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

Luís Rei
Instituto Superior Técnico, Lisboa, Portugal.

November 2018

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.

1. Introduction

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. 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 [3]. 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 patient's 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.

2. Beyond Tandem Analysis

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. This kind of approach was firstly named "tandem analysis" by Arabie and Hubert and was already disputed by De Sarbo et al. [4] 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.

2.1. 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. 1, 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.

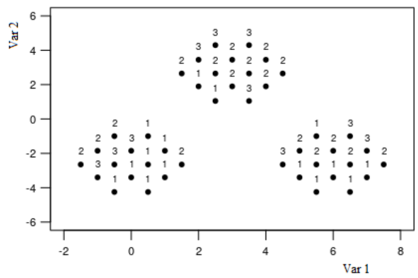


Figure 1: K-means classification of 42 objects described by six variables, two of which determine the location of the points in the plot (three classes) and the other four variables are randomly generated by normal distribution.

The 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 Fig. 2, 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 2b, 2c, and 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 1. In Fig. 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. 2d.

In our example, the substitution of the standardized scores by those multiplied by the square roots of the eigenvalues to avoid the case of distorted mutual distances 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 the sourced in a of bad space representation, but instead it originates from the simple fact that the principal components are not upbringing the best achievable agglomerations.

Table 1: Explained total variance and cumulative variance.

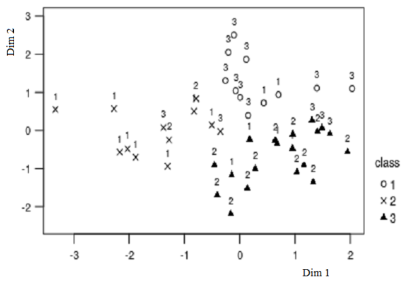
Components	Eig	% variance	% cumulative
1	1.23	25.33	25.33
2	1.03	17.52	42.85
3	1.00	16.78	59.63
4	0.98	16.15	75.78
5	0.89	13.21	88.99
6	0.81	11.01	100.00

2.2. Factorial and Reduced K-means

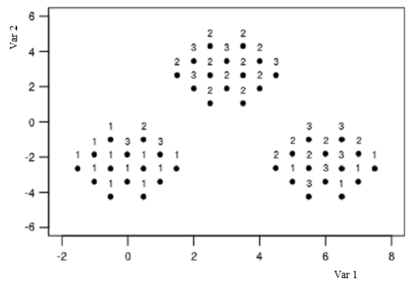
Before advancing with any comparison to a sequential tandem approach we explore the theoretical intricacies of two integrated methods: Reduced K-means (RKM) proposed by De Soete [5]; and Factorial K-Means [6].

2.2.1 Optimization of the Loss Functions

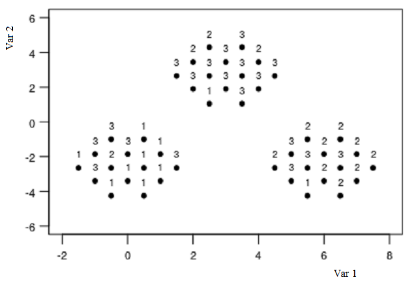
Both integrated approaches aim at identifying the best partition of objects described by the best orthogonal linear combination of variables following the least squares criterion. A dual objective is attempted to be achieved: optimal data synthesis of objects and attributes occur simultaneous to variable 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



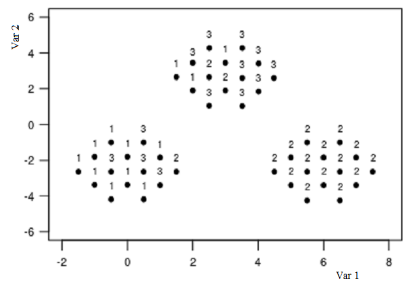
(a) First two principal components



(b) First three principal components



(c) First four principal components



(d) First five principal components

Figure 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. 1.

orthonormal components' score matrix $\mathbf{A}$ that expresses the loadings of the variables, and a cluster centroid score matrix $\bar{\mathbf{Y}}$. 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

$$F_{RKM}(\mathbf{U}, \mathbf{A}, \bar{\mathbf{Y}}) = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}||^2, \tag{2.1}$$

and the model that is fitted by it is

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}_R, \tag{2.2}$$

where $\mathbf{E}_R$ is an $(I \times J)$ residual matrix. Whereas the FKM minimizes the loss function

$$\begin{aligned} F_{FKM}(\mathbf{U}, \mathbf{A}, \bar{\mathbf{Y}}) &= ||\mathbf{X}\mathbf{A}\mathbf{A}^{\mathrm{T}} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}||^2 \\ &= ||\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{Y}}||^2. \end{aligned} \tag{2.3}$$

and the model that is fitted by it is

$$\mathbf{X}\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}_F, \tag{2.4}$$

where $\mathbf{E}_F$ is also an $(I \times J)$ residual matrix.

