UNIVERSITY OF THESSALY

Faculty of Medicine

«Research Methodology in Biomedicine, Biostatistics and Clinical Bioinformatics at University of Thessaly»

Postgraduate Dissertation

Comparison of multivariate methods in group/cluster identification

Σύγκριση πολυμεταβλητών μεθόδων στην αναγνώριση ομάδων / συμπλεγμάτων

Scientific Committee:

Apostolos Batsidis, *Assistant Professor, Department of Mathematics, University of Ioannina* (Supervisor)

Ioannis Stefanidis, *MD, PhD, Professor of Internal Medicine/Nephrology, Faculty of Medicine, University of Thessaly*

Chrysoula Doxani, *MSc, MD, PhD, Research Fellow in Genetic Pharmacoepidemiology, University of Thessaly*

Sofia D. Anastasiadou

2018

Acknowledgements

I express my deep and profound gratitude to my supervisor Dr. Apostolos Batsidis, Assistant Professor of Statistics, Department of Mathematics, University of Ioannina, for his support, guidance and constructive feedback, in completing this demanding but meaningful learning journey.

I would also like to thank the other two members of my committee Dr. Ioannis Stefanidis, Professor of Internal Medicine/Nephrology, Faculty of Medicine, University of Thessaly and Mrs Chrysoula Doxani, Ph.D, Research Fellow in Genetic Pharmacoepidemiology, University of Thessaly for their valuable advices and assistance.

I am highly grateful to the distinguished Professor, Dr. Elias Zintzaras, Head of the Programme of Postgraduate Studies (MSc) «Research Methodology in Biomedicine, Biostatistics and Clinical Bioinformatics», Professor of Biomathematics-Biometry, Faculty of Medicine, University of Thessaly, who has strongly promoted research excellence and offered many opportunities for advanced learning to all students through this scientific, innovative and exciting programme.

I also want to thank Mr. Theodoros Mprotsis, Teaching and Research Fellow, Laboratory of Biomathematics, who has provided teaching and technical support.

Finally, I want to thank Mrs. Aikaterini Kalogianni, Head of the Secretariat, Laboratory of Biomathematics, for her politeness, and professionalism, especially for the provision of valuable information concerning the processes of the programme, as well as her willingness to help all the students to solve any problems.

Devoted to my brother…

Contents

Abstract 1

Περίληψη 2

1. Introduction 3

2. Data Clustering Techiques 3

2.1. Principal Components Analysis *3*

2.2. Discriminant Analysis *6*

2.3. K-Means *8*

2.4. Biochemical Analyses *11*

2*.5. Participants* *13*

2.6. Limitations *13*

3. Results *13*

3.1. Principal Components Analysis (PCA) results *13*

3.2. *Discriminant Analysis results* *18*

3.3. *Discriminant Analysis results for 3 predictors related to kidney function* *22*

3.4. *K-Means Results* *26*

4. Conclusions 28

5. References *29*

COMPARISON OF MULTIVARIATE METHODS IN GROUP/CLUSTER IDENTIFICATION

Abstract

Introduction: Even though there is a substantial development and utilization of pattering methods in medicine, a direct comparison of multivariate methods in group/cluster identification for biomarkers has not been carried out.

Objective: This Msc Thesis analyses three different statistical techniques: i.e the Principal Components Analysis (PCA), the Discriminant Analysis and the K-Means clustering. The main objective is to compare patterns derived from Principal Components Analysis (PCA), Discriminant Analysis and K-Means procedures with respect to biochemical measurements.

Design: The study included 303 patients, 151 cases and 152 controls. The 151 patients (cases) were diagnosed as suffering from kidney disease. Concentrations of AST (SGOT), ALT (SGPT), Glucose Serum, Urea, Creatinine, Serum Uric Acid, Serum Calcium, Potassium Serum, Sodium Serum, Total Albumins (TP), Albumin, Alp, γ-GT, CRP, LDH and CPK were measured.

Methods: The Msc Thesis focuses on the presentation of the three main types of clustering methods, Principal Components Analysis (PCA), Discriminant Analysis and K-Means.

Results: PCA's results showed the existence of 5 Components, amongst which the third is shown to be the Component for renal function. This Component comprises of variables: Urea, Creatinine and Serum Uric Acid, which are also the variables which are clinically measured to determine the existence or not of kidney disease. From the scatter plots for all combinations of Components, it was established that the Component for renal function was indeed the one with respect to which controls differentiated from cases. Discriminant Analysis was applied twice. It was initially applied on all 16 variables measuring the concentrations in the participants' biochemical analyses and showed that Urea is indeed the best predictor, followed by Creatinine and then Serum Uric Acid, all with respect to separating controls from cases. The accuracy of Predicted Group Membership was verified. Moreover, analysis exhibits high sensitivity and high specificity. It was then applied only for aforementioned three variables and showed that they are, indeed, the appropriate predictors for the separation of the two groups, controls from cases. More specifically, Creatinine was shown to be the best predictor, followed by Urea and Serum Uric Acid, with respect to the separation of controls from cases. Predicted Group Membership accuracy was verified in this analysis as well, as were the high sensitivity and high specificity of the data.

K-Means was applied only on these three variables and showed that Urea predictor, Creatinine and Serum Uric Acid predictors can satisfactorily separate controls from cases.

Conclusion: The goal of the Msc Thesis was to compare 3 types of clustering techniques and Principal Components Analysis (PCA), Discriminant Analysis and K-Means, three statistically different procedures, on real data with respect to concentration measurements of biochemical analyses indexes.

Results were shown to be comparable in relation to plasma biomarkers and kidney disease. In addition, they showed that the three methods operate complementary, each one accentuating a different dimension for the interpretation of data, the interpretation of which would not have been determinative without the import of clinical doctors and medicine.

Key words: Principal Components Analysis, Discriminant Analysis, K-Means, Clustering

Σύγκριση πολυμεταβλητών μεθόδων στην αναγνώριση ομάδων / συμπλεγμάτων

Περίληψη

Εισαγωγή: Παρά το γεγονός ότι υπάρχει σημαντική ανάπτυξη και αξιοποίηση πολυμεταβλητών μεθόδων στην αναγνώριση ομάδων / συμπλεγμάτων στην Ιατρική, δεν έχει πραγματοποιηθεί σύγκριση αυτών σε ό,τι αφορά σε βιοδείκτες.

Στόχοι: Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία αναλύει και συγκρίνει τρεις πολυμεταβλητές μεθόδους αναγνώρισης ομάδων / συμπλεγμάτων, όπως είναι η K-Means, η Διαχωριστική Ανάλυση (Discriminant Analysis) και η Ανάλυση σε Κύριες Συνιστώσες (Principal Components Analysis, PCA). Κύριο στόχο αποτελεί η σύγκριση προτύπων που προκύπτουν από τις διαδικασίες ανάλυσης των στατιστικών μεθόδων, της K-Means, της Διαχωριστικής Ανάλυσης και της Ανάλυσης σε Κύριες Συνιστώσες (PCA) σε δεδομένα που αφορούν σε συγκεντρώσεις βιοδεικτών.

Σχεδιασμός: Στην έρευνα συμμετείχαν 303 άτομα, 151 πάσχοντες and 152 μάρτυρες. Οι 151 ασθενείς/ πάσχοντες είχαν διαγνωστεί με νεφρική ανεπάρκεια. Για το σκοπό της έρευνας μετρήθηκαν οι συγκεντρώσεις των βιοδεικτών AST (SGOT), ALT (SGPT), Glucose Serum, Creatinine (Serum Creatinine), Urea, Serum Uric Acid, Serum Calcium, Potassium Serum, Sodium Serum, Total Albumins, Albumin, Alp, γ-GT, CRP, LDH και CPK.

Μέθοδοι: Η Μεταπτυχιακή Διπλωματική Εργασία επικεντρώνεται στην παρουσίαση τριών διαφορετικών μεθόδων ταξινόμησης της K-Means, της Διαχωριστικής Ανάλυσης και της Ανάλυσης σε Κύριες Συνιστώσες (PCA).

