

MEGI

Master Degree Program in
Statistics and Information Management

CONTRIBUTION OF DISCRIMINANT ANALYSIS IN THE CLASSIFICATION OF ANXIETY DISORDERS

Bruno Krithinas Guedes

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CONTRIBUTION OF DISCRIMINANT ANALYSIS IN THE CLASSIFICATION OF ANXIETY DISORDERS

By

Bruno Krithinas Guedes

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Information Analysis and Management

Supervisor: Paulo Jorge Mota de Pinho Gomes, PhD

July 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Bruno Krithinas Guedes

Lisbon, 12th of July of 2023

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Professor Paulo Gomes, for all the knowledge shared, motivation, and guidance throughout these past months. Your expertise has played a crucial role in my development and the realization of this dissertation. To all other professors who have been an influence on my academic and personal growth, I also express my gratitude.

To Rafaela, for always encouraging me and being a source of inspiration and unconditional love. There will never be enough words to thank you and what you have done for me.

To my brother, for being my role model.

To my mother and father, for always being by my side and encouraging me never to give up, and to always pursue my dreams.

To my friends, for being a source of joy and true friendship, many years follow.

Last but not least, I want to thank my family for being a source of strength and a safe place I can always rely on.

ABSTRACT

In recent decades, the importance given to the treatment of mental disorders (MDs) has increased. As some MDs can be prevented if diagnosis occurs in the early stages of the disease, this thesis aims essentially to revisit some statistical discriminant methods emphasizing their potential role to classify patients into different MD categories. Three groups (nervous, psychotic, and healthy) of fifty individuals each were analyzed with the purpose of best separating the three groups. Factorial discriminant analysis (FDA) was used in an initial set of thirty variables, obtaining a misclassification rate of 4.(6)% in the training set and 10.(6)% in the testing set. Each variable was analyzed over its contribution to the discrimination of the groups under matter in which, succeeding the first FDA performed, the correlation between each variable and each discriminant function (DF) was examined, the partial F-test was made for every variable, and the standardized DF coefficients were analyzed, where fifteen variables were selected to take part in a new analysis. The new FDA model obtained a misclassification apparent rate of 10% and 14.(6)% in the training test and testing test, respectively. This model was compared to a stepwise discriminant method of eighteen variables, where it was obtained a misclassification apparent rate of 10% in the training set and 16% in the testing set. The reduced model is then recommended over the two other models, as it is more adequate for the main purpose of this thesis, as its complexity is lower when compared to the full model, and also outperformed the stepwise discrimination model. In conclusion, FDA has proven itself to be a potential good method to diagnose patients considered in this study.

KEYWORDS

Mental disorders; Nervous; Psychotic; Principal Component Analysis; Factorial Discriminant Analysis

Sustainable Development Goals (SGD):



INDEX

1. Introduction	1
1.1. Objectives of the Study	1
2. Literature review	3
2.1. Discriminant Analysis.....	3
2.2. Problematic Context.....	3
3. Methodology	5
3.1. Data	5
3.1.1. Data Collection	8
3.2. Statistical Analysis	9
3.2.1. Introduction.....	9
3.2.2. Factorial Discriminant Analysis	9
3.2.3. Probabilistic Context	12
3.3. Contribution Of Variables.....	13
3.4. Stepwise Discriminant Analysis	14
4. Results and Discussion.....	16
4.1. Exploratory Statistical Analysis.....	16
4.2. Factorial Discriminant Analysis.....	17
4.3. Reduced Model.....	25
4.4. Stepwise Discrimination	28
4.5. Evaluating Models	32
5. Conclusions And Future Works	34
Bibliographical References.....	35
Appendix A	38
Annexes	48
A) Henze-Zirkler Test.....	48
B) Likelihood Ratio Test	48
C) Box M Test.....	48
D) Partial F-Values.....	49
E) Standardized Coefficients.....	49
F) Stepwise Discriminant Analysis	49
G) Variance Inflation Factors.....	50

