MULTIVARIATE ANALYSIS APPLICATION IN CARDIOVASCULAR DISEASE

RIGENA SEMA[1], ERALDA GJIKA (DHAMO)[2], LULE HALLAÇI (BASHA)[3]

[1]Department of Mathematics, Faculty of Natural Science, Tirana University

[2]Department of Applied Mathematics, Faculty of Natural Science, Tirana University

[3]Department of Applied Mathematics, Faculty of Natural Science, Tirana University

e-mail**:** rigena.sema@fshn.edu.al

Abstract

Multivariate analysis is a powerful statistical methodology used to identify correlations between random variables when dealing with large dimension of data. It relay's on basic and more sophisticated properties of matrix algebra. Multivariate analysis statistical test enables us to pre-process and analyze the distribution of variables which we will consider in our study. Normality assumption are checked using different numerical and statistical test. In this study we present an approach of how this statistical test may be used in clinical research. We consider a dataset of 150 individuals observed in a cardiology ward. Their physical and clinical signs have been observed and used to further select the most important variables which will then be used to build a statistical model that will assist doctors to quickly and accurately diagnose the risk of a patient needing an intervention. The analysis proposed here may be used as an initial tool for other studies in the future which may increase the accuracy by enlarging the number of observations and variables.

Key words: MVN distribution, bivariate distribution, modeling, cardiovascular diseases, clinical research, PCA.

Përmbledhje

Analiza shumëpërmasore është një metodë e rëndësishme statistikore e përdorur për të identifikuar korrelacionin midis variablave të rastësishëm kur kemi të bëjmë me dimensione të mëdha të të dhënave. Ajo lidhet me vetitë themelore dhe më të sofistikuara të algjebrës së matricës. Testi statistikor i analizës shumëpërmasore na mundëson të-përpunojmë dhe analizojmë shpërndarjen e variablave që do të shqyrtojmë në studimin tonë. Supozimi i normalitetit kontrollohet duke përdorur teste të ndryshme numerike dhe statistikore. Në këtë studim ne paraqesim një qasje se si ky test statistikor mund të përdoret në kërkimin klinik. Ne shqyrtojmë një grup të dhënash prej 150 individësh të vëzhguar në një pavijon të kardiologjisë. Shenjat e tyre fizike dhe klinike janë vëzhguar dhe përdorur për të zgjedhur më tej variablat më të rëndësishëm, të cilët më pas do të përdoren për të ndërtuar një model statistikor që do të ndihmojë mjekët të diagnostikojnë shpejt dhe saktë rrezikun e një pacienti që ka nevojë për një ndërhyrje. Analiza e propozuar këtu mund të përdoret si një mjet fillestar për studime të tjera në të ardhmen, të cilat mund të rrisin saktësinë duke zgjeruar numrin e vëzhgimeve dhe variablave.

Fjalë kyçe: Shpërndarja MVN, shpërndarje dy përmasore, modelime, sëmundje kardiovaskulare, të dhëna klinike, PCA.

Introduction

Cardiovascular disease (CVD) depends on many risks factor. It is one of the diseases where the mortality is very high. Patients with type 2 diabetes are the most at risk by (CVD) (Heart Diseases and Stroke Statistics, 2021; Saxon DR & Rasouli *et al.* (2018). To have the best possible description and interpretation of data we can use multivariable methods. Studies has shown multivariate analysis of risk factors for coronary heart diseases (Jeanne Truett, 1967; Cerutti,, 2009; Gianpaolo Reboldi, 2013; Carlo Ricciardi, *et al*. 2020). Multivariate Analysis (MVA) is deal with data on more than one variable. Its techniques allow that more than two variables to be analyzed at once. These variables often are described by their joint probability distribution (Katz & Mitchell, 2003). The identification of patterns in clinical data in patients with chest discomfort is done in medicine using multivariate modeling (Oliver Hirsch *et a*l., 2011). Multidimensional scaling (MDS) and multiple correspondence analysis (MCA) have both been used by the writers. In the analysis, six factors were found. They demonstrated that there is no consistent relationship between the signs and symptoms of chest discomfort. Chest discomfort is thus a diverse clinical group.

Another study which uses multivariate analyses is Akash Mishra, *et. al.*, (2021). In this investigation, various outcome characteristics were associated using clinical data from the ACCOARD lipid trial. In order to statistically decide whether to accept or reject the hypothesis, they used multivariate and univariate analysis. After that, they collated the data. The multivariate technique has the most potential to alter statistical conclusions and produce exact findings in medicine, according to the authors' research.

Multivariate analysis is not used only in medicine. So, Böhm, K., Smidt, E., *et al*. (2013) have studied the weather, which, like other natural phenomena, depends on several factors. By using multivariate analysis, it is shown that PCA and PLS-R were the most applied methods in waste management. David Nez-Alonso, Luis Vicente, *et al*. (2019) have studied and analyzed the air quality of the city of Madrid (Spain). They used various multivariate analysis methods and concluded that the results of PCA, CA, and ordinary kriging applied to air pollution data may be useful in assessing air pollution while also providing a deeper understanding of the major mechanisms involved. Mikhailov & Tupicina, *et al*. (2007) have proved that multivariate data analysis is a useful tool for ecological monitoring. They discovered that chemometric methods enabled them to investigate the structure of waste disposal in specific areas. In economy multivariate and univariate nonlinear prediction methods may be used, including polynomial and BP neural network. Junhai Ma, & Lixia Liu *et al*., (2008) have shown that multivariable nonlinear prediction method was useful to study stock price prediction in emerging markets, especially to Shanghai stock market.

One of the most important distributions used is multivariate normal (MVN) distribution and bivariate normal (BVN) distribution. Yue *et al*. (1999) has

applied BVN to analyze the joint distributions of two correlated random variables flood peaks and volumes as well as flood volumes and durations. They showed that BVN model can contribute to solving problems of hydrological engineering design and management. Gurprit Grover & Alka Sabharwal *et al*., (2014) have applied to model to estimate the duration of diabetes. They applied MVN model for three random variables Duration of diabetes (t), Serum Creatinine (SrCr) and Fasting blood glucose (FBG) The second model BVN was applied for only two random variables Duration of diabetes (t) and SrCr. They showed that the MVN model is preferred over BVN model as more the information the better would be the estimates. Zhen Xue & Liangliang Zhang *et al*., (2021) by using centering matrix and based on its properties which consisted of the quadratic form, spectral decomposition, null spaces, projection Kronecker square have explored the role of centering matrix in the principal component analysis (PCA) and regression analysis theory. It is show that the sum of deviation squares is expressed by the quadratic form with the centering matrix as the kernel matrix.