Αποτελέσματα: Τα αποτελέσματα της PCA κατέδειξαν την ύπαρξη 5 Κύριων Συνιστωσών, μεταξύ των οποίων η τρίτη αναδεικνύεται ως η Κύρια Συνιστώσα της νεφρικής λειτουργίας και απαρτίζεται από τις μεταβλητές ουρία (Urea), κρεατινίνη (Creatinine) και ουρικό οξύ (Serum Uric Acid), οι οποίες είναι και οι μεταβλητές που αναδεικνύουν την ύπαρξη ή όχι της νεφροπάθειας. Από τα scatter plots όλων των συνδυασμών των αξόνων αναδείχτηκε ότι πράγματι η Συνιστώσα της νεφρικής λειτουργίας είναι αυτή που καταδεικνύει τη διαφοροποίηση των πασχόντων από τους μάρτυρες. Η Διαχωριστική Ανάλυση εφαρμόστηκε 2 φορές. Αρχικά εφαρμόστηκε στο σύνολο των 16 μεταβλητών που μετρούν τις συγκεντρώσεις των βιοχημικών αναλύσεων των συμμετεχόντων στην έρευνα και κατέδειξε ότι πράγματι η ουρία είναι ο καλύτερος προγνωστικός παράγοντας, η κρεατινίνη ο επόμενος και το ουρικό οξύ ο τρίτος αναφορικά με το διαχωρισμό των πασχόντων από τους μάρτυρες. Η Predicted Group Membership Accuracy επαληθεύτηκε, όπως ακριβώς επαληθεύτηκε και η υψηλή ευαισθησία και η ειδικότητα των δεδομένων.

Στη συνέχεια εφαρμόστηκε μόνο για τις τρεις αυτές μεταβλητές και κατέδειξε ότι πράγματι είναι οι κατάλληλοι προγνωστικοί παράγοντες για το διαχωρισμό των δύο ομάδων, των πασχόντων από τους μάρτυρες. Ειδικότερα, η κρεατινίνη αναδείχθηκε ως ο καλύτερος προγνωστικός παράγοντας, η ουρία ο επόμενος και το ουρικό οξύ ο τρίτος αναφορικά με το διαχωρισμό των πασχόντων από τους μάρτυρες. Η Predicted Group Membership Accuracy επαληθεύτηκε, όπως ακριβώς επαληθεύτηκε και η υψηλή ευαισθησία και η ειδικότητα. Η K-Means που εφαρμόστηκε μόνο για τις τρεις αυτές μεταβλητές κατέδειξε ότι η ουρία, η κρεατινίνη και το ουρικό οξύ διαχωρίσουν ικανοποιητικά τους πάσχοντες από τους μάρτυρες.

Συμπεράσματα: Στοχοθεσία της έρευνας αποτέλεσε η σύγκριση 3 κατηγοριών τεχνικών ταξινόμησης όπως είναι η Principal Components Analysis (PCA), η Discriminant Analysis και η K-Means που είναι στατιστικά διαφορετικές προσεγγίσεις σε πραγματικά δεδομένα τα οποία αναφερόταν στις μετρήσεις συγκεντρώσεων δεικτών βιοχημικών αναλύσεων. Τα αποτελέσματα κατέδειξαν τη συμπληρωματικότητά τους σε σχέση με τους βιοδείκτες πλάσματος και σε ό,τι αφορά στη νεφρική ανεπάρκεια. Επιπρόσθετα, κατέδειξαν ότι οι τρεις μέθοδοι λειτουργούν συμπληρωματικά αναδεικνύοντας η καθεμία μια διαφορετική διάσταση της ερμηνείας των δεδομένων, η ερμηνεία και αποσαφήνιση των οποίων δεν θα ήταν καθοριστική χωρίς τη συμβολή της Ιατρικής και των κλινικών ιατρών.

1. Introduction

Patterning methods are well known methodological tools. PCA reduces data in patterns based on correlations between them and a factor score for all of them that can be assigned to a participant. K-Means data are defined by managing differences in means of groups. The interpretation of the findings is based on the fact that a participant is associated with only one cluster that has a specific structure. Discriminant analysis is a type of profile analysis or an analytical predictive technique. The Classification Results show the Predicted Group Membership accuracy in the original sample and demonstrates the sensitivity and specificity measurements. Discriminant analysis establishes the predictors that contribute most to group discrimination.

Although there is an extensive budding of pattering methods in medicine, a direct comparison of multivariate methods in group/cluster identification has not been carried out with respect to biomarkers. Thus, our primary objective was to compare the outcomes of K-Means, Discriminant Analysis and PCA in relation to measurements of AST (SGOT), ALT (SGPT), Glucose Serum, Urea, Creatinine Serum, Uric Acid, Serum Calcium, Potassium Serum, Sodium Serum, Total Albumins, Albumin, Alp, γ-GT, CRP, LDH and CPK Concentrations. More especially, the primary objective was to compare the patterns related to Urea, Creatinine and Serum Uric Acid, the biomarkers relating to the existence of kidney disease.

The secondary objective was to put, PCA Analysis, Discriminant Analysis and K-Means clustering in order to accentuate the similarities and differences of the three methodologies.

2. Data Clustering Techniques

This section is dedicated to the presentation of the three main types of clustering methods, Principal Components Analysis (PCA), Discriminant Analysis and K-Means.

2.1. Principal Component Analysis

Principal Component Analysis or PCA is a method for the analysis of multivariate data, and it is considered as a part of factor Analysis.

The principal objectives of PCA are:

- Data Reduction. PCA aims to replace highly correlated variables with a small number of correlated variables (Dafermos, 2013).
- Data detection and establishment of a structure/model. The goal of PCA is, namely, to accentuate structures or fundamental relations existing between the existing variables (Dafermos, 2013). Moreover, PCA aims to bring to light and assess latent variables, to detect and assess latent sources of variability and co-variability in observable measurements.

- Detection of patterns. The goal of PCA is to detect prototype correlations which may potentially determine causality relations between the examined variables (Dafermos, 2013).

PCA is a descriptive or explanatory method and does not rest on conditions. In reality, PCA rests on the spectrum analysis of the variance or correlation matrix. Principal Components Analysis (PCA) is by far the most widespread pattern recognition tool**.** It is a method for compressing a lot of data into patterns that capture the essence of the original data. More specifically, it constitutes a multivariate statistical analysis that is often used to reduce the dimension of data for easy exploration. Its objectives include: 1) to reduce the original into a lower number of orthogonal (uncorrelated), synthesized variables; 2) to visualize correlations among the original variables and between these variables and the components and 3) to visualize proximities among statistical units. Furthermore, PCA is considered to be a change of variable space.

It rests on the study of eigenvalues and eigenvectors in the correlations or covariance matrix.

As a multivariate analysis technique for dimension reduction, PCA aims to compress the data without losing much of the information contained in the original data. The process regards explaining the variance-covariance structure of a set of variables through a few new variables. All principal components are specific linear combinations of the p random variables exhibiting three important properties:

1. The principal components are uncorrelated. There are also orthogonal uncorrelated, linear combinations of standardized variables.
2. The first principal component has the highest variance; the second principal component has the second highest variance, and so on.
3. The total variation, if all the principal components combined, is equal to the total variation in the original variables.

In reality, PCA converts data to a set of linear components and, as it is characteristically alluded by Field (2009), it converts them to measurable ones.

Each component has the form: $Component_i=b_1X_1+ b_2X_2+\ldots b_nX_m$. It is evident that PCA forecasts components based on measured variables. It is rendered that PCA breaks down the original data to a model of linear variables. PCA brings to light which linear components exist in the data and the manner by which one particular variable contributes to the shaping of each component (Field, 2009).

PCA rests on the overall variance of the variables in descending order. The first Principal Component (PC1) captures the most variance of the data; the second Principal Component (PC2), which is not correlated with PC1, captures the second variance etc.

The number of the components extracted is equal to the original variables and the sum of their variance is the sum of the variance of the original variables.

The sum of the squares of loadings to a principal component signifies the participation of the component to the overall variance of the variables. The value of the sum for each principal component is called eigenvalue. Eigenvalues are presented in descending order and allow the exclusion of components which do not interpret a satisfactory percentage of the overall variance, resulting in only components interpreting a satisfactory percentage of the overall variance to be employed for the interpretation of the results.