When the centroids are located in the full space ($Q = J$) the methods trace back to the original k-means algorithm. As is observable from (2.1), 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 (2.3) 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.

2.2.2 Finding the Ideal Data

From (2.3) it follows that the FKM's loss function is null if and only if $\mathbf{X}\mathbf{A} = \mathbf{U}\bar{\mathbf{Y}}$. We can consider the data matrix $\mathbf{X}$ expressed in terms of the loadings matrix $\mathbf{A}$ and $\mathbf{A}^{\perp}$, as

$$\mathbf{X} = \mathbf{B}\mathbf{A}^{\mathrm{T}} + \mathbf{C}\mathbf{A}^{\perp\mathrm{T}}. \tag{2.5}$$

Replacing (2.5) in $\mathbf{X}\mathbf{A} = \mathbf{U}\bar{\mathbf{Y}}$ we find $\mathbf{X}\mathbf{A} = \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{C}\mathbf{A}^{\perp\mathrm{T}}\mathbf{A} = \mathbf{B} = \mathbf{U}\bar{\mathbf{Y}}$. Returning to the ideal FKM data we have $\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{C}\mathbf{A}^{\perp\mathrm{T}}$, and evoking the general matrix $\mathbf{C}$ by $\mathbf{E}^{\perp}$, the full class is

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}^{\perp}\mathbf{A}^{\perp\mathrm{T}}. \tag{2.6}$$

Equation (2.4) 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.

From (2.1) 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 (2.2) illustrates the ideal RKM data 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).

2.3. Integrated Approach Application

Following the discussion of RKM and FKM, performance is assessed by applying the methodology to a simulated data already analyzed with tandem analysis. 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, the centroid specified to be located in a two-dimensional subspace. The result exposed in Fig. 3 reveals structure recovery and data classification failure of RKM.

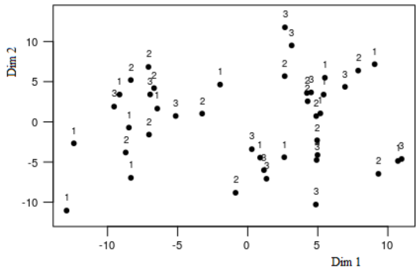


Figure 3: Classification of 42 objects in a low-dimensional space represented by the first two dimensions of the reduced k-means analysis.

Turning to the FKM procedure, Table 2 introduces the correlations between the model defined factors and the six original features. Figure 2 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 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

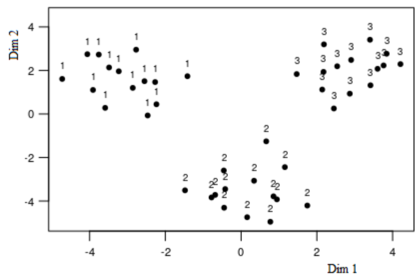


Figure 4: Classification of 42 objects in a low-dimensional space represented on the first two dimensions of the factorial k-means analysis.

3. Clustering and Disjoint Principal Component Analysis

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 or 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 [7], 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 to a data set of small round blue cell tumors.

3.1. 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 [7]. 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

$$\begin{aligned} \mathbf{X} &= \mathbf{U}\bar{\mathbf{X}} + \mathbf{E}_1 && \text{(K-means step on } \mathbf{X}\text{)} \quad (3.1a)\\ &= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E}_1 + \mathbf{E}_2 && \text{(PCA step on } \mathbf{U}\bar{\mathbf{X}}\text{)} \quad (3.1b)\\ &= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}} + \mathbf{E} && (3.1c) \end{aligned}$$

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. (3.1c) 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 $\min_{\mathbf{U},\bar{\mathbf{Y}},\mathbf{A}} \|\mathbf{X}-\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2$, 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 to the equivalent maximization of the between cluster deviance in the reduced space follows as $\|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2 = \|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\|^2$ and $\|\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2 = \|\mathbf{X}\|^2$ - $\|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}\|^2$ [7].