In this paper we study cardiovascular diseases by using multivariate analysis. The main objectives of multivariate analysis are exploratory data analysis, classification and parameter prediction. To have a clear view for these methods it is important to check and describe the dependence of the random variables. We will describe the random variables by using matrix algebra, vectors and multivariate and bivariate distributions. Another purpose is to understand, among others, which statistical models can be used to classify a patient as a low, medium and high risk for stenosis intervention.

The paper is organized in three sections. In the first section we present the methodology, in the second section we present the application of model in real data and at the end, we present our findings and conclusions with open discussions for future studies.

1. Methodology

a. Multivariate Normal (MVN) distribution

In this section we give some notions and present some auxiliary results that will be used throughout our methodological developments in this paper. For the notion and result we are referring Rencher *et al*., (2002) and Johson and Wichern *et al*., (2007) and relay's on basic and more sophisticated properties of matrix algebra.

A random vector $\underset{\sim}{X} = (x_1, x_2, \dots, x_p)^T$ is said to have a multivariate normal distribution (MVN) if $x_1, x_2, \dots, x_p$ have join density of the form

$$f_{\underset{\sim}{X}}\left(\underset{\sim}{X}\right) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}\left(\underset{\sim}{X}-\mu\right)^T \Sigma^{-1}\left(\underset{\sim}{X}-\mu\right)\right\}, \underset{\sim}{x} \in \mathbb{R}^p \qquad (1.1)$$

where $\mu \in \mathbb{R}^p$ is a constant vector and $\Sigma_{p\times p}$ is a constant positive definite matrix. The p-dimensional MVN is denoted as $N_p(\mu, \Sigma)$. A MVN distribution

is fully specified by its mean and covariance matrix or different we can stay that:

If $\underset{\sim}{X} \sim N_p(\mu, \Sigma)$ then $E\left(\underset{\sim}{X}\right) = \mu$ and $cov\left(\underset{\sim}{X}\right) = \Sigma$.

If $\underset{\sim}{X} \sim N_p(\mu, \Sigma)$ then the moment generating function of $\underset{\sim}{X}$ is given by

$$M_{\underset{\sim}{X}}(t) = E\left(e^{t^T \underset{\sim}{X}}\right) = exp\left\{t^T u + \frac{1}{2} t^T \Sigma t\right\}, t \in \mathbb{R}^p$$

If $\mu = 0$ and $\Sigma = I_p$, then $N_p(0, I_p)$ is called standard MVN. If $\underset{\sim}{X} \sim N_p(0, I_p)$, then any two entries, X_i, X_j for $i \neq j$, of $\underset{\sim}{X}$ satisfy $cov(x_i, x_j) = 0$ and $Var(x_i) = 1$ for all $i = 1,2, \dots, p$.

If $X_1, X_2, \dots, X_p$ are IID $N(0,1)$, then from independence we have that the join density of $X_1, X_2, \dots, X_p$ is

$$f_{\underset{\sim}{X}}\left(\underset{\sim}{X}\right) = \prod_{i=1}^{p} f_{X_i}(X_i) = \prod_{i=1}^{p} \left\{\frac{1}{(2\pi)^{1/2}} exp\left(-\frac{1}{2} {X_i}^2\right)\right\} =$$

$$= \frac{1}{(2\pi)^{p/2}} exp\left\{-\frac{1}{2}\sum_{i=1}^{p} {X_i}^2\right\} \qquad (1.2)$$

but we have that $|I_p| = 1$ and $\sum_{i=1}^{p} {x_i}^2 = \left(\underset{\sim}{X} - 0\right)^T {I_p}^{-1}\left(\underset{\sim}{X} - 0\right)$ so

$$f_{\underset{\sim}{X}}\left(\underset{\sim}{X}\right) = \frac{1}{(2\pi)^{p/2}|I_p|^{1/2}} exp\left\{-\frac{1}{2}\left(\underset{\sim}{X} - 0\right)^T {I_p}^{-1}\left(\underset{\sim}{X} - 0\right)\right\}$$

which is that $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \sim N_p(0, I_p)$.

We present some of the properties of a MVN random vector:

1.If $\underset{\sim}{X} \sim N_p(\mu, \Sigma)$ with PD (positive definite) Σ and $v \in \mathbb{R}^p$ is a non-zero constant vector then the vector $y = v^T \underset{\sim}{X}$ is an univariate random variable and for any $t \in \mathbb{R}$ we have

$$E\{\exp(ty)\} = E\left\{\exp\left(u^T \underset{\sim}{X}\right)\right\} \text{ where } u = tv$$

We noticed that $E\left\{\exp\left(u^T \underset{\sim}{X}\right)\right\}$ is a MGF of vector $\underset{\sim}{X}$ in u, but $\underset{\sim}{X} \sim N_p(\mu, \Sigma)$ so

$$M_{\underset{\sim}{X}}(u) = E\left(e^{u^T \underset{\sim}{X}}\right) = exp\left\{u^T u + \frac{1}{2} u^T \Sigma u\right\}, u \in \mathbb{R}^p.$$

Therefore

$$E\{\exp(ty)\} = M_{\underset{\sim}{X}}(u) = exp\left\{tv^T u + \tfrac{1}{2} t^2 (u^T \Sigma u)\right\}, t \in \mathbb{R},$$

that is MGF of $N(v^T u, u^T \Sigma v)$

so

$$v^T \underset{\sim}{X} \sim N(v^T u, v^T \Sigma v).$$

2.If for every non-zero constant p-vector v, $v^T \underset{\sim}{X}$ has a normal distribution (with positive variance), by using MGF of MVN and univariate normal distribution we have that MGF of vector $y = v^T \underset{\sim}{X}$ is

$$M_y(t) = exp\left\{\mu_y t + \frac{1}{2}\sigma_y^2 t^2\right\}, t \in \mathbb{R}$$