The assumptions of PCA:

- Sample, size of sample and sampling adequacy: The sample must be random. The size of the sample must be at least 300 cases. A sample comprising of 50 cases is far too small, one with 100 cases small, 200 cases are fairly satisfactory, while a sample with 500 is very good and one with 1000 cases, excellent. Sampling adequacy is checked using the Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy. In order for it to be accepted, the KMO index must exceed value 0.70. More specifically, it is deemed to be excellent when it exceeds 0.90, very good when it ranges from 0.80-090, good in the interval 0.70-0.80, insignificant should it range from 0.60 to 0.70, very small if between 0.50 and 0.60 and unacceptable if below 0.50 (Kaiser, 1974).
- Data: Data must be quantitative; they must be of the interval ratio type. The case of the 5-point Likert scale is considered to be a fixed ratio scale. In addition, dichotomous variables may be employed, where 0 signifies the absence of the measurable characteristic, while 1 signifies its presence (Dafermos, 2013, p. 32).
- Linearity: Data must exhibit high correlation coefficients.
- Absence of extreme values, outliers: Data distribution must not be asymmetrical and contain extreme values or outliers.
- Misusing values: Care must be shown as to the management of misusing value and the distribution of such values must be investigated, together with the possibility of them following nonrandom patterns.
- Bartlett's Test of Sphericity: Testing for Sphericity checks the null hypothesis: the null (Ho) hypothesis: all correlation factors are not far removed from zero. The Ho hypothesis must be rejected so as to allow for PCA.
- Rotation: In case where components are uncorrelated, the orthogonal rotation will be employed, where at the rotation components intersect vertically. In case where components exhibit a correlation greater than 0.30, then the oblique rotation is employed, where components do not form a 90^{o} angle between them. The goal of the

rotation is for large loadings of the variable to become larger and the small ones, smaller.

The criteria for selecting components:

- Components are selected based on the variance percentage they explain. Selected is the set of components explaining a total overall variance percentage greater or equal to 70%. A percentage in the region of 70-80% is deemed satisfactory.
- Selected are components whose eigenvalues are equal or greater than one (Kaiser, 1960) or equal or greater than 0.70 (Jolliffe, 1972, 1986).
- The scree plot criterion, which also constitutes the graph for each eigenvalue, depicted on the yy′ axis, and of the components, depicted on xx′ axis, is used for the selection of components (Cattell, 1966). According to this criterion the turning point, i.e the point where the slope of the curve levels off is considered as the limit for the selection of factors. The factors before this point are selected.
- Communalities must be over 0.40. After having determined the number of factors communalities are also redetermined.
- The rotation of factors improves the interpretation of data.

2.2. Discriminant Analysis

Discriminant Analysis has been defined as a multivariate technique disturbed with the classification of a new object x, with x a random vector expressed by a set of attributes (x_1, x_2,…,$x_{p)}$, into one of two or more distinct populations (Batsidis & Zografos, 2006). Discriminant Analysis allows for two or more groups of cases to be distinguished or, better, to be separated, based on the variables measured in each case. These groups are known beforehand. More specifically, Discriminant Analysis caters for the successful examination and classification of cases in the groups to which they belong. The goal of Discriminant Analysis is to establish rules for deciding with respect to the classification of observations across various populations. It is method aiming for pattern recognition.

In medicine, for example, it is frequently requested a rule be constructed, taking account of the symptoms of an illness in order a new patient to get the appropriate diagnose. This rule will guide decision-making in the future (Karlis, 2005).

Discriminant Analysis is used for the better understanding of the importance of multivariate analysis of variance (MANOVA) (Howitt & Cramer, 2010) and in reality, it produces a classification matrix that depicts the accuracy of the determination of the quality of some member of a group based on the independent variables. It must be emphasized at this point, that the independent variable in MANOVA is the dependent one in Discriminant Analysis,

thus a number of numerical variables are required and one unique nominal variable (categorical variable) that will define the groups used (Field, 2009).

Hair et al. (2005, p.256) designate the objectives of Discriminant Analysis to be:

- to ascertain if there are differences that are statistically significant between the score profiles on a set of variables for two or more a priori defined groups;
- to establish which independent variables, explain most of the differences in average score profiles between these two or more groups;
- to set up procedures for the classification of objects into groups based on their scores on a set of independent variables;
- to determine the number and composition of discrimination's dimensions between groups formed from the set of independent variables (Chatsidis, 2015).

The Assumptions of Discriminant analysis:

The most popular method of Discriminant Analysis shares the same assumptions as MANOVA (Batsidis & Zografos, 2011; Chatsidis, 2015), while it is quite sensitive to extreme values, outliers, and predictor variables that must always be less in size from that of the smallest group.

- Multivariate normality: Multivariate normality assumes that the joint effects of a pair of variables are normally distributed.
- Homoscedasticity or homogeneity of variance/covariance: Box's M statistic test in the Equality of Covariances procedure can be used to test it, or, alternative, it can be tested by looking for equal slopes in Probability Plots. Having said this, it has been suggested for linear Discriminant analysis to be used when covariances are equal and for quadratic Discriminant analysis to be employed when they are not equal.
- Equality of Variance-Covariance Matrices: It assumes the equivalence of covariates matrices across the groups (Hair et al., 2005).
- Multicollinearity: An increased correlation between predictor variables can cause a decrease of the predictive power.
- Independence: It is assumed that participants are randomly sampled and it is also assumed that a participant's score on one variable has to be independent of all other participants' scores on that variable (Hair et al., 2005).
- Sensitivity to outliers: Outliers impact is possible to disproportionate in the overall results, and thus, they are ought to be eliminated (Hair et al., 2005).

A tolerance of slight transgressions of the assumptions above has been suggested for Discriminant Analysis, while its reliability has been shown even when using dichotomous variables (multivariate normality is often violated in such cases).

Discriminant Function: Discriminant Analysis works by creating one or more linear combinations of predictors, creating a new latent variable for each function. These functions are called discriminant functions.

The first function that is created maximizes the differences between groups with respect to that function. The second function maximizes differences with respect to it, but must also not be correlated with the previous function. And so on for subsequent functions, the requirement being that each new function is not correlated with any of the previous ones.

A discriminant score is assigned to each function in order to determine how well it predicts group placement.

- Structure Correlation Coefficients indicate the correlation between each predictor and the discriminant score of each function.
- Standardized Coefficients indicate each predictor's unique contribution to each function. This is, therefore, a partial correlation, indicating the relative importance of each predictor in predicting group assignment from each function.
- Functions at Group Centroids: Mean discriminant scores for each grouping variable are given for each function. The further apart the means are, the less errors twill be in classification.

Discriminant Analysis involves the derivation of a variate, the variate being the linear combination of two or more variables that best discriminate between groups that have been a priori defined. Achieving discrimination entails setting the weights of the variate for each variable so that between-group variance is maximized relatively to within-group variance. The discrimination function, which is the linear combination for a discrimination analysis, is derived from an equation of the form that follows:

$Z_{ij}=a+W_1X_{1k}+ W_2X_{2k+\ldots+} W_nX_{nk}$

where Z_{ij} = discriminant Z score of discriminant function j for object k

a= intercept

W_i = discriminant weight for independent variable i

X_k = independent variable i for object k

This score is a metric variable providing a direct means of comparing observations on each function. A measure of the group deference is a comparison of the group cancroids, the average discriminant Z score for all groups' members. The difference between cancroids is measured in terms of Mahalanobis D^2 measure.

2.3. K-Means

Amongst the various partitioning-based data clustering methods, K-Means is one of the simplest ones (Karlis, 2005) and has been adapted to many problem domains. Used when

there are unlabeled data, K-Means clustering is a type of supervised learning and specifically, K-Means is one of the simplest supervised learning algorithms that solve the clustering problem (MacQueen, 1967).

The K-Means method functions satisfactorily with large samples. This, of course, depends on the initial values employed.

The K-Means algorithm is also known as a partitioning algorithm. The algorithm partitions the multilevel plane created by the data in places, and corresponds an area to each group (Karlis, 2005).