LIST OF FIGURES

Figure 4.1 – Representation of all thirty variables in the first factorial principal plane generated by the two discriminant functions.....	18
Figure 4.2 – Biplot representing a multiple of the coefficients for all standardized covariates	20
Figure 4.3 – Representation of the proposed variable selection (15 variables) in the first factorial principal plane generated by the two discriminant functions	25
Figure 4.4 – Biplot representing a multiple of the coefficients for the fifteen standardized covariates.....	27
Figure 4.5 – Representation of the eighteen variables obtained from stepwise discrimination, represented in an orthogonal axis generated by the two discriminant functions	30
Figure 4.6 – Biplot representing a multiple of the coefficients for the eighteen standardized covariates.....	31

LIST OF TABLES

Table 3.1 – List of variables to be studied	5
Table 3.2 – Mean values of each variable for each group	6
Table 3.3 – Standard deviation values of each variable for each group	7
Table 4.1 – Standardized scoring coefficients of the thirty variables in each discriminant function	19
Table 4.2 – Classification count in the training set with thirty variables	21
Table 4.3 – Classification count in the testing set with thirty variables	21
Table 4.4 – Partial F test' statistic and p-value performed on all thirty variables	22
Table 4.5 – Correlation of each of the thirty variables to each discriminant function	23
Table 4.6 – Standardized scoring coefficients of the fifteen variables in each discriminant function	26
Table 4.7 – Classification count in the training set with fifteen variables	28
Table 4.8 – Classification count in the testing set with fifteen variables	28
Table 4.9 – Tolerance and VIF values for each of the thirty variables	28
Table 4.10 – Classification count in the training set with eighteen variables	32
Table 4.11 – Classification count in the testing set with eighteen variables	32

LIST OF ABBREVIATIONS AND ACRONYMS

AD	Anxiety disorder
DA	Discriminant Analysis
DF	Discriminant Function
FDA	Factorial Discriminant Analysis
GAD	Generalized Anxiety Disorder
MD	Mental disorder
TND	Truncated Normal Distribution
VIF	Variance Inflation Factors

1. INTRODUCTION

Mental disorders (MDs) are conditions that can disrupt a person's behavior, well-being, and emotional state. When left untreated or misdiagnosed, they can have tragic consequences, including self-harm and suicide. Research has shown that people with mental health issues have a higher mortality rate compared to those who have not (Gissler et al., 2013; Walker et al., 2015).

According to the World Health Organization (2022), MDs include anxiety disorders (ADs), post-traumatic stress disorder, disruptive behavior disorders, bipolar disorder, dissocial disorders, depression, schizophrenia, neurodevelopmental disorders, and eating disorders. Research suggests that ADs are particularly common, with estimates of their global prevalence ranging from 3.8% to 25% (Remes et al., 2016). A systematic review of 87 studies from 44 countries conducted by Baxter et al. (2013) found that the prevalence of ADs ranged from 0.9% to 28.3%, indicating that both estimates fall within a similar range.

ADs are one of many mental illnesses that harm the quality of life, which Olatunji et al. (2007) defined as "aspects of life that make it fulfilling and worthwhile and extends beyond anxiety symptoms themselves to include patients' subjective well-being and life satisfaction". People with this type of disorder may experience panic attacks, changes in appetite, sweats, and palpitations, among other symptoms (Anxiety Care UK, 2017). According to Batelaan et al. (2016), individuals with anxiety are 26% more susceptible to the risk of developing coronary heart disease and almost 50% more susceptible to the risk of cardiac death.

Accurate and early diagnosis of MDs is crucial for effective treatment and can help prevent people from experiencing more severe states of mind. However, diagnosing them can be challenging, and misdiagnosis can lead to inappropriate treatment and continued suffering (Nasrallah, 2015). The use of standardized assessment tools, such as questionnaires and interviews, can also help mental health professionals accurately diagnose MDs. Symptoms of mental illness can often overlap, making it difficult to differentiate between disorders and leading to misdiagnosis (Milberger et al., 1995; Renard et al., 2017). Early intervention is important for the treatment of mental health conditions, as precocious treatment can help reduce the severity of secondary effects of psychopathologies (Kessler et al., 2007). Patients can also play a role in their diagnosis by providing accurate and detailed information about their symptoms and medical history.