Note that $M_y(1) = M_{\underset{\sim}{X}}(v), \mu_y = v^T \mu_{\underset{\sim}{X}}$. Further we note that $Var\left(v^T \underset{\sim}{X}\right) > 0$ for all non-zero $v \in \mathbb{R}^p$ and so $Var\left(v^T \underset{\sim}{X}\right) = v^T \Sigma_x v > 0$ and Σ_x is a PD matrix. Therefore, the MGF of an MVN distribution is

$$M_{\underset{\sim}{X}}(v) = exp\left\{u^T \mu_x + \frac{1}{2} v^T \Sigma_x v\right\}$$

where $v \in \mathbb{R}^p$. Hence $\underset{\sim}{X}$ has a MVN distribution.

3. Let $A_{g\times p}$ be a constant matrix with full row rank, i.e. $rank(A) = g$, $\underset{\sim}{X} \sim N_p(\mu, \Sigma)$ for a PD Σ, and $b \in \mathbb{R}^g$ a constant vector. For the vector $\underset{g\times 1}{Y} = A\underset{\sim}{X} + b$ and for every vector $v \in \mathbb{R}^g$ we have that

$$v^T \underset{g\times 1}{Y} = \left(v^T A \underset{\sim}{X}\right) + v^T b = u^T \underset{\sim}{X} + c.$$

Since matrix A has full row rank and v is a non-zero vector, then u is non-zero vector and $u^T \underset{\sim}{Y}$ is normal and so the vector $v\underset{\sim}{Y}$ is normal. From the above results we have that $\underset{g\times 1}{Y} = A\underset{\sim}{X} + b$ is MVN and

$$A\underset{\sim}{X} + b \sim N_g(Au + b, A\Sigma A^T) \qquad (1.3)$$

4.Let $\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X_1} \\ \underset{\sim}{X_2} \end{pmatrix} \sim N_p(\mu, \Sigma)$, where $\underset{\sim}{X_1} \in \mathbb{R}^p$ and $\underset{\sim}{X_2} \in \mathbb{R}^{q-p}$, and

$A = \begin{pmatrix} I_p & 0_{p\times(q-p)} \end{pmatrix}$. We observe that matrix A has full row rank, then from properties 3 we have that $\underset{\sim}{X_1} = A\underset{\sim}{X}$ is a MVN and

$$E\left(\underset{\sim}{X_1}\right) = E\left(A\underset{\sim}{X}\right) = AE\left(\underset{\sim}{X}\right) = \begin{pmatrix} I_p & 0_{p\times(q-p)} \end{pmatrix}\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = I_p\mu_1 = \mu_1$$

$$cov\left(\underset{\sim}{X}_1\right) = cov\left(A\,\underset{\sim}{X}\right) = Acov\left(\underset{\sim}{X}\right)A^T =$$
$$= \left(I_p \;\; 0_{p\times(q-p)}\right)\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix} I_p \\ 0_{p\times(q-p)}\end{pmatrix} =$$
$$= \left(\Sigma_{11} \;\; \Sigma_{12}\right)\begin{pmatrix} I_p \\ 0_{p\times(q-p)}\end{pmatrix} = \Sigma_{11}$$

In the same manner we have that $E\left(\underset{\sim}{X}_2\right) = \mu_2$ and $cov\left(\underset{\sim}{X}_1\right) = \Sigma_{22}$. So, every sub vector of $\underset{\sim}{X}$ is normally distributed. The converse is not true in general.

Thus, we can ask: *Under what condition normality (MVN) of* $\underset{\sim}{X}_1$ *and normality of* $\underset{\sim}{X}_2$ *imply joint normality of* $\begin{pmatrix}\underset{\sim}{X}_1\\ \underset{\sim}{X}_2\end{pmatrix}$?

The following results show the relationship of 0 covariance with independence.

Let $\underset{\sim}{X}_1 \in \mathbb{R}^p$ and $\underset{\sim}{X}_2 \in \mathbb{R}^{q-p}$ be two random vectors and $\underset{\sim}{X} = \begin{pmatrix}\underset{\sim}{X}_1\\ \underset{\sim}{X}_2\end{pmatrix}$ has distributed $as\ N_p(\mu, \Sigma)$ with some mean and some PD covariance matrix, denoted $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ and $cov\left(\underset{\sim}{X}_1, \underset{\sim}{X}_2\right) = 0$, then

$$|\Sigma| \neq 0, \Sigma = \begin{pmatrix}\Sigma_{11} & 0\\ 0^T & \Sigma_{22}\end{pmatrix}$$

and the join density of $\underset{\sim}{X} = \begin{pmatrix}\underset{\sim}{X}_1\\ \underset{\sim}{X}_2\end{pmatrix}$ has the form

$$f_{\underset{\sim}{X}}\left(\underset{\sim}{X}\right) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\, exp\left\{-\frac{1}{2}\left(\underset{\sim}{X}-\mu\right)^T\Sigma^{-1}\left(\underset{\sim}{X}-\mu\right)\right\}, \underset{\sim}{x} \in \mathbb{R}^p$$

but Σ is positive definite matrix, then Σ^{-1} exists and $\Sigma^{-1} = \begin{pmatrix}{\Sigma_{11}}^{-1} & 0\\ 0^T & {\Sigma_{22}}^{-1}\end{pmatrix}$.