K-Means clustering is intended to partition *n* objects into *k* clusters, where each object belongs to the cluster with the nearest mean.

Produced by these methods are exactly *k* different clusters of greatest possible distinction. The best number of clusters, *k,* leading to the greatest separation (distance) is not a priori known and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

$J=\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^j - c_j\right\|^2$

Where $\left\|x_i^j - c_j\right\|^2$ is the measurement of a chosen distance between a data point x_i^j and the cluster centre and functions as an indicator of the distance between the n data points and their respective cluster centers (MacQueen, 1967).

So, centroids are assigned to every cluster (Karlis, 2005). This is a partition method technique identifying mutually exclusive clusters of spherical shape. It generates a specific number of disjoint, flat (non-hierarchical) clusters. Static methods can be employed in order to cluster and assign rank values to the clustered categorical data. In this case, categorical data have been converted into numeric, by assigning rank values (MacQueen, 1967).

K-Means algorithm organizes objects into k-partitions (k-clusters) where each partition represents a cluster. We start out with an initial set of means and classify cases based on their distances to their centers. The center of a cluster is nothing more than the mean value of all observations for each variable in the cluster. It, essentially, corresponds to the vector of the means. Should the data be ordinal, the medoid will be employed, which is the top for nominal data, namely the most frequent value (Karlis, 2005). In mixed type data, the center for each cluster may comprise of the peaks of categorical variables and the means of continuous ones.

Next, we compute cluster means again, using the cases that are assigned to the clusters and then, we reclassify all the cases based on the new set of means. This step is repeated until cluster means don't change between successive steps/ repetitions. We calculated the cluster's means once more and assign cases to their permanent clusters. The distance employed for the

assignments is the Euclidean distance, although any other types of distance may also be employed. Below the algorithm of k-means methodology is going to be presented.

Thus, the algorithm of k-means is:

a) Decide on a value for *k*, the number of clusters.

b) Initialize the *k* cluster centers.

c) Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.

d) Re-estimate the *k* cluster centers, by assuming the memberships found above are correct.

e) Repeat 3 and 4 steps until none of the *N* objects changed membership in the last iteration.

K-Means clustering is a relatively efficient and scalable process for huge sums of data sets while it is easy to be understood and implemented. Some of its drawbacks include the fact that the process commences only after the mean of a cluster is initialized, while user defined clusters are constant and find it hard to handle data with noise and outliers (Karlis, 2005; Field, 2009). Unfortunately, there is no globally accepted theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different *k* and choose the best one, based on a preselected criterion. In general, a large *k* will potentially decrease errors but will increase the risk of overfitting.

The following table (Table 1) presents some of the visual look at the clusters of each method.

Table 1: Visual look at the clusters

Principal component analysis	**Discriminant Analysis**	**K-Means clustering**
PCA is a frequently employed statistical technique for unsupervised dimension reduction.	Discriminant Analysis is a supervised learning statistical technique or a supervised classification.	A popular data clustering method in the case of supervised learning tasks is K-Means clustering.
PCA is often used to transform high dimensional data into lower dimensional ones (Jolliffe, 2002) (singular value decomposition). Coherent patters can be detected more clearly in lower dimensional data.	Irrespective if one considers Discriminant Analysis as a type of profile analysis or an analytical predictive technique, it provides a basis for classifying both the sample used in order to estimate the discriminant function as well as any other observations having values for all independent variables.	One of the most popular and efficient clustering methods is the K-Means method (Hartigan & Wang, 1979; Lloyd, 1957; MacQueen, 1967) which uses prototypes (centroids) to represent clusters by optimizing the squared error function.
The number of components is unknown.	The number of groups is known.	The number of groups is known.
PCA-based dimension reduction is based on the ability of PCA to pick up those dimensions that exhibit the largest variances. In mathematical	LDA is also closely related to PCA in the sense that they both seek the linear combinations of variables best explaining the data.	

terms, this is equivalent with identifying the best low rank approximation (Santhanalakshi, R., & Alagarsamy) (in L2 norm) of the data via the singular value decomposition (SVD) (Eckart & Young, 1936). However, this noise reduction property cannot alone adequately explain the effectiveness of PCA.		
PCA is similar to MANOVA.	Discriminant Analysis is a reversed statistical technique related to MANOVA.	
PCA usually deal with correlation matrices and it is possible to analyse a variance-covariance matrix.		
It reduces data into patterns based on correlation between variables and individuals received a factor score for all derived factors.	As a profile analysis, it provides an objective assessment of differences between groups on a set of impendent variables and it is similar to multivariate analysis of variance.	It reduces data into patterns based on the individual's differences in many items and individuals belong to only one cluster.
Principal Component Analysis is one of the most useful data analysis and machine learning methods which can be used to identify patterns in highly complex datasets, letting one know which of the variables in one's data the most important ones are. Finally, it can let one see. Lastly, it can tell you how accurate your new understanding of the data actually is.	Assists in the understanding of group differences while providing insight into the role of individual variables. It also defines combinations of such variables to represent dimensions of discrimination between groups.	

2.4. Biochemical analyses

Biochemical testing was carried out in the sera of 151 patients (cases) suffering from kidney disease and 152 controls, which, among others, included the analysis and calculation for the following items/variables. The data was derived and given from a hospital data basis, in order to be used only for didactical purposes.

1. Transaminases: AST (SGOT) Aspartate aminotransferase / Serum glutamic-oxaloacetic transaminase (SGOT) and ALT (SGPT) Alanine aminotransferase / Serum

glutamate-pyruvate transaminase (SGPT) with reference values 10-37U/L and 10-45U/L respectively.

2. Glucose Serum: Known to the general public as blood sugar, glucose serum refers to concentration or glucose in the bloodstream with reference values for fasting glucose serum levels <100mg/dL and 101-125mg/dL for prediabetes (impaired) glucose serum levels.

3. Urea (kidney function) and Serum Creatinine are the items/laboratory tests/measurements which check kidney/renal function with reference values 10-43mg/dL for Urea and with respect to creatine A<50: 0.84-1.25mg/dL, A>0.81-1.44mg/dL and Γ: 0.66-1.10 mg/dL.

4. Serum Uric Acid: with reference values ranging from 3.5 to 7.2mg/dL.

5. Electrolytes, and more specifically, Serum Calcium, Phosphorus serum (P), Potassium serum (K), Sodium serum (Na), Magnesium serum, with reference values 8.8-10.6 mg/dL for Calcium, 2.5-4.5 mg/dL for Phosphorus, 3.5-5.1 for mmol/L for Potassium and 136-145 mmol/L for Sodium.

6. Albumin and Total Albumins (TP).

7. Total bilirubin, namely Indirect bilirubin and Direct bilirubin, which constitute the items/laboratory tests/measurements that check for Jaundice and whose reference values are 0.3-0.2 mg/dL for Total bilirubin and 0.00-0.20 mg/dL for Direct bilirubin.

8. Alkaline phosphatase (Alp) and Gamma-glutamyl transferase (γ-GT), where increased Alp values signify a bone problem and increased values for both Alp as well as γ-GT signify a Hepatopathy with reference values 30-120U/L for Alkaline phosphatise and <55U/L for Gamma-glutamyl transferase.

9. C-reactive protein (CRP), which is an enzyme protein that facilitates to extract chemical changes in the body found in your heart, brain and skeletal muscles with reference values <6 mg/L.

10. Lactate Dehydrogenase (LDH) with reference values <248 U/L.

11. *Creatine phosphokinase* (CPK*)* with reference values <170 U/L.

However, only 16 of these items were used for the needs of this Msc Thesis, and more specifically: transaminases AST and ALT, Serum glucose, Urea, Serum keratinise, Alp and γ-GT, Serum Uric Acid, TP, Albumin, Serum Calcium Potassium Serum, while testing for them was common for the cases and the controls.

2.5. Participants

303 patients participated in this survey, 151 cases and 152 controls. The 151 patients (cases) were diagnosed for renal or kidney disease. Of these 151 cases, 71 were males and 80 females. With respect to the 152 controls, 69 were males and 83 females.