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 [8] given by the total variance of the data in the reduced space $\max F = \|\mathbf{U}\bar{\mathbf{Y}}\|^2$, and constrained to the allocation of I objects and P clusters, to the allocation of J variables into Q disjoint components and constrained to the PCA implementation. The maximum dissimilarity of centroids is represented by the maximization of the objective function.

To solve the problem an iterative algorithm called alternating least-squares algorithm (ALS) is proposed in [7]. Figure 5 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 F_k and F_{k+1} is smaller than a specified tolerance threshold.

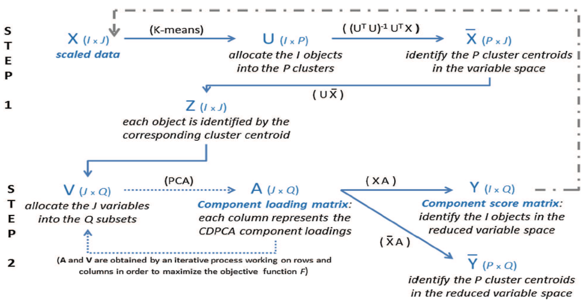


Figure 5: The two basic steps of one iteration of the Alternating Least-Squares algorithm for performing CDPCA (extracted from [8]).

Since F is bounded above the output of each iteration will converge to a stationary point - at least a local maximum of the problem [7]; 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.

3.2. Algorithms

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

3.2.1 Initialization

At the beginning, the data matrix $\mathbf{X}$ is standardized and the parameters to perform CDPCA are initialized. 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 $\mathbf{V}_0$.

3.2.2 General Iteration

After performing the initialization steps, at the beginning of the $(k+1)$-th iteration, 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. 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 not then half the objects on the biggest cluster are assigned into one of the empty clusters.

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. 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 $F_{k+1} = \mathrm{tr}((\mathbf{U}_{k+1}\bar{\mathbf{Y}}_{k+1})^T\mathbf{U}_{k+1}\bar{\mathbf{Y}}_{k+1}$.

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.

3.3. Empirical Examples

The clustering and disjoint PCA has been applied to a real data set describing Small Round Blue Cell Tumors (SRBCT) of a childhood cancer study by Khan et al. [9] and takes into account microarray experiments to show the performances of the methodology.

The gene expression data[1] 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.

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 and classifying patients on the basis of first 21 component scores. The results are shown in Fig. 6. The k-means algorithm was run on the first two PCA starting from random partitions. It was necessary to run it for a large number of initial random starts for the presence of several local optima (optimal 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 solution equalled to 28.8% of total deviance and 23 correct predictions.

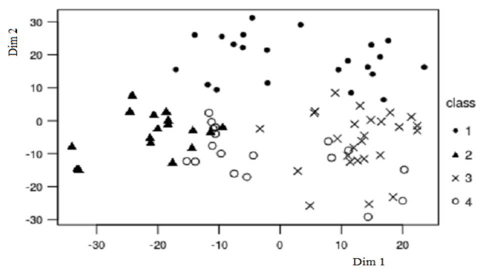


Figure 6: Tandem analysis results on the SRBCT dataset. Overlapping clusters make interpretation more difficult or impossible.

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 results of the CDPCA are reported in Fig. 7. The optimal solution was found 2 times in the 250 runs, the algorithm converging after between 9 iterations (tolerance of 10^{-3}). The two components of the CDPCA explain double the variance of PCA, albeit presenting a low value nonetheless (9% and 7%, respectively).

Despite the fact that the components of CDPCA possess rather consistent low scores like what was witnessed in the PCA components, the disjoint PCA

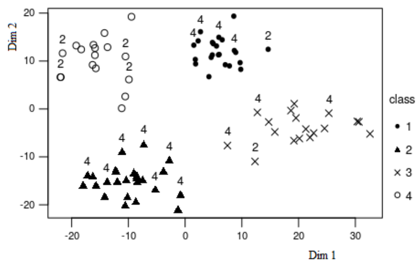


Figure 7: Clustering and disjoint PCA results on the SRBCT dataset. Numbers above points showcase themissclassifications of even classes, the ones that casted the most doubts in determining the corresponding cluster.

clearly shows more homogeneous clusters (between cluster deviance of CDPCA is 83% of total deviance.