Since

$$\left(\underset{\sim}{X}-\mu\right) = \begin{pmatrix}\underset{\sim}{X}_1-\mu_1\\ \underset{\sim}{X}_2-\mu_2\end{pmatrix} \text{ and } \left(\underset{\sim}{X}-\mu\right)^T = \begin{pmatrix}\left(\underset{\sim}{X}_1-\mu_1\right)^T & \left(\underset{\sim}{X}_2-\mu_2\right)^T\end{pmatrix}$$

we can write

$$\left(\underset{\sim}{X}-\mu\right)^T\Sigma^{-1}\left(\underset{\sim}{X}-\mu\right) = \begin{pmatrix}\left(\underset{\sim}{X}_1-\mu_1\right)^T & \left(\underset{\sim}{X}_2-\mu_2\right)^T\end{pmatrix}\begin{pmatrix}{\Sigma_{11}}^{-1} & 0\\ 0^T & {\Sigma_{22}}^{-1}\end{pmatrix}\begin{pmatrix}\underset{\sim}{X}_1-\mu_1\\ \underset{\sim}{X}_2-\mu_2\end{pmatrix}$$
$$= \begin{pmatrix}\left(\underset{\sim}{X}_1-\mu_1\right)^T{\Sigma_{11}}^{-1} & \left(\underset{\sim}{X}_2-\mu_2\right)^T{\Sigma_{22}}^{-1}\end{pmatrix}\begin{pmatrix}\underset{\sim}{X}_1-\mu_1\\ \underset{\sim}{X}_2-\mu_2\end{pmatrix} =$$
$$= \left(\underset{\sim}{X}_1-\mu_1\right)^T{\Sigma_{11}}^{-1}\left(\underset{\sim}{X}_1-\mu_1\right) + \left(\underset{\sim}{X}_2-\mu_2\right)^T{\Sigma_{22}}^{-1}\left(\underset{\sim}{X}_2-\mu_2\right)$$

We know that

$$\Sigma_{p\times p} = \begin{pmatrix} \Sigma_{11_{q\times q}} & 0_{q\times(p-q)} \\ 0^T{}_{(p-q)\times q} & \Sigma_{22\,(p-q)\times(p-q)} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0^T & I_{p-q} \end{pmatrix}\begin{pmatrix} I_q & 0 \\ 0^T & \Sigma_{2\times 2} \end{pmatrix}$$

and

$$|\Sigma_{p\times p}| = \left|\begin{pmatrix} \Sigma_{11} & 0 \\ 0^T & I_{p-q} \end{pmatrix}\right| \left|\begin{pmatrix} I_q & 0 \\ 0^T & \Sigma_{2\times 2} \end{pmatrix}\right|$$

From the determinant of a matrix, we have

$$\left|\begin{pmatrix} \Sigma_{11} & 0 \\ 0^T & I_{p-q} \end{pmatrix}\right| = |I_{p-q}||\Sigma_{11}| = |\Sigma_{11}| \text{ and } \left|\begin{pmatrix} I_q & 0 \\ 0^T & \Sigma_{2\times 2} \end{pmatrix}\right| = |I_q||\Sigma_{22}| = |\Sigma_{22}|,$$

thus $|\Sigma_{p\times p}| = |\Sigma_{11}||\Sigma_{22}|$

Thus, the joint density can be written as

$$f_{\underset{\sim}{X}}(\underset{\sim}{X}) = \frac{1}{(2\pi)^{p/2}||\Sigma_{11}||\Sigma_{22}||^{1/2}} exp\left\{-\frac{1}{2}\left(\left(\underset{\sim}{X_1}-\mu_1\right)^T {\Sigma_{11}}^{-1}\left(\underset{\sim}{X_1}-\mu_1\right)+ \left(\underset{\sim}{X_2}-\mu_2\right)^T {\Sigma_{22}}^{-1}(\underset{\sim}{X_2}-\mu_2)\right)\right\}$$

Consequently, the random vector $X_1 \in \mathbb{R}^{q\times 1}$ and $X_2 \in \mathbb{R}^{(p-q)\times 1}$ are independent. The converse of this results is true.

The following results for normal condition are true:

Let $\underset{\sim}{X}_{p\times 1} = \begin{pmatrix} \underset{\sim}{X_1}_{q\times 1} \\ \underset{\sim}{X_2}_{(p-q)\times 1} \end{pmatrix} \sim N_p(\mu.\Sigma)$ with $\mu = \begin{pmatrix} \mu_{1_{q\times 1}} \\ \mu_{2_{(p-q)\times 1}} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11_{q\times q}} & \Sigma_{12\,q\times(p-q)} \\ \Sigma_{21\,(p-q)\times q} & \Sigma_{22(p-q)\times(p-q)} \end{pmatrix}$

Then the conditional distribution of $\underset{\sim}{X_1}$ given that $\underset{\sim}{X_2} = \underset{\sim}{x_2}$ is $N_p(\mu_{1|2}.\Sigma_{1|2})$ where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(\underset{\sim}{x_2}-\mu_2\right)$$

and

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Similarly, $\underset{\sim}{X_2} \mid \underset{\sim}{X_1} = \underset{\sim}{x_1} \sim N_{q-p}(\mu_{2|1}.\Sigma_{2|1})$, where

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}\left(\underset{\sim}{x_1}-\mu_1\right)$$

and

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

b. Bivariate Normal (BVN) distributor

Another case of multivariate normal distribution is the bivariate normal distribution (BVN) which is focused on two normal random variables X and Y with parameters $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ and ρ. These random variables have bivariate normal distribution if their joint PDF is given by

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\} \quad (1.4)$$

where $\mu_X, \mu_Y \in \mathbb{R}, \sigma_X, \sigma_Y > 0$ and $\rho \in (-1 \quad 1)$ are all constant and for $X = x$, Y is normally distributed with

$$E[Y|X = x] = \mu_Y + \rho\sigma_Y\frac{x-\mu_X}{\sigma_X} \text{ and } Var[Y|X = x] = (1-\rho^2)\sigma_Y^2$$

c. PCA method

In data analysis is used the method of principal components analysis (PCA) which consists of dimensionality reduction of a data. Relying on linear and matrix algebra we can identify the *i*-th patient by a vector x_i.The parameters for the *i*-th patient that we evaluate are defined by x_{ji}. Therefore, the following formula is used to calculate the mean of the *m* variables:

$$\mu = \frac{1}{n}(x_1 + x_2 + \cdots x_n)$$

assuming that there are *n* patients.

Let's define a matrix with real elements such that the *i*-th column is $x_i - \mu$

$$A = (x_1 - \mu \quad x_2 - \mu \quad \cdots \quad x_n - \mu)$$

and then the covariance matrix is $B = \frac{1}{n-1}AA^T$. From linear algebra, we know that B is symmetric matrix and consequently orthogonal diagonalization. To determine which of the given variables is most crucial, we must reorder the eigenvalues from greatest to smallest and calculate their associated orthonormal eigenvectors.