2.6. Limitations

1. The research included participants who underwent biochemical analyses. 151 participants were patients and the results originated from the Nephrology clinic of the hospital, while 152 were controls from other clinics and out-patient departments. This fact shows the existence of bias, which constitutes the most important limitation for the research.
2. The representativeness of the sample.
3. Data processing relates more to the demonstration of the methodology for teaching purposes and for the interpretation of medical data.

3. Results

In this section Descriptive statistics for controls cases are presented. Results from the application of PCA, K-means and Discriminant Analysis are followed.

3.1. Principal Component Analysis (PCA) results

Principal Component Analysis (PCA) results*:* Kaiser-Meyer-Olkin (KMO) Measure of the Sampling Adequacy and Bartlett's Test of Sphericity Measure for the suitability of the method were both tested before the analysis of the factor analysis results.

Both the Kaiser-Meyer-Olkin (KMO) factor, equal to 0.650 and deemed very satisfactory as it exceeds the acceptable value of 0.60, as well as Bartlett's Test of Sphericity (x^2=1414,953, df=120, p<0.001) have shown that the application of the Principal Component Analysis with oblique rotation method is permitted.

The application of Principal Component Analysis with varimax rotation for all variables on the basis that the characteristic root or eigenvalue criterion is over one (eigenvalue≥ 1) was

verified for 5 Components. These specific factors explained 60.296% of variance. Similarly, according to the Scree Plot criterion, the steep descending trend of eigenvalues begins after the 5th Principal Components (PC5) (Cattel, 1996). Consequently, the existence of the 5 Components was verified.

The first Principal Component (PC1), with an eigenvalue equal to 3.004, interprets 16.531% of the total variance of data, a percentage deemed satisfactory (Hair, 2005), gathers values for variables AST, γ_GT, Alp_Phosphatase and ALT with very high loadings, whose values amount to 0.829, 0.810, 0.771 and 0.707 (Table 2).

The values of the Communalities of items AST, γ_GT, Alp_Phosphatase και ALT, take on values 0.743, 0.725, 0.651 and 0.608, exceeding the 0.40 value criterion posed as the limit for the verification of the satisfactory quality for the variables of the First Component (PC1). The First Component (PC1) is constructed and interpreted by transaminases AST (SGOT) (Aspartate aminotransferase/Serum glutamic-oxaloacetic transaminase (SGOT)) and ALT (SGPT) (Alanine aminotransferase/Serum glutamate-pyruvate **t**ransaminase (SGPT)), enzyme Alp_Phosphatase and enzyme γ_GT. The First Component (PC1) is shown to essentially be the Component of renal function.

The Second Component (PC2) refers to all blood proteins called Total Albumins (TP) and Albumin, one of the two blood protein categories, the content of which with respect to all albumin amounts to approximately 60% of Total Protein, and to electrolytes Calcium (Ca) and Pottasium (K).

This Component has an eigenvalue of 2.660 and interprets 14.422% of total data variance. The eigenvalue criterion, eigenvalue over one, verifies that the 4 variables TP, Albumin, Calcium and Potassium, which exhibit very high loadings 0.815, 0.775, 0,755 and 0.480 correspondingly, are represented by the same conceptual construct (Table 2). The values for the Communalities of TR, Albumin, Calcium and Potassium take on prices 0.714, 0,682, 0.572 and 0.501 respectively and exceed the 0.40 value criterion posed as the verification limit for the satisfactory quality of statements of Second Component (PC2).

The Third Component (PC3) (Table 2) refers to Urea, which is the final product from the metabolism of proteins, Creatinine which is a nitrogen product of metabolism and Uric_acid, and exhibit high loadings of 0.827, 0.730 and 0.679 respectively, with an eigenvalue of 1.723, that interprets 13.382% of total data variance, a percentage deemed satisfactory (Hair et al., 2005), while falling under it are, in order, elements Urea, Creatinine and Uric_acid. The values of the Communalities of Urea, Creatinine and Uric_acid take on prices 0.742, 0.727 and 0.517 exceeding the 0.40 value criterion posed as the limit for the verification of the satisfactory quality of Third Component (PC3). The Third Component (PC3) is essentially shown to be the Component of renal function.

Table 2: Rotated Component Matrix

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
AST	,829				
γ_GT	,810				
Alp_Phosphatase	,771				
ALT	,707				
TP		,815			
Albumin		,775			
Calcium		,755			
Pottasium_K		,480	,393		
Urea			,827		
Creatinine			,730		
Uric_acid			,679		
CRP				-,721	
LDH				,511	
Sodium_Na				,493	
Serum_glucose					-,845
CPK					,435

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

The Fourth Component (PC4) (Table 2) that refers to CRP, LDH and Sodium (Na), with an eigenvalue of 1.171, interprets 8.486% of the total data variance, a percentage deemed satisfactory (Hair et al., 2005). More specifically, falling under the Fourth Component (PC4) are, in order, items CRP, LDH and Sodium_Na, with loadings -0.721, 0.511 and 0.493 respectively (Table 2). One must note, at this point, that the negative sign regarding loading CRP means that it runs contrary to the other ones and, more specifically, to the Fourth Component (PC4). The values of Communalities for elements CRP, LDH and Sodium (Na) take on prices 0.608, 0.524 and 0.328 respectively and exceed the 0.40 value criterion as the limit for the verification of the satisfactory quality Fourth Component but Sodium (Na).

The last Fifth Component (PC5) (Table 2) of this analysis, has an eigenvalue of 1.090, and interprets 7.475% of total data variance. The eigenvalue criterion (eigenvalue over one) verifies that Serum_glucose and enzyme CPK which is an en enzyme, protein that facilitates to extract chemical changes in the body found in your heart, brain and skeletal muscles represent the same conceptual construct. Communalities values for Serum_glucose and CPK take on prices 0.715, 0.682, and 0.291 respectively. Communality of the enzyme CPK did not

exceed the 0.40 value criterion and maybe its participation in the analysis should be reconsidered. More specifically, Serum_glucose and enzyme CPK fall, in order, under the fifth component, their loadings being -0.845 and 0.435 respectively. The negative sign of the loading for Serum_glucose means that it runs contrary to enzyme CRP and the Fifth Component (PC5).

The charts that follow present the results of the components with respect to the differences between the cases (patients) and the controls (cases-controls).

It follows from the first scatter plot (PC1xPC2) (Figure 1) that:

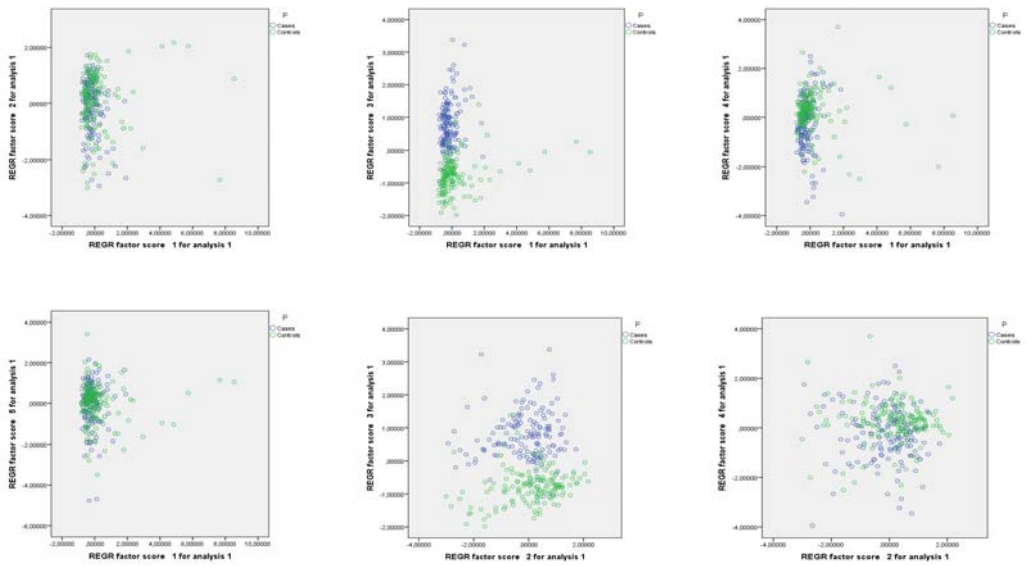
- There is some small distinction between the two groups.
- Cases tend to have a little more smaller values from controls.
- Controls are more homogeneous but exhibit straightforward outliers.