In this difficult dataset the classification on the two dimensions defined by the CDPCA model is similar to the one of the tandem analysis with 36 correct predictions. The 43.4% accuracy translates a less than optimal subspace recovery and can 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.

4. Integrated Methods for Increase in Data Interpretability

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 method developed is named Relaxed Clustering and Disjoint Principal Component Analysis, or RCDPCA, stemming from the inclusion of a fuzzy model in the object assignment phase of CDPCA. The second method is denominated Nested Clustering and Disjoint Principal Component Analysis, or 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.

4.1. 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. It tries to circumvent hiccups of the k-means discussed in the previous section by attempting to overcome poor performance on badly delineated or overlapping groups, or on symmetric data. The algorithm maximizes its objective function by virtue of a greedy search and retains the properties of the work introduced in Section 3.

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

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

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.

4.1.1 Understanding Soft Clustering

Soft or fuzzy clustering allows gradual memberships of data points to clusters, 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. Constraints guarantee that no cluster is empty, and 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 partitions exhaustive). The combination of the two conditions imply that no cluster can contain full membership of all data points and that membership degrees for a datum resemble probabilities of being member of its own cluster.

The fuzzy clustering criterion we discuss generalizes the within groups sum of square errors function initially reported by Dunn in [10] as an algorithm akin to hard c-means,

$$J_f(\mathbf{U}, \bar{\mathbf{y}}) = \sum_{i=1}^{I}\sum_{p=1}^{c}(u_{ip})^m (d_{ip})^2, \tag{4.1}$$

where $(d_{ip})^2 = ||\mathbf{x}_i - \bar{\mathbf{y}}_p||^2$, and weighting exponent m belongs in $[1, \infty[$. Because the terms of J_f are proportional to $(d_{ip})^2$, it is a squared error clustering criterion, and its least-squared error stationary points are solutions of $\min_{M_{fc}\times\mathbb{R}^{cp}}\{J_f(\mathbf{U}, \bar{\mathbf{y}})\}$. 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 fuzzy c-means 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.

4.1.2 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. What is the physical meaning of the parameter m? 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" [11] 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. Defining the local loss function as the within-group sum of squared error the algorithm is set to minimize the objective function

$$\sum_{i=1}^{I}\sum_{p=1}^{c} u_{ip}(d_{ip})^2 + \gamma\sum_{i=1}^{I}\sum_{p=1}^{c} u_{ip}\ln u_{ip}, \tag{4.2}$$

where u_{ip} denotes the grade of membership of the i-th data pairs in the p-th cluster.

The functional (4.2) 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.

To maximize MEI the alternating minimization procedure between membership matrix $\mathbf{U}$ and cluster center matrix $\bar{\mathbf{y}}$ can be applied to (4.2), resulting in the solutions

$$u_{ip} = \frac{e^{-\frac{d_{ip}^2}{2\sigma^2}}}{\sum_{j=1}^{c} e^{-\frac{d_{ij}^2}{2\sigma^2}}}, \quad \bar{\mathbf{y}}_p = \frac{\sum_{i=1}^{I} u_{ip}\mathbf{x}_i}{\sum_{i=1}^{I} u_{ip}} \qquad \forall\, i, p,$$

where σ is the Lagrangian multiplier from the loss function.

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 k-means).

4.2. Nested CDPCA

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. 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. In Fig. 8 a visual display of the cycle focusing on two layers optimized by the same F is provided. Firstly, we compute the first layer partitions and sequentially advance and examine each individual cluster originated in the first phase of the program.

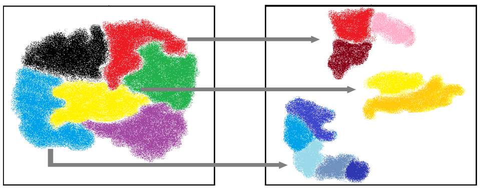


Figure 8: Illustrative graphical display of the NCDPCA algorithm. In this particular case, the first partition reveals six clusters represented by different colors. In the second phase of the nested process three of the clusters are reevaluated and further divisions can be retrieved for analysis.

5. Results and Discussion

Cancer classification is intimately connected to cancer treatment and enhancements in this area contribute significantly to advances in patient recovery processes. The main challenges of cancer treatment were always 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 appearances to different clinical courses with different responses to therapy is extremely limiting. For more tumors 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 section 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.

