigg),$$

where λ are the Lagrange multipliers, one per constraint. Since a necessary condition for a minimum of the Lagrange function is that the partial derivatives w.r.t. the membership degrees vanish, we obtain

$$\frac{\partial F_i}{\partial \lambda}(\lambda, \mathbf{u_i}) = \sum_{p=1}^{c} u_{ip} - 1 = 0, \tag{B.1a}$$

$$\frac{\partial F_i}{\partial u_{ts}}(\lambda, \mathbf{u_i}) = (d_{ts})^2 - \lambda + f^{'}(u_{ts}) = 0. \tag{B.1b}$$

Developing (B.1b) we arrive at

$$u_{ts} = f^{'-1}\Bigg(\frac{\lambda - d_{ts}^2}{\gamma}\Bigg), \tag{B.2}$$

where $'$ denotes taking the derivative w.r.t. the argument of the function and $f^{'-1}$ denotes the inverse of the derivative of the function f.

In analogy to Appendix A, constraints on the membership degrees are now exploited to obtain

$$\sum_{p=1}^{c} u_{ip} = \sum_{s=1}^{c} f^{'-1}\Bigg(\frac{\lambda - d_{ts}^2}{\gamma}\Bigg) = 1.$$

This equation has to be solved for λ and the result has to be used to substitute λ in (B.2). In order to do so, we introduce the exact form of the regularization function f. With $f(u_{ts}) = u_{ts} \ln u_{ts}$ we have that $f^{'}(u_{ts}) = 1 + \ln u_{ts}$, and therefore $F^{'-1}(y) = e^{y-1}$. Using the latter in the formulas obtained above for deriving the update rule for the membership degrees yields

$$u_{ts} = \frac{e^{-\frac{d_{ts}^2}{2\sigma^2}}}{\sum_{j=1}^{c} e^{-\frac{d_{tj}^2}{2\sigma^2}}}. \tag{B.3}$$

It should be noted that $f^{'}(u_{ts}) = 1 + \ln u_{ts}$ implies $f^{'}(1) - f^{'}(0) = \infty$; Despite the fact that Shannon entropy regularization always yields graded assignments this drawback is less harmful here because $e^{-\frac{d_{ts}^2}{2\sigma^2}}$ is much "steeper" than $(d_{ts})^2$ hence being less prone to producing undesired results.