The second scatter plot (PC1xPC3) (Figure 1) shows that:

- There is a clear distinction between the two groups.
- Cases tend to have larger values from controls.
- Controls are relative more homogeneous but exhibit clear outliers. Cases also exhibit some outliers.

It follows from the third scatter plot (PC1xPC4) (Figure 1) that:

- There is no distinction between the two groups and in addition there is fairly extensive overlapping between the two groups.
- Cases tend to have lower values from controls.
- Controls are more homogeneous but exhibit clear outliers.



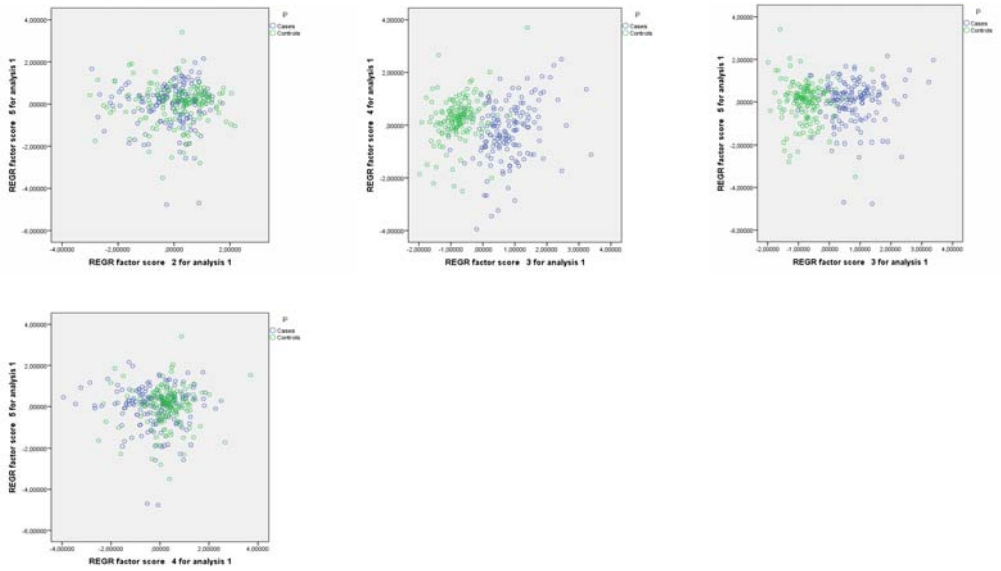


Figure 1: Scatter plot

It follows from the fourth scatter plot (PC1xPC5) (Figure 1) that:

- There is no distinction between the two groups and, additionally, there is extensive overlapping between them.
- Cases tend to exhibit greater values than controls.
- Controls are more homogeneous than cases.

It follows from the fourth scatter plot (PC1xPC5) (Figure 1) that:

- There is no distinction between the two groups and, additionally, there is extensive overlapping between them.
- Cases tend to exhibit greater values than controls.
- Controls are more homogeneous than cases.

It ensues from the fifth scatter plot (PC2xPC3) (Figure 1) that:

- There is a great distinction between the two groups.
- Controls tend to have more concentrated values than cases.
- Controls are more homogeneous but exhibit clear outliers.

It follows from the sixth scatter plot (PC2xPC4) (Figure 1) that:

- There is no distinction between the two groups.
- Controls tend to have more concentrated values than cases.
- Both controls and cases exhibit clear outliers.

The seventh scatter plot (PC2xPC5) shows that:

- There is no clear distinction between the two groups.
- Cases tend to exhibit a similar concentration of values to controls.

- Controls are somewhat more homogeneous but exhibit clear outliers. Cases also exhibit some outliers.

It follows from the eighth scatter plot (PC3xPC4) (Figure 1) that:

- There is a clear distinction between the two groups.
- Cases tend to exhibit values that are more remote than those of controls.
- Controls are more homogeneous but exhibit unambiguous outliers. Cases exhibit some outliers as well.

It ensues from the ninth scatter plot (PC3xPC5) (Figure 1) that:

- There is a clear distinction between the two groups.
- Cases tend to have a similar dispersion of values compared to controls.
- Controls are more homogeneous but exhibit some outliers. Cases exhibit some outliers as well.

It follows from the tenth scatter plot (PC4xPC5) (Figure 1) that:

- The two groups are not clearly distinguished.
- Controls tend to exhibit a higher concentration, namely a small dispersion of values compared to cases.
- Controls are more homogeneous but exhibit certain outliers. Outliers are also present in cases.

3.2. Discriminant Analysis results

In this section two Discriminant Analysis results are followed. The first one includes the 16 examined variables and the second one only Urea, Creatinine and Serum Uric Acid related to kidney function.

Pursuant to the table of the Tests of Equality of Group Means, Wilks' Lambda is statistically significant for each predictive variable except Serum_glucose, Calcium, Alp_Phosphatase, LDH and γ_GT (Table 3).

Table 3: Tests of Equality of Group Means

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
AST	,919	26,472	1	301	,000
ALT	,936	20,563	1	301	,000
Serum_glucose	,999	,298	1	301	,585
Urea	,413	428,598	1	301	,000
Creatinine	,425	407,404	1	301	,000
Uric_acid	,764	92,871	1	301	,000
Calcium	,994	1,715	1	301	,191
Pottasium_K	,977	6,990	1	301	,009

Sodium_Na	,974	7,992	1	301	,005
TP	,923	24,979	1	301	,000
Albumin	,959	12,734	1	301	,000
Alp_Phosphatase	,999	,296	1	301	,587
CRP	,926	24,144	1	301	,000
LDH	1,000	,000	1	301	,993
CPK	,950	15,714	1	301	,000
γ_GT	,994	1,813	1	301	,179

From the Log Determinants table, the Log Determinant values are similar, 64.879 for cases, 62.424 for controls and 69.107 for Pooled within-groups, very close, thus there is no problem in the analysis of the data (Table 4).

Table 4: Log Determinants

Log Determinants

p1	Rank	Log Determinant
0	16	64,879
1	16	62,424
Pooled within-groups	16	69,107

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

As far as Box's results are concerned the hull hypothesis of equal population covariance matrices is rejected since $p<0.001$ (Table 5).

Table 5: Test Results

Test Results

Box's M		1643,368
F	Approx.	11,418
	df1	136
	df2	279759,853
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Account is taken of the Eigenvalue for each function in the analysis (Table 6). The Eigenvalue is converted into percentage of variance account for, and the first variate accounts for 100%. The larger the Eigenvalue, in this case values 3.268, the more variance the functions explains. Thus, the higher the Eigenvalue the better the Fit is, the better the data fits the model. The Canonical Correlation is high and is equal to 0.875. Its square is used as an effect size (Field, 2009).

Table 6: Eigenvalues

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3,268[a]	100,0	100,0	,875

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda shows how well the prediction model fits. In the present case the prediction model is statistically significant (Wilks' Lambda=0,234, x^2=425.190, df=16, p<0.001) (Table 7). Thus, it can be noted that the first variate alone significantly discriminate the groups and consequently the prediction model is significant.

Table 7: Wilks' Lambda

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,234	425,190	16	,000

Standardized Canonical Discriminant Function Coefficients table shows the standardized discriminate function coefficients for 16 variates (Table 8). In fact, these values represent the standardized versions of the eigenvectors' values. Standardized Canonical Discriminant Function Coefficients shows the importance of the 16 predictors, Urea is the best predictor, Creatinine is the next and etc.

Table 8: Standardized Canonical Discriminant Function Coefficients

Standardized Canonical Discriminant Function Coefficients			
	Function 1	Variable	Function 1
AST	-,042	Sodium_Na	,017
ALT	-,099	TP	-,278
Serum_glucose	-,049	Albumin	,067
Urea	,662	Alp_Phosphatase	,100
Creatinine	,618	CRP	,224
Uric_acid	,172	LDH	-,052
Calcium	,210	CPK	-,170
Pottasium_K	,080	γ_GT	-,099

The Structure Matrix below (Table 9) demonstrates the same information which is in some extent in different form. The values are the canonical variate correlation coefficients, and they are comparable to PCA loadings and designate the substantive character of the variates.