2. Application

a. Motivation

Aortic valve stenosis may lead to heart failure. These symptoms include chest pain, fatigue, shortness of breath, swollen ankles and feet, heart murmur or palpitations etc. Often these causes are hidden, and patients do not understand, so they may present themselves in the emergency room. The study includes the observations of 150 patients presented in a cardiology ward at times when these signs became distinct. So, the patients are not on a routine visit. Many quantitative variables are measured among them: Age,

single-nucleotide polymorphisms located in the collagen (COL), triglycerides (TG) level as a high-level risk factor, fasting blood glucose (FBG), glycated hemoglobin (HbA1c) and the target variable coronary artery disease (CAD). The age of the patients included in the study varies from 30 to 82 years old with an average of about 61 years old. Almost half of the patients are less than 61 years old. Fasting blood glucose for these patients goes from 80 to 300, with an average of 145.7. Triglycerides level is on average 161, with a minimum of 10 and a maximum of 600, where half of the patients has a triglycerides level greater than 137.5. Glycated hemoglobin varies from 5 to 12, with mean 6.688, where in 25 percent of patients, this indicator has a value lower than 5.7.

Since some tests that can be performed are expensive such as cardiovascular imaging techniques including: Transthoracic echocardiogram (TTE); Electrocardiogram (ECG); Exercise stress testing; Magnetic resonance imaging (MRI); Cardiac catheterization; Transesophageal echocardiogram (TEE); CT scan doctors are often forced to use rapid external indicators that can be measured quickly and at no cost to the patients presented in emergency room.

The purpose of this paper is to understand, among others, which statistical models can be used to classify a patient into one of the following categories: No- there is no need for stenosis, Single- only 1, Multiple-several stenosis. Also, to understand the importance of quantitative variables in the main variable that is CAD. So, which is the risk of an individual to be affected by cardiovascular disease and to have an intervention for stenosis. Obtaining the correlation coefficient between our quantitative variables we may understand their linear relationship. Observing the correlation matrix below we do not have a clear view of the correlation between our variables. In this case a PCA analysis may help us to analyze which variables may be of interest to investigate and use as support variables in our model of CAD variable.

Table 1. Correlation coefficient

	AGE	FBG	TG	HbA1c	COL
AGE	1	0.020526	-0.07154	0.137691	0.085475
FBG	0.020526	1	0.238245	0.773296	0.136723
TG	-0.07154	0.238245	1	0.341096	0.338351
HbA1c	0.137691	0.773296	0.341096	1	0.149286
COL	0.085475	0.136723	0.338351	0.149286	1

To obtain a clear view of this behavior we may observe the distribution of HbA1c and the CAD values: the highest the value of HbA1c the highest the level of risk this is something we were expected. The group of patients from No–Cad has a normal distribution as observed from the symmetry of the

boxplot. From the boxplot we also observe a similar behavior of HbA1c and FBG together, COL and TG together.

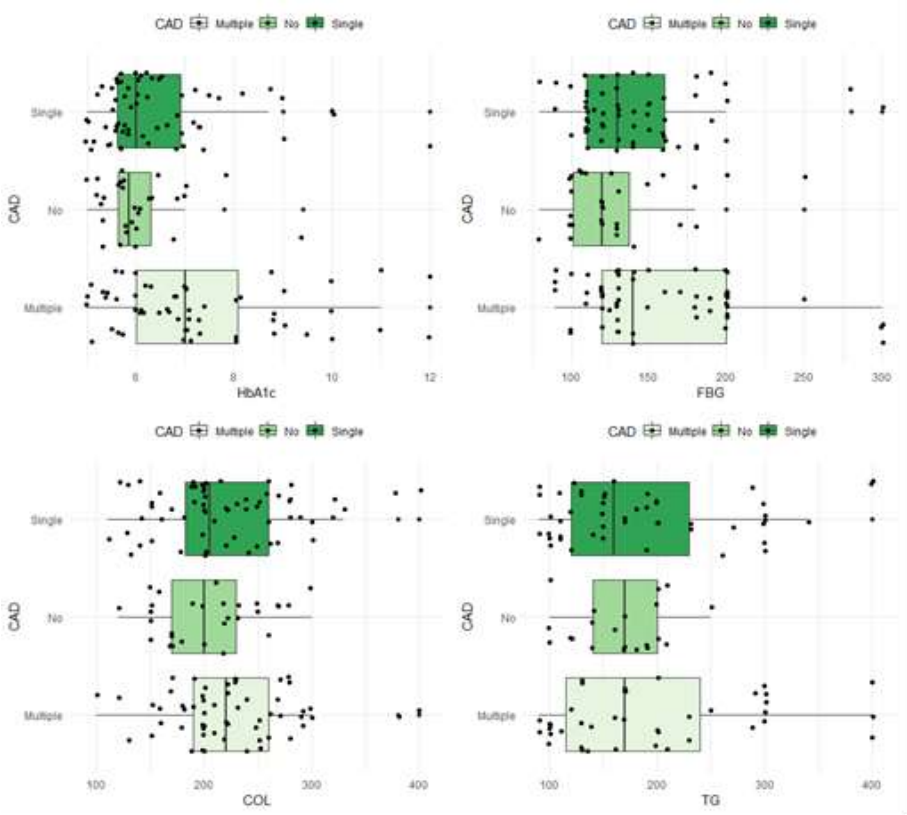


Figure 1. Boxplot of individuals and distributions for HbA1c, FBG, COL, TG according to CAD categories

b. Multivariate analysis of the variables

To analyze normality assumptions, we will use in this study MVN package in R (Korkmaz *et al*., 2014).