1.1.OBJECTIVES OF THE STUDY

This study aims to use Factorial Discriminant Analysis (FDA) to develop rules for classifying new individuals into one of three categories: nervous, psychotic, or healthy. The study aims to identify the minimum number of variables that best discriminate the three categories. Having the critical quantitative variables, it is possible to create discriminant functions (DFs) based on them. These functions can be used to classify future individuals into the appropriate group.

Taking this into consideration, the objectives of this thesis will be:

- Identifying the variables that best discriminate the three classes under study;
- Identify the number of DFs needed to represent the differences among the groups;
- Creating a rule to classify future observations into one of the three groups;

- Develop a misclassification apparent rate matrix.

With these objectives in mind, the expected contributions will be to help identify individuals who are likely to have an AD based on a set of relevant predictors, as well as help healthcare professionals better understand how to approach patients, further explored in section 2.2 of this paper.

2. LITERATURE REVIEW

ADs are a specific group of MDs, commonly characterized by constant fear, worry, and anxiety. Certain events in life can trigger this condition, such as health problems, money problems, and work-related issues, among others (National Institute of Mental Health, 2022).

To diagnose patients who are prone to ADs, it is proposed to use discriminant analysis (DA) as a decisional method to predict a new individual to one of the predefined groups.

2.1. DISCRIMINANT ANALYSIS

DA is the most suitable method when the problem can be reduced to determine how groups differ concerning the predictive variables. The main objective of this method is to identify an equation or rule that can effectively classify future observations into one of the groups (Sharma, 1995).

As examples of DA applications, the study of Ma et al. (2016) used a stepwise discriminant procedure to determine the geographic origins of a famous Chinese green tea - *Dongting Biluochun*. The authors of the study reached astonishing results, reaching a perfect score on the classification of *Dongting Biluochun* in the Non-Geographical Indication region, and reaching a 95.2% classification rate in other Chinese regions. In addition, it was proven that linear discriminant analysis (LDA) was suitable to use in other regions too, as it “gave an average correct classification rate of 100% for three green teas”. In another study, and because consumers are becoming more and more concerned about food quality throughout the food chain, Esteki et al. (2018) have proven that the ability of ultraviolet-visible spectroscopy and nuclear magnetic resonance spectroscopy, vibrational spectroscopy, fluorescence spectroscopies, and hyperspectral imaging combined with LDA was essential to quantify food contaminants and determine food products. In addition, the authors concluded that the combination of these multivariate techniques could be generalized to help food manufacturers and health and safety executives to verify the quality of the food. In the document written by Ali et al. (2019), a new intelligent system is proposed to help physicians diagnose Parkinson’s disease. LDA took a critical role in the method, being used as a tool for dimensionality reduction. The new algorithm proposed discriminated Parkinson’s disease patients and healthy patients with an accuracy of 95% in the learning set and 100% on the validation set using all characteristics available in the data. One of the limitations of the study was gender imbalance, reducing the discriminant capability of the method to 80% and 82.14%, respectively for the same databases as previously mentioned.

DA has proven to be a good decision-making method in the articles described above. Its broad applicability is remarkable and has led the authors to interesting outcomes in favor of their expected results.

2.2. PROBLEMATIC CONTEXT

In the field of mental health, DA may be used to identify specific patterns of symptoms or characteristics that are associated with different MDs. This can be useful for accurately diagnosing mental health problems, as well as for predicting which treatment options may be most effective for a particular individual.