Structure Matrix shows consistency in relation to importance to best predictors' importance. The dependent predictors Urea, Creatinine, Uric_acid have high canonical variate correlation and they contribute most to group discrimination.

Urea is the best predictor, Creatinine is the next and Uric_acid has the lowest value. Canonical variate correlation coefficients for Urea, Creatinine, Uric_acid value for 0,660, 0.644 and 0.307 respectively.

Table 9: Structure Matrix

Structure Matrix

Variable	Function 1	Variable	Function 1
Urea	,660	Albumin	-,114
Creatinine	,644	Sodium_Na	-,090
Uric_acid	,307	Pottasium_K	,084
AST	-,164	γ_GT	-,043
TP	-,159	Calcium	-,042
CRP	,157	Serum_glucose	,017
ALT	-,145	Alp_Phosphatase	-,017
CPK	-,126	LDH	,000

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function

Canonical Discriminant Function Coefficients table demonstrate the Canonical Discriminant Function Coefficients which are the unstandardized versions of the standardized Coefficients, less useful than the standardized Coefficients (Table 10). The specific values are the value of b in equation $D=a+b_1x_1+b_2x_2+\ldots+b_{16}x_{16,}$ where a represents the Constant.

Table 10: Canonical Discriminant Function Coefficients

Canonical Discriminant Function Coefficients

Variable	Function 1	Variable	Function 1
AST	-,002	Sodium_Na	,005
ALT	-,004	TP	-,352
Serum_glucose	-,001	Albumin	,117
Urea	,017	Alp_Phosphatase	,001
Creatinine	,330	CRP	,006
Uric_acid	,101	LDH	-,001
Calcium	,270	CPK	-,002
Pottasium_K	,138	γ_GT	-,001
		(Constant)	-4,332

Unstandardized coefficients

Functions at Group Centroids represent the mean variate scores for each group. In fact, they represent the unstandardized canonical discriminant functions evaluated at group means. The Centroid score for cases equals to 1.808 and for controls -1.796. This means that these groups with values opposite in sign are being discriminated by that variate (Table 11).

Table 11: Functions at Group Centroids

Functions at Group Centroids

p1	Function 1
0	1,808
1	-1,796

Unstandardized canonical discriminant functions evaluated at group means

The Classification Results table shows the Predicted Group Membership accuracy in the original sample and demonstrates the sensitivity and specificity measurements. Sensitivity counts for 92.1% and it is high. High sensitivity means that there are few false negatives. Specificity counts for 99.3% and it is high. High specificity means that there are few false positives. In addition, 95.7% of original grouped cases correctly classified (Table 12).

Table 12: Classification Results

Classification Results[a]

			Predicted Group Membership		
		p1	*0*	*1*	*Total*
Original	Count	0	139	12	151
		1	1	151	152
	%	0	92,1	7,9	100,0
		1	,7	99,3	100,0

a. 95,7% of original grouped cases correctly classified.

On the whole the predictors Urea, Creatinine, Uric_acid have high canonical variate contribute most to the group discrimination.

It ought to be mentioned at this point that logistic regression could be employed.

3.3. Discriminant Analysis results for 3 predictors related to kidney function

It has been suggested by doctors a Discriminant Analysis results for the 3 predictors related to kidney function to be applied.

Wilks' Lambda test revels that each predictive variable (Urea, Creatinine, Uric_acid) is statistically significant according to the following table (Table 13).

Table 13: Tests of Equality of Group Means

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Urea	,413	428,598	1	301	,000
Creatinine	,425	407,404	1	301	,000
Uric_acid	,764	92,871	1	301	,000

Log Determinant values are similar, 10.927 for cases, and 9.624 for Pooled within-groups, but for controls are smaller, only 2.611 (Table 14).

Table 14: Log Determinants

Log Determinants

p1	Rank	Log Determinant
0	3	10,927
1	3	2,611
Pooled within-groups	3	9,624

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

The hull hypothesis of equal population covariance matrices is rejected since $p<0.001$ according to the Box's test (Table 15).

Table 15: Test Results

Test Results

Box's M		863,449
F	Approx.	142,353
	df1	6
	df2	656352,247
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

In the analysis the Eigenvalue is considered for each function. The Eigenvalue, in this case values 2.770, and first variate accounts for 100%. The Canonical Correlation is high and equals to 0.857 (Table 16). Its square is an effect size indicator (Field, 2009).

Table 16: Eigenvalues

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2,770[a]	100,0	100,0	,857

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda shows that prediction model fits well (Wilks' Lambda =0.265, x^2=397.463, df=3, p<0.001) (Table 17).

Table 17: Wilks' Lambda

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,265	397,463	3	,000

The standardized discriminate function coefficients show the importance of the 3 predictors. Coefficients for Urea, Creatinine, Uric_acid value for 0.652, 0.676 and 0.180 respectively Thus Creatinine is the best predictor, Urea is the next and Uric_acid comes last (Table 18).

Table 18: Standardized Canonical Discriminant Function Coefficients

Standardized Canonical Discriminant Function Coefficients	
Variable	Function 1
Urea	,652
Creatinine	,676
Uric_acid	,180

Structure Matrix shows consistency in relation to importance to best predictors' importance. The dependent predictors Urea, Creatinine, Uric_acid have high canonical variate correlation and there contribute the most to group discrimination.

Urea has the highest value, Creatinine is the next and Uric_acid has the lowest value. Canonical variate correlation coefficients for Urea, Creatinine, Uric_acid value for 0.717, 0.699 and 0.334 respectively (Table 19).

Table 19: Structure Matrix

Structure Matrix	
Variable	Function 1
Urea	,717
Creatinine	,699
Uric_acid	,334

.

The Discriminant Function equation takes the form (Table 20):

D=-3,054 +0.016 Urea+0.361Creatinine +0.106 Serum Uric Acid.

Table 20: Canonical Discriminant Function Coefficients

Canonical Discriminant Function Coefficients	
Variabe	Function 1
Urea	,016
Creatinine	,361
Uric_acid	,106
(Constant)	-3,054

Unstandardized coefficients

The Centroid score for cases equals to 1.664 and for controls -1.653. This means that these groups with values opposite in sign are being discriminated by that variate (Table 21).

Table 21: Functions at Group Centroids

Functions at Group Centroids	
p1	Function 1
0	1,664
1	-1,653

Unstandardized canonical discriminant functions evaluated at group means

According to the following table (Table 22), 95% of original grouped cases correctly were classified. Sensitivity counts for 90.7% and it is high. High sensitivity means that there are few false negatives. Specificity counts for 99.3% and it is high. High specificity means that there are few false positives. The Predicted Group Membership accuracy was confirmed.

Table 22: Classification Results

Classification Results[a]					
			Predicted Group Membership		
		p1	*0*	*1*	*Total*
Original	Count	0	137	14	151
		1	1	151	152
	%	0	90,7	9,3	100,0
		1	,7	99,3	100,0

a. 95,0% of original grouped cases correctly classified.

Finally, the predictors Urea, Creatinine, Uric_acid have high canonical variate contribute the most to group discrimination.

3.4. K-Means results

The application of K-Means which is a non-hierarchical method has given the following results on the examined standardized values of the variables. So, clusters are based on the standardized values of the measurements. The application of K-Means was limited to these three variables: Urea, Creatinine and Serum Uric Acid which also constitute the criteria for kidney disease, while it showed that they are indeed the appropriate predictors for the separation of the two groups. The Initial Cluster Centers present the Zscore. For Cluster 1 all Zscores have a negative sign in contrast with Cluster 2 where all Zscores have a positive sign (Table 23).

Table 23: Initial Cluster Centers

Initial Cluster Centers		
	Cluster	
	1	2
Zscore(Urea)	-1,14137	1,65127
Zscore(Creatinine)	-,91915	4,26190
Zscore(Uric_acid)	-2,26262	1,20301

The Final Cluster Centers table shows how far the relative centers are. Cluster 1 has the lowest highest scores of Urea, Creatinine and Serum Uric Acid the negatives one Cluster 2 presents the highest scores of Urea, Creatinine and Serum Uric Acid (Table 24).