Multivariate outliers

Multivariate outliers are the common reason for violating MVN assumption. Thus, before starting to multivariate analysis it is important to check whether the data have multivariate outliers. The MVN package includes two multivariate outlier detection methods which are based on robust Mahalanobis distances as a metric which calculates how far each observation is to the center of joint distribution. The joint distribution can be thought of as the centroid in multivariate space. The two approaches, defined as Mahalanobis distance and adjusted Mahalanobis distance in the R MNV package, detect multivariate outliers and will display: robust Mahalanobis distances and Adjusted Mahalanobis Distance; the 97.5 percent quantile (Q)

of the chi-square distribution; possible outliers. Below are shown these results for the two cases:

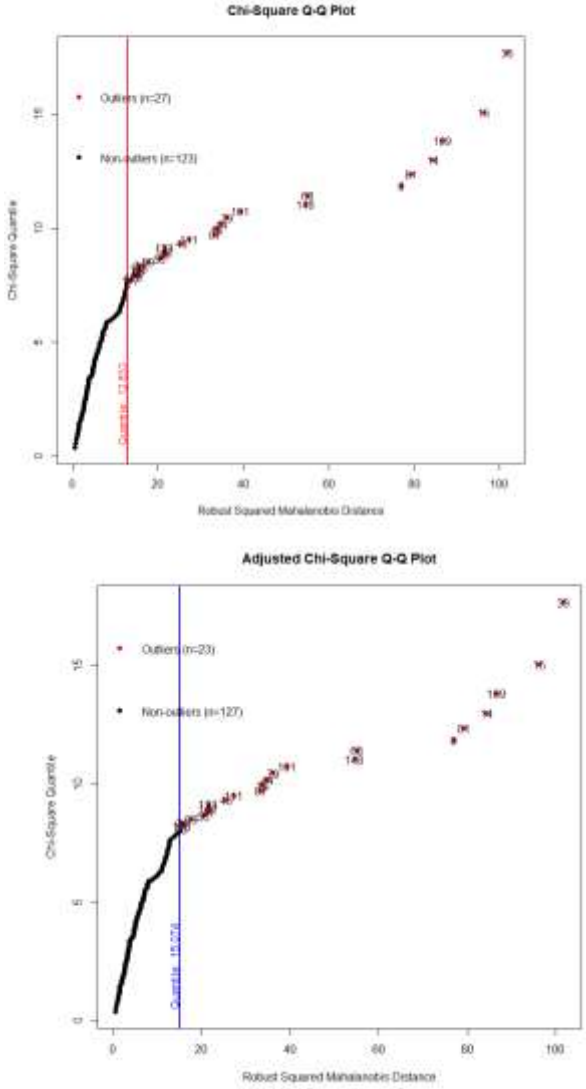


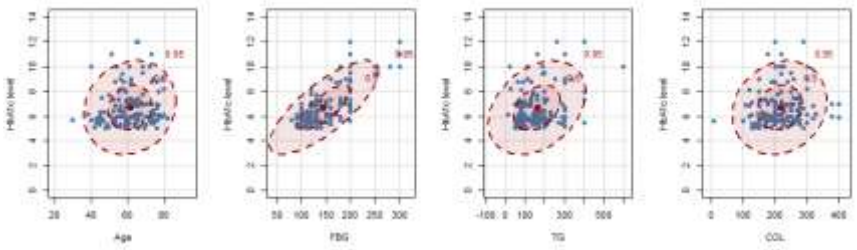
Figure 2. Chi square QQ-plot and adjusted for Mahalanobis distance

We observe from Chi-Square Q-Q plot that 27 individuals are classified as outliers and only 123 are not outliers. And from the adjusted Chi-Square Q-Q plot we still have 23 observations classified as outliers.

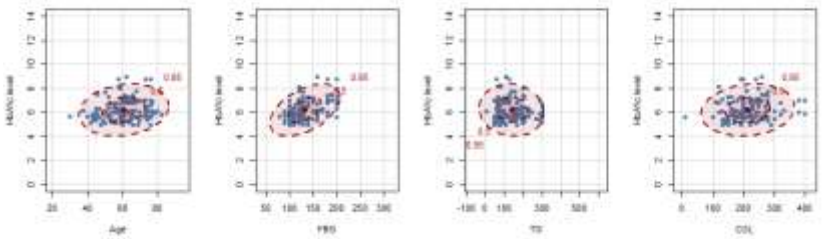
Univariate normal marginal densities are necessary but not a sufficient condition for MVN. Hence, in addition to univariate plots visualizing contour plots will be useful for further investigation of our dataset. Also, the perspective plot is an extension of the univariate probability distribution curve into a three-dimensional probability distribution surface related with

bivariate distributions. This graph is useful for a better understanding of the correlation between two variables. It will be close to a bell-shaped graph if normality is met. The "contour plot" involves the projection of the perspective plot into a 2-dimensional space and this can be used to check multivariate normality assumption. The values will be within the ellipses if normality is met and no ellipse shape if normality is missing.

The contour plot for the four combinations when the main variable is HbA1c show no presence of normality in our data. But when we work with the data after removing outliers the situation changes for some of the pairs, Figure 3.



(a)



(b)

Figure 3. Contour plot before (a) and after (b) removing outlier's

Subset analysis

If we want to check for any of the three groups of CAD, the assumptions for normality we observe that these assumptions are meet only for the group of individuals which had no need of stenosis. The same results are obtained even after removing the outliers based on mahalanobis distance Chi-square QQ-plot. To test the assumptions for normality, we use the Shapiro–Wilk test statistic and Henze-Zirkler's test. The Shapiro–Wilk test statistic is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.1)$$

where $x_{(i)}$ are the ordered sample values and a_i are constants generated from the means, variances and covariance of the order statistics of the data. The following equation gives the Henze-Zirkler Test for normality:

$$HZ = \left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}D_{ij}}\right] - \left[2(1+\beta^2)^{-\frac{d}{2}}\sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}D_i}\right] + \left[n(1+2\beta^2)^{-\frac{d}{2}}\right] \quad (2.2)$$

where D_i is the squared Mahalanobis distance of the ith observation to the centroid and D_{ij} gives the squared Mahalanobis distance between ith and jth observations.