5.1. Leukemia Data

This clinical dataset contains gene expression data from the leukemia microarray study accessible from the R package *multtest*. The data contained 38 cases of human 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 22 component scores addressing 82% of the total variance. The results are shown in Fig. 9. The k-means algorithm was performed on the first two PCA starting from random partitions 500 times and found the present optimal solution after 409 runs. The first two components explain 15.6% and 9.4% of the total variance. The between cluster deviance of the optimal solution was equal to 13.6% of the total deviance.

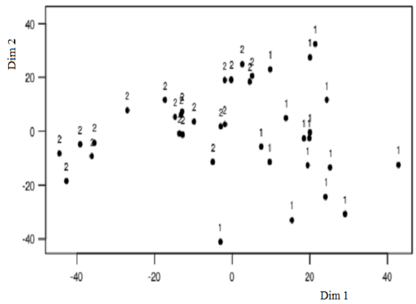


Figure 9: Tandem analysis. Classification of leukemia patients represented on the first two principal components.

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 (32 wrong labels).

Without knowledge on the conditions of the data, and the disposition of the residuals, it is wise to attempt an multitude of angles and methods to extract the widest range of pertinent information. The integrated approaches studied in Section 2 are revisited in this experiment with a distinct parameter Q of 10 for RKM and FKM. The between cluster deviance of the first is 78.4% and its output's sensitivity increase led to a jump to 31.6% of good labeling (26 wrong decisions); however, it's visualization was not much superior to the tandem approach and its image was omitted. In the case of FKM, Fig. 10, a subspace was found such that the data points were aligned and perfectly stratified into two clusters (error function reached its lower limit and 100% within cluster deviance was observed) product of an 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 an inflexible model to further object introductions.

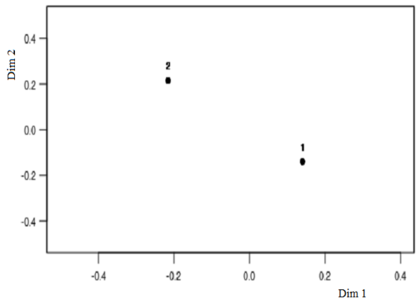


Figure 10: FKM clustering of leukemia patients in low dimensional space.

The results of the Relaxed CDPCA are reported in Fig. 11. The two components of the relaxed clustering and disjoint PCA explain a low variance of 5.2% and 5.0%, respectively. Despite the fact that the components

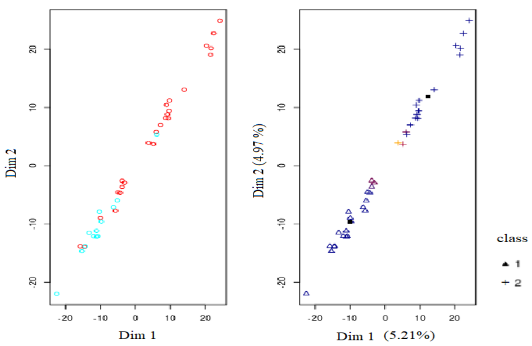


Figure 11: Fuzzy model incorporated in CDPCA applied to leukemia data. (left) Real classification with the blue color symbolizing the first class and red the second; (right) RCDPCA classification with brighter colors corresponding to less certain affirmations of class memberships.

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 79% of total deviance. The classification on the two dimensions defined by the model is a step above of the other ones with 28 correct predictions. The 73.7% accuracy translates a less than optimal subspace recovery and can be explained by the deficient definition of Cluster 1 and the proximity of label 2's.

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.

5.2. 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, Fig. 12.

RCDPCA took 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 number of missclassifications in Section 3. 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.

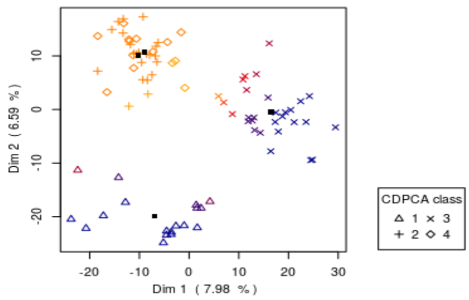


Figure 12: Relaxed clustering and disjoint PCA results on the SRBCT dataset.

An application of a tandem PCA + fuzzy c-means was performed but ommited as it translated all the issues discussed in the previous sections. Even with adjusted "fuzziness" parameters the doubts cast by this test are too much to bare and the system evaluation was dimmed unusable.