Table 24: Final Cluster Centers

Final Cluster Centers		
	Cluster	
	1	2
Zscore(Urea)	-,72634	,88009
Zscore(Creatinine)	-,65937	,79895
Zscore(Uric_acid)	-,57125	,69217

The following graph (Figure 2) gives a visual look at the clusters. The blue Colum represents Urea, Green represents Creatinine and finally Grey Colum represents Serum Uric Acid. Cluster 2 presents the highest scores of Urea, Creatinine and Serum Uric Acid. Finally, cluster 1 represents the lowest, the negatives ones, below zero. Thus, that's how all lays on the graph. The graph presents how these variables used in order to determine which cluster each participant landed in.

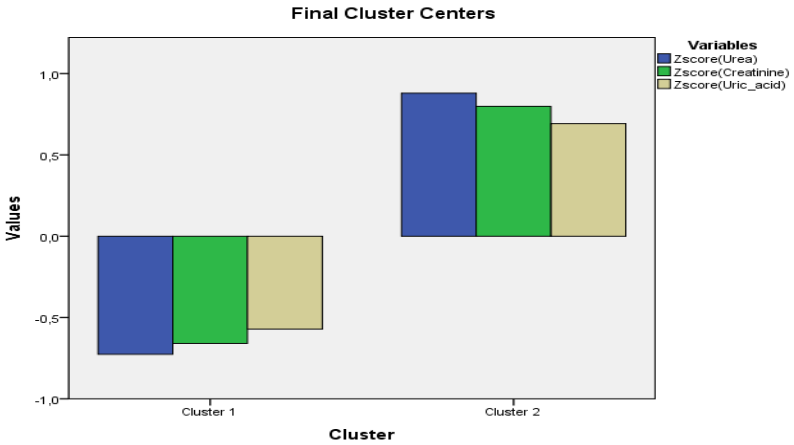


Figure 2: Clusters' visualization

Another way to look at the important of each variable and determined cluster is ANOVA table. F scores have very high values, 538.271, 337.451 and 197.930 for Urea, Creatinine and Serum Uric Acid respectively, which are statistically significant. F scores are relative weight given to a particular variable in order to determine in which cluster a participant was allocated to. All these F values are very large, all statistically significant and thus, all these variables have significant impact on determine which cluster a patient is allocated to (Table 25).

Table 25: ANOVA

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(Urea)	193,689	1	,360	301	538,271	,000
Zscore(Creatinine)	159,621	1	,473	301	337,451	,000
Zscore(Uric_acid)	119,806	1	,605	301	197,930	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The following table Number of Cases in each Cluster gives the distribution, so 166 landed on cluster one and 137 on cluster 2 (Table 26).

Table 26: Number of Cases in each Cluster

Number of Cases in each Cluster

Cluster	1	166,000
	2	137,000
Valid		303,000
Missing		,000

It is worth to be noted that the application of K-Means gives the opportunity to save the Cluster Membership Number in a new variable, QCL_1 (Cluster Membership Number), and thus the individuals' classification in their corresponding cluster could be cheeked. Variable QCL_1 could also be used to test whether there is a significant difference between the clusters.

4. Conclusions

Statistical techniques employed for the analysis and interpretation of data that are widely used in epidemiology and medicine belong to Multivariate Methods. Biochemical analyses and biomarkers identification utilize these methods to a great extend, which constitute pattern recognition methods and are distinguished into two major categories:

Unsupervised pattern recognition methods

Supervised pattern recognition methods

The first category is based on the principal that there is no a priori information about the membership of the sample examined. PCA also falls under this category, since the Principal Components Analysis is not known beforehand, but ensues from the application of the method. Principal Components are hierarchically calculated.

The second category is based on the principal that there is a priori information about the membership of the sample examined. K-Means and Discriminant Analysis fall under this category. The number of classes is based on which variables will be categorized and known and defined.

With respect to PCA, each individual is assigned a unique score for every Principal Component. With respect to K-Means and Discriminant Analysis each individual belongs only to one group. In the case of Discriminant Analysis (Group centroinds) we also get the discriminant function.

To investigate the primary and secondary objectives, the three aforementioned methods were applied on a cases-controls sample (151-152) with respect to the measurements of 16 bio-indexes which ensued from the biochemical analyses. Cases suffered from kidney disease. For this reason, the interpretation of the data was directed to the bio-indexes which relate to the disease and which are: the Urea, the Creatine and Serum Uric Acid. The objective of the paper was to apply the methods with an educational and not clinical orientation.

The results from the application of the methods have pointed at their differences and similarities but also their complementarity. One can concisely cite that the application of PCA resulted to a data reduction and showed that there are five Principal Components (Latent Variables) which interpret all of the total variability/information of data, as well as their structure. It is worth noting that the third Component emerges as the Component for kidney

function and comprises of variables Urea, Creatinine and Serum Uric Acid, which are also the variable comprising the clinical measurement that show the existence or not of kidney disease. The scatter plots of all combinations of Components showed that the Component of kidney function is indeed the one showing the differentiation of controls from cases, while the scatter plots offered the best visualization of the data.

Discriminant Analysis showed that Urea, Creatinine, and Serum Uric Acid are, indeed, the best predictors with respect to the separation of controls from cases. It offered the potential to assess and evaluate the accuracy of the Predicted Group Membership, which was verified. In addition, Discriminant Analysis evaluated also the high sensitivity and high specificity, where high values were ascertained for both. It also offered the potential to determine the Discriminant Function. Finally, the K-Means which was applied with respect to only the three variables, Urea, Creatinine, and Serum Uric Acid has shown that they satisfactorily separate the controls from cases and, among others, classified each individual.

It could be noted at this point that other similarities and differences between the methods could also be cited, such as, for example, the role of loadings for PCA, etc., but the scope of this paper does not permit us to undertake this task.

However, the posterior application of Discriminant on PCAs and of K-Means on PCAs is recommended, so that a better visualization of the clusters may be obtained.

5. References

Batsidis, A., & Zografos, K. (2006). Dicrimination of observations into one of two elliptic populations based on monotone training samples. *Metrica*, 64, 221-241.

Batsidis, A., & Zografos, K. (2011). Errors of misclassification in discrimination of dimensional coherent elliptic random field observations. *Statistica eerlandica,* 65, 446-461.

Cattel, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-275.

Chatsidis, D. (2015) Predicting business failure of industrial firms in Greece using discriminant analysis and financial statements data. Dissertation Thesis. Greek Open University.

Dafermos, B. (2013). *Factor Analysis*, Thessaloniki: Ziti.

Eckart, C., & Yong, G. (1936). The approximation of the matrix by another of lower rank. *Psychometrika,* 1, 183-187.

Field. A. (2009). *Discovering statistic using SPSS*. SAGE Publications India Pvt Ltd.

Hair, J. F., Anderson, R.E, Tatham, R.L., & Black, W. C. (2005). *Multivariate data analysis*. New Jersey: Prentice-Hall.

Hartigan, J., & Wang, M. (1979). A K-menas clustering algorithm, *Applied Statistics*, 28, 100-108.

Howitt, D., & Cramer, D. (2010). Applications with SPSS 16. Athens: Kleidarithos.

Jollife, I. T (1972). Discarding variables in the principal components analysis, I: Artificial data. *Applied Statistics*, 1, 57-93.

Jollife, I. T (1986). *Principal components analysis*, New York: Springer.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurements*, 20, 141-151.

Kaiser, H. F. (1974). In order of factorial simplicity, *Psychometrika*, 39, 31-36.

Karlis, D. (2005). *Multivariate Statistical Analysis*. Athens: Stamoulis.

Lloyd, S. (1957). *Least squares quantization in pcm.* Bell Telephone Laboratories Paper, Marray Hill.

MacQueen, S. (1967). Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symposium*, 281-295.

Santhanalakshi, R., & Alagarsamy, K. *International J. Comp. Tech. Aplli*.2(1), 193-198. Retr. http://www.ijcta.com/documents/volumes/vol2issue1/ijcta2011020118.pdf.