Dugas et al. (1998) demonstrated that using DFs derived from symptom measures can effectively classify subjects with a generalized anxiety disorder (GAD) and non-clinical control subjects. In this study, 91% of subjects were correctly classified using symptom-based DFs, with 22 out of 24 in the GAD group and 18 out of 20 in the non-clinical group being correctly classified. In contrast, using DFs based on four process variables (cognitive avoidance, problem orientation, intolerance of uncertainty and beliefs about worry, and cognitive avoidance) resulted in a lower classification rate of 82%, with 18 out of 24 in the GAD group and 18 out of 20 in the non-clinical group being correctly classified.

In a more recent study, Chen et al. (2022) showed the effectiveness of fivefold cross-validation and of the LDA model, a method for evaluating the performance of a machine learning model in classifying individuals diagnosed with panic disorder, bipolar disorder, major depressive disorder, and schizophrenia using a multi-protein serum. The results obtained for both methodologies were 96.9% and 96.5%, respectively.

In the present study, DA will be used as a key method to differentiate nervous individuals, people who have had psychotic episodes, and healthy individuals. Considering the studies mentioned before, it is expected that the discriminant power of the separability between the three groups will be high, leading to a low misclassification probability. Gender will not be considered, as ADs may be more prevalent and disabling in women due to differences in socialization, upbringing, or exposure to stressors rather than any inherent difference in anxiety susceptibility (Shear, 1997).

Several limitations underlined to DA should be noted. Firstly, the DFs developed may not be generalizable to other populations or samples, as the model is based on the specific characteristics of the sample used. However, Ma et al. (2016) have demonstrated that DFs can be effectively applied across similar samples, and Esteki et al. (2018) suggest a generalization to help in the verification of food quality. Additionally, this study is subject to the limitations inherent to DA, such as the reliance on certain assumptions about the distribution of variables and the relationship between predictor variables and group membership, as well as the sensitivity to missing data and the limited ability to handle non-linear relationships. Furthermore, the more groups that are considered, the greater the risk of misdiagnosis for individual subjects. It is recommended to minimize the number of predictor variables to avoid creating an overly complex model, which would also increase the difficulty in interpreting the conclusions.

While DA is prone to be a useful tool in the diagnosis of ADs, it should not be used as the sole method of diagnosis. It is important to consider a range of factors, including the patient's medical history and the results of other exams, to make an accurate diagnosis. It is also important to note that DA is a statistical method and is subject to certain limitations and assumptions, as discussed earlier. Therefore, it is advisable to use DA in conjunction with other methods and to carefully evaluate the results in the context of the individual patient.

3. METHODOLOGY

Bold letters will be represented as vectors throughout this topic.

3.1. DATA

This paper will be supported by a study conducted by Professor P. Pichot at the Saint-Anne Hospital, presented in Romeder's (1973) book. The study involved a survey of thirty questions, evaluating thirty distinct characteristics enumerated in Table 3.1. Each question was rated on a continuous scale from zero to four based on the frequency and intensity of symptoms in the days before the examination. It was known beforehand that out of a hundred and fifty young adults, fifty were classified as nervous, fifty were classified as psychotic, and fifty were classified as not having any mental health condition – therefore classified as healthy. In the study mentioned forty individuals from each group were selected for the training sample and were assigned to groups one, two, and three, while the remaining thirty took part in the test sample.

Table 3.1 – List of variables to be studied

Number	Variable
1.	Fatigue
2.	Nightmares
3.	Muscle twitching
4.	Cramps
5.	Tremors
6.	Tension
7.	Muscle pain
8.	Knotted throat
9.	Satiety
10.	Heartburn
11.	Diarrhea
12.	Horripilation
13.	Palpitations
14.	Headaches
15.	Dizziness
16.	Tingling
17.	Pulse
18.	Fainting
19.	Eye floaters
20.	Oppression
21.	Indifference

Number	Variable
22.	Attention
23.	Memory
24.	Indecision
25.	Vague anxiety
26.	Fear of the future
27.	Apprehension of the worst
28.	Fear of loneliness
29.	Other fears
30.	Fear of crowds

Out of the hundred and twenty questionnaires, the answers from each participant in the training sample were displayed in two separate tables, showing the mean of each variable for each group, and the standard deviation of each variable and for each group as shown in Table 3.2 and Table 3.3, respectively.