Table 2. Henze-Zirkler test for subsets of CAD variable before and after outlier remove

	Before outlier remove			After outlier remove		
Subset	*HZ*	*p-value*	*MVN*	*HZ*	*p-value*	*MVN*
Multiple	1.0128	0.016249	NO	0.93024	0.09185	YES
Single	1.2433	4.50E-06	NO	0.87060	0.17805	YES
No	0.8669	0.19613	YES	0.93317	0.10208	YES

We have used the adjusted result to filter those observations which are not considered outlier and proceed again with the analysis. After removing the outlier's, we observe that each subset of CAD is now following a multivariate normal distribution. Now we may proceed with different multivariate logistic models to predict our target variable CAD. In this study our aim is to understand the multivariate normality assumptions and investigate the data.

c. PCA and cluster analysis

For a better understanding on the importance of the variables and classification of patients in groups we applied the PCA analysis aiming to get as much information about the dimension of components used to explain variability. The PCA analysis results show us that the first three components explain up to 80% of the variation in our data, Figure 4. and also, the variable PCA plot shows us these three components: HbA1c and FBG are in the same group which has the highest contribution and is part of the 2nd component; TG and COL are in the second group which is part of the 1st

component and AGE is in the third group with the lowest value of the contribution. Combining together the information provided from the three graphs below (PCA analysis) we understand that there are some individuals which are highly affected by these variables and other which are not. Further we need to investigate the effect of any categorical variable and understand how this behavior may be interpreted.

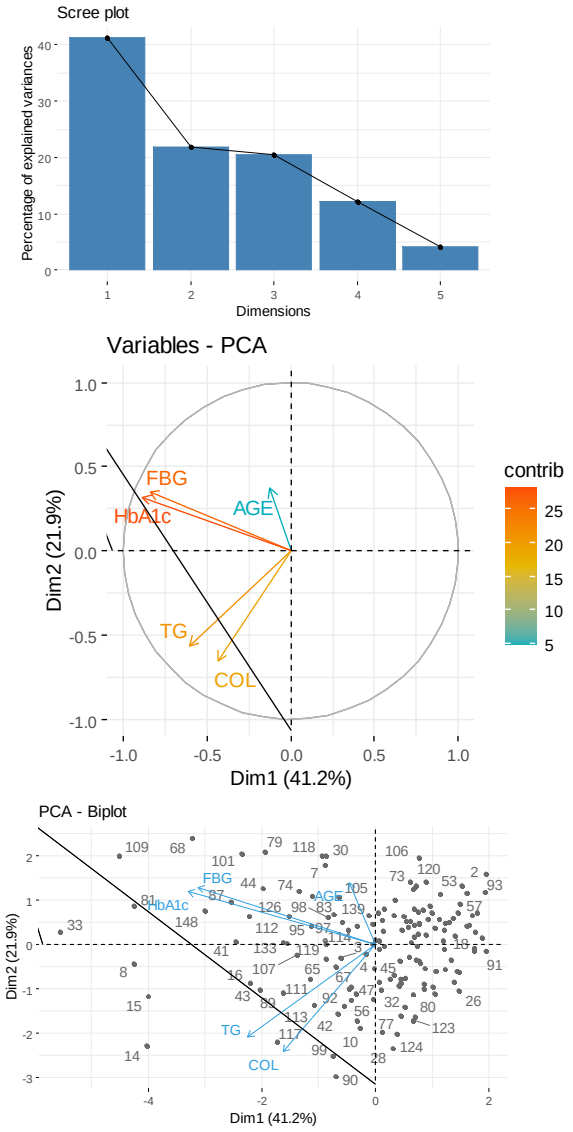


Figure 4. PCA analysis of variable and individuals

Below a cluster correlation plot is created using the quantitative variables and CAD qualitative target variable, Figure 5. What we observe in the correlation graph is: individuals which have 0 stenosis have lower observed values also in quantitative variables. And individuals with multiple stenosis

have higher observed values of quantitative variables compared to those with single stenosis.

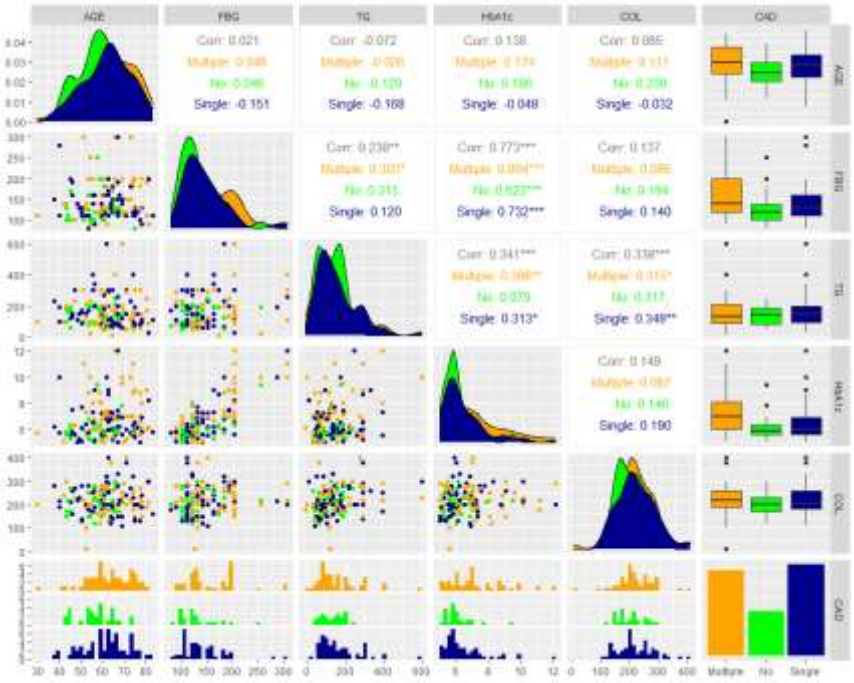
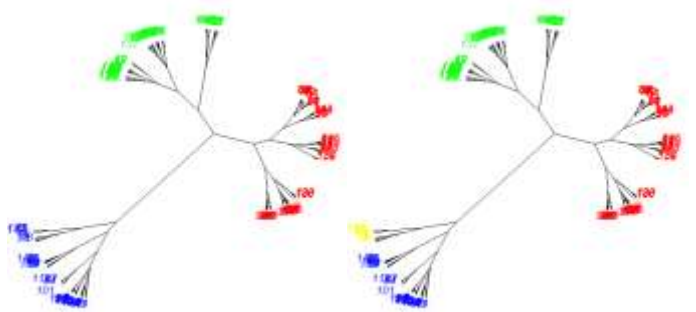


Figure 5. Correlation plot of quantitative variables and CAD

After analyzing the above results from the correlation, we tried also to understand if there is a significant classification of patients in clusters. The below graph shows how our individuals may be organized in three or four clusters. The two graphs show a significant difference when the number of clusters is three, so we may doubt that an individual may be classified in one of the groups which may be related to the three values of CAD variable (No, Single, Multiple), Figure 6. For four clusters the differences are not significant.