5.3. 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. Three initial sets of collected cancer data were united with 1204 and 547 cases of breast and prostate cancer over 19660 genes, and 84 instances of melanoma cases over 52746 genomes (after scaling and intercepting common genes a (1835 x 19435) matrix was left).

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.

In Fig. 13, the NCDPCA achieved a between cluster deviance of 92.1% in an average of 4 iterations for the first stratification of data and 86.0% after 6 iterations in the following sublayer analysis, explaining almost 20% of the total variance with 2 components. The

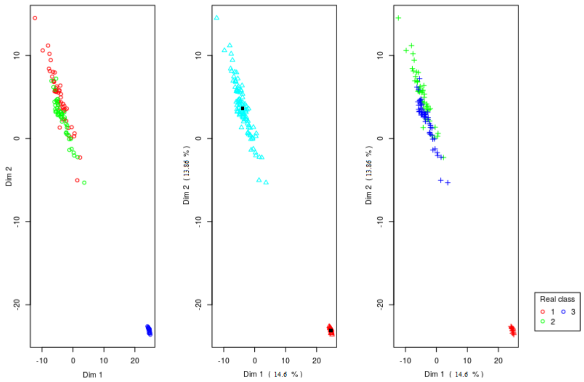


Figure 13: Nested clustering and disjoint PCA results on the TCGA custom cancer dataset. (left) Real classification; (middle) first layer partition; (right) and second layer partition represented in the computed feature space. Note that not all objects are drawn (removed to facilitate the perception of the clusters' overlap).

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 initial layer's feature space. These results are 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. The second layer allows for the reinterpretation of the amalgam 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 (fraction of relevant instances that have been retrieved over the total amount of relevant instances) and 91.3% precision (fraction of relevant instances among the retrieved instances), summarized in a 87.5% of correct predictions.

In Fig. 14 we studied the same data obtained 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 29% and 10% of the variance.

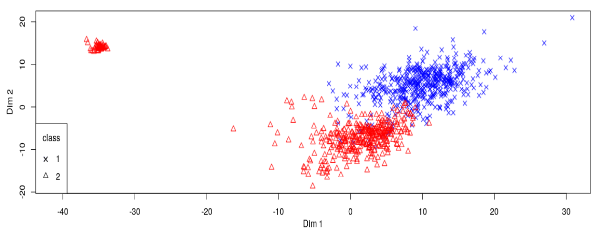


Figure 14: Nested tandem analysis fails to separate hormonal from non-hormonal tumors.

6. Conclusions

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 a reduced pace. We proposed the incorporation of a fuzzy model in CDPCA, 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. We also proposed a technique used to find groupings in a sublayer of the data. The solution achieved with this system is improved over the obtained with the tandem analysis. The general framework can be extended to benefit any methodology and may be used in a wide set of systems with differing characteristics.

Acknowledgements

The author gratefully acknowledges the support of the Portuguese Foundation of Science and Technology (FCT) that allowed the thesis to be performed in the framework of project PTDC/EMS-SIS/0642/2014.

References

[1] Y. Atilgan and F. Dogan, "Data mining on distributed medical databases: Recent trends and future directions," in *Inst. for Comp. Sci., Social-Informatics and Telecom. Eng.*, vol. 11, pp. 216–224, 2009.

[2] O. Liu Sheng and H. Garcia, " Information Management in Hospitals: An Integrating Approach," in *9th IEEE Int. Phoenix Conf. on Comp. and Comms.*, (Scottsdale, AZ, USA), pp. 296–303, 1990.

[3] H. Jiawei et al., *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, 2012.

[4] W. Desarbo et al., "Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups," *Marketing Letters*, 1991.

[5] 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.

[6] M. Vichi and H. A. L. Kiers, "Factorial k-means analysis for two-way data," *Computational Statistics and Data Analysis*, 2001.

[7] M. Vichi and G. Saporta, "Clustering and disjoint principal component analysis," *Comp. Stat. and Data Analysis*, vol. 53, no. 8, pp. 3194–3208, 2009.

[8] E. Macedo and A. Freitas, "The alternating least-squares algorithm for CDPCA," in *Comm. in Comp. and Inf. Sci.*, vol. 499, pp. 173–191, 2015.

[9] J. Khan et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat. Med.*, 2001.

[10] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, 1973.

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, 1948.