Table 3.2 – Mean values of each variable for each group

Variable	Group 1 Nervous group	Group 2 Psychotic group	Group 3 Healthy group
Variable 1.	2.300	2.100	1.550
Variable 2.	1.325	0.525	0.300
Variable 3.	0.525	0.775	0.250
Variable 4.	0.325	0.425	0.350
Variable 5.	1.075	1.025	0.150
Variable 6.	2.675	1.725	0.675
Variable 7.	0.975	0.500	0.650
Variable 8.	1.125	0.925	0.200
Variable 9.	1.350	0.800	0.750
Variable 10.	0.400	0.275	0.050
Variable 11.	0.325	0.325	0.150
Variable 12.	1.000	0.550	0.150
Variable 13.	1.275	1.550	0.100
Variable 14.	1.200	0.825	0.525
Variable 15.	0.850	0.975	0.275
Variable 16.	0.800	0.625	0.400
Variable 17.	0.700	0.700	0.100

Variable	Group 1 Nervous group	Group 2 Psychotic group	Group 3 Healthy group
Variable 18.	0.850	0.775	0.025
Variable 19.	0.625	0.700	0.125
Variable 20.	1.550	1.325	0.200
Variable 21.	2.000	1.225	0.375
Variable 22.	2.575	2.375	0.800
Variable 23.	1.925	1.575	0.675
Variable 24.	2.175	1.875	0.700
Variable 25.	2.425	2.300	0.750
Variable 26.	2.450	2.000	0.475
Variable 27.	1.650	1.700	0.225
Variable 28.	1.200	1.425	0.125
Variable 29.	1.500	1.725	0.675
Variable 30.	1.350	0.975	0.075

Table 3.3 – Standard deviation values of each variable for each group

Variable	Group 1 Nervous group	Group 2 Psychotic group	Group 3 Healthy group
Variable 1.	1.646	1.513	1.431
Variable 2.	1.349	1.024	0.748
Variable 3.	1.072	1.235	0.733
Variable 4.	0.755	0.863	0.823
Variable 5.	1.349	1.351	0.654
Variable 6.	1.385	1.549	1.034
Variable 7.	1.405	1.025	1.130
Variable 8.	1.615	1.273	0.678
Variable 9.	1.542	1.288	1.318
Variable 10.	0.970	0.547	0.218
Variable 11.	0.848	0.565	0.421
Variable 12.	1.342	0.947	0.527
Variable 13.	1.323	1.431	0.300
Variable 14.	1.470	1.302	0.894
Variable 15.	1.295	1.294	0.922
Variable 16.	1.187	0.967	0.943

Variable	Group 1 Nervous group	Group 2 Psychotic group	Group 3 Healthy group
Variable 17.	1.030	1.005	0.300
Variable 18.	1.333	1.084	0.156
Variable 19.	1.177	0.980	0.640
Variable 20.	1.642	1.403	0.400
Variable 21.	1.703	1.214	0.827
Variable 22.	1.563	1.576	1.145
Variable 23.	1.523	1.579	0.985
Variable 24.	1.783	1.646	0.954
Variable 25.	1.611	1.676	1.199
Variable 26.	1.627	1.673	0.894
Variable 27.	1.696	1.600	0.474
Variable 28.	1.520	1.563	0.458
Variable 29.	1.565	1.396	1.081
Variable 30.	1.636	1.313	0.346

3.1.1. Data Collection

In this study, and due to the need of having every individual result from each participant to perform a discriminant analysis technique, a simulation was made of hundred and fifty participants, fifty for each group. As previously mentioned, the survey was evaluated between a minimum of zero and a maximum of four, classifying the frequency of the appearance of symptoms and their intensity. Knowing that the values of the questionnaire could only be between zero and four, the simulation would have to be through a truncated distribution, namely the Truncated Normal Distribution (TND).