(a) three clusters (b) four clusters

Figure 6. Root dendrogram for three (a) and four (b) clusters

After this detailed analysis, other multivariate logistic regression models or similar methods can be examined to gain more information and to model the target variable CAD. The above analysis will help us on determining the most important factors for an accurate classification of a patient being or not in risk for stenosis ranging from no, mild or severe. Using PCA as a dimension reduction methodology will help on proceeding with other supervised or unsupervised methods.

Conclusion

Multivariate analysis is a powerful statistical methodology used to identify correlations between random variables when dealing with large dimension of data. It relay's on basic and more sophisticated properties of matrix algebra. In this study we presented an approach of how this statistical test may be used in clinical research. We consider a dataset of 150 individuals observed in a cardiology ward. Their physical and clinical signs have been observed and used to further select the most important variables which will then be used to build a statistical model that will assist doctors to quickly and accurately diagnose the risk of a patient needing an intervention.

First, we have dealt with outliers using Chi-Square Q-Q plot which suggest 23 observations classified as outliers. After removing the outlier's, we observe that each subset of CAD is now following a multivariate normal distribution. Then we proceed with identifying clusters using PCA. The PCA analysis results showed that the first three components explain up to 80% of the variation in our data. These three components are organized as follow: HbA1c and FBG are in the same group which has the highest contribution and is part of the 2^{nd} component; TG and COL are in the second group which is part of the 1^{st} component and AGE is in the third group with the lowest value of the contribution.

A correlation plot showed individuals which have no stenosis intervention have lower observed values also in quantitative variables and individuals with multiple stenosis have higher observed values of quantitative variables compared to those with single stenosis. The analysis proposed here may be used as an initial tool for other studies in the future which may increase the accuracy by enlarging the number of observations and variables. After this detailed analysis using multivariate distribution tests, other multivariate logistic regression models or similar methods can be examined to gain more information and to model the target variable CAD. This analysis helped us on determining the most important factors for an accurate classification of a patient being or not in risk for stenosis ranging from no, mild or severe. Using PCA as a dimension reduction methodology helped us on proceeding with other supervised or unsupervised methods.

References

Heart Diseases and Stroke Statistics— 2021 Update (2021): A Report from the American Heart Association, Volume 143, Issue 8: Pg254-743, https://doi.org/10.1161/CIR.0000000000000950

Alvin C. Rencher (2002): Methods of multivariate analysis; Second edition, J. Wiley

David Núñez-Alonso., Luis Vicente Pérez-Arribas., Sadia Manzoor., Jorge O. Cáceres (2019): "Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region", Journal of Analytical Methods in Chemistry, vol. 2019, Article ID 9753927, 9 pages https://doi.org/10.1155/2019/9753927

Böhm, K., Smidt, E., & Tintner, J. (2013): Application of Multivariate Data Analyses in Waste Management. In L. a. de Freitas, & A. a. B. R. de Freitas (Eds.), Multivariate Analysis in Management, Engineering and the Sciences. Intech Open. https://doi.org/10.5772/53975

Cerutti, S.; Hoyer, D.: Voss A., (2009): Multiscale, multiorgan and multivariate complexity analyses of cardiovascular regulation, In Philosophical Transactions Mathematical Physical & Engineering Sciences, volume 367, issue 1892.

Grover G., Sabharwal A., Mittal J., (2014):" Applications of Multivariate and Bivariate Normal Distributions to Estimate Duration of Diabetes"; International Journal of Statistics and Applications; 4(1):pg 46-57

Gabrielsson J., Lindberg N.O., Lundstedt T., (2002): Multivariate methods in pharmaceutical applications. In Journal of Chemometrics; volume 16, issue 3

Hirsch O., Bösner S., Hüllermeier E., Senge R., Dembczynski K., Donner-Banzhoff R ,(2011): "Multivariate modeling to identify patterns in clinical data: the example of chest pain"; In BMC Medical Research Methodology, volume 11, issue 1

Junhai Ma., Lixia Liu (2008): "Multivariate Nonlinear Analysis and Prediction of Shanghai Stock Market", Discrete Dynamics in Nature and Society, Article ID 526734, 8 pages, https://doi.org/10.1155/2008/526734

Korkmaz S, Goksuluk D, Zararsiz G.(2014): MVN, An R Package for Assessing Multivariate Normality. In the R Journal. 6(2):151-162; http://www.biosoft.hacettepe.edu.tr/MVN/

http://cran.nexr.com/web/packages/MVN/MVN.pdf (per R command)

Mishra K., K.T., Harichandrakumar, Binu VS., Satheesh S., Sreekumaran N (2021): Multivariate approach in analyzing medical data with correlated multiple outcomes: An exploration using ACCORD trial data; Clinical Epidemiology and Global Health, volume 11.

Mikhailov, E. V., Tupicina, O. V., Bykov, D. E., Chertes, K. L., Rodionova, O. Y., & Pomerantsev, A. L. (2007): Ecological assessment of landfills with multivariate analysis—A feasibility study.Chemometrics and intelligent laboratory systems, 88(1), 3-10

Mitchell H. K (2003): Multivariable Analysis: A Primer for Readers of Medical Research, In Annals of Internal Medicine, volume 138, issue 8

Richard A. Johson and Dean W.Wichern (2007): Applied Multivariate Statistical Analysis; Sixth edition

Reboldi G., Angeli F., Verdecchia P, (2013): Multivariable Analysis in Cerebrovascular; In Cerebrovascular Diseases, volume 35, issue 2