3.1.1.1. Truncated Normal Distribution

The TND is a normal distribution limited to a specific interval. There can be three types of truncations: a left/lower truncation, where the minimum value of the distribution is pre-established; a right/upper truncation, where the maximum value of the distribution is pre-established; or a double truncation, where the upper and lower value of the distribution are two known constant values. In this study, the data was simulated generating a TND as the distribution associated with the provenience of the data.

Supposing that X has a TND with mean μ , standard deviation σ , inferiorly truncated by a and superiorly truncated by b , its density function is given by

$$f(x) = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, a \leq x \leq b \quad (3.1)$$

where Φ is the standard normal distribution cumulative distribution function, which is defined by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt. \quad (3.2)$$

3.2. STATISTICAL ANALYSIS

3.2.1. Introduction

Exploratory statistical analysis, namely univariate, and multivariate analysis, are a preliminary step to overview the differentiation between the three groups under study. This analysis will help to uncover patterns, anomalies, or relationships between all variables. The study aims to best discriminate each group considered, and using descriptive statistics such as boxplot analysis will provide an initial understanding of the best variables to achieve the objectives outlined in section 1.1.

To test the multivariate normality subjacent to the samples generated, Henze-Zirkler's test for normality was used, which is based on a non-negative functional distance. Under the null hypothesis, the test statistic is approximately log-normally distributed (see Henze-Zirkler Test in the Annexes).

3.2.2. Factorial Discriminant Analysis

Let X be the table with the p quantitative variables, and A a logic table associated with a qualitative variable with r modalities. Let p_i be the weight attributed to each individual i , and let

$$P_r = \sum_{i \in C_r} p_i \quad (3.3)$$

be the weight of class C_r . Let be

$$\mathbf{g} = \sum_{i=1}^n p_i \mathbf{x}_i \quad (3.4)$$

the center of gravity of the total cluster. As a result,

$$\mathbf{g}_r = \frac{1}{p_r} \sum_{i \in C_r} p_i \mathbf{x}_i \quad (3.5)$$

will be the center of gravity associated with the class C_r and $\mathbf{g}$ can also be calculated as

$$\mathbf{g} = \sum_{k=1}^r P_k \mathbf{g}_k. \quad (3.6)$$

The Variance and Covariance matrix is defined by

$$V = {}^t X D X = \sum_{i=1}^n p_i (\mathbf{x}_i - \mathbf{g}) * {}^t (\mathbf{x}_i - \mathbf{g}) \quad (3.7)$$

where D is the matrix of the weights attributed to the individuals. The variance between groups (B) measures the variability between the means of each group under study, quantifying how much each group differs from one another, while the variance within groups (W) measures the variability of individuals within each group. The variance between groups can be calculated by

$$B = \sum_{k=1}^r P_k (\mathbf{g}_k - \mathbf{g})^* (\mathbf{g}_k - \mathbf{g}) \quad (3.8)$$

while the variance within groups can be calculated by

$$W = \sum_{k=1}^r \sum_{i \in C_r} p_i (\mathbf{x}_i - \mathbf{g}_k)^* (\mathbf{x}_i - \mathbf{g}_k). \quad (3.9)$$

It can be shown that the sum of both between variance and within variance represents the total variability of the data, previously referred to as V .

To reach this paper's main purpose, one of the settled objectives is to obtain the linear combination of the thirty variables under study that best discriminates the classes. Being α the vector of the coefficients associated with each independent variable, then C can be represented as the linear combination

$$C = \sum_{j=1}^p \alpha^j x^j \quad (3.10)$$

which will be used to maximize the variance between groups. The respective variance can be calculated by

$$s^2(C) = \|C\|_D^2 = {}^t C D C \quad (3.11)$$

and by equivalence, is equal to ${}^t \alpha V \alpha$, demonstrated in Gomes (1993). As the total variance, V , can be decomposed into the within-group variance, W , and between-group variance, B , then

$${}^t \alpha V \alpha = {}^t \alpha W \alpha + {}^t \alpha B \alpha \quad (3.12)$$

and so

$$s^2(C) = {}^t \alpha W \alpha + {}^t \alpha B \alpha. \quad (3.13)$$

To obtain the linear combination C that best discriminates the classes, the optimal solution will be the vector α under the condition $Max \frac{{}^t \alpha B \alpha}{{}^t \alpha V \alpha}$, where ${}^t \alpha V \alpha$ is constant.

Without loss of generality, it is considered ${}^t \alpha V \alpha = 1$, simplifying the optimal solution to maximizing ${}^t \alpha B \alpha$. The maximum is obtained when the vector α is the eigenvector of $V^{-1}B$ associated with the highest eigenvalue λ_1 :

$$V^{-1}B \alpha = \lambda_1 \alpha \quad (3.14)$$

so $\frac{t_{\alpha B \alpha}}{t_{\alpha V \alpha}} = \lambda_1$. As a result, the highest eigenvalue λ_1 will measure the discriminant power associated with the first DF

$$C_1 = \sum_{j=1}^p \alpha^j x^j \quad (3.15)$$

where α^j are the coordinates of the eigenvector α . To test the significance of each DF, the likelihood ratio test will be performed to determine whether or not the DF under analysis is relevant to discriminate the individuals. The practical terms are analyzed in Likelihood Ratio Test in the Annexes.

Like maximizing $\frac{t_{\alpha B \alpha}}{t_{\alpha V \alpha}}$, the maximization of $\frac{t_{\alpha B \alpha}}{t_{\alpha W \alpha}}$ will produce equivalent results. Its solution will be the eigenvector of $W^{-1}B$ associated with the highest eigenvalue

$$W^{-1}B \alpha = \gamma_1 \alpha.^1 \quad (3.16)$$

The eigenvector will remain the same, but the eigenvalue will be

$$\gamma = \frac{\lambda}{1 - \lambda}, \lambda < 1 \quad (3.17)$$

The new metric will be W^{-1} as in opposition to the previous metric V^{-1} , the metric of Mahalanobis.

There can be at most $r - 1$ DFs, being r the number of classes under study. In the most complex problems, to separate r groups from each other, $r - 1$ boundaries will be needed.

Supposing that a new observation, x_0 , is obtained, the objective will be to allocate the observation to the nearest group recurring to the Mahalanobis distance (with metric W^{-1} or the equivalent metric V^{-1}) calculated by

$$\begin{aligned} d^2(x_0, g_i)_{W^{-1}} &= {}^t(x_0 - g_i)W^{-1}(x_0 - g_i) \Leftrightarrow \\ d^2(x_0, g_i)_{W^{-1}} &= {}^t x_0 W^{-1} x_0 + {}^t g_i W^{-1} g_i - 2 * {}^t x_0 W^{-1} g_i \end{aligned} \quad (3.18)$$

The new individual will then be attributed to the group to which $d^2(x_0, g_i)$ is minimum.

If the dispersion of the values subjacent to each group differs significantly from one another, the exploratory analysis reaches its limits, as a new observation could be keen to be attributed to the group whose dispersion is the highest. To overcome this limitation, a statistical distribution hypothesis of the repartition of the observations in the space is usually needed, which will imply a probabilistic model underlying the multivariate sample.

¹ In Principal Component Analysis, the principal factors are the eigenvectors of matrix QV , where Q represents the metric on individual space. From the coordinates of the first eigenvector of QV , the principal components can be written as a linear combination of the centered variables under study. If V is replaced by B and the metric Q is replaced by W^{-1} , then FDA can be considered a PCA of center of gravity of classes with the metric $Q = W^{-1}$.

3.2.3. Probabilistic Context

Let p_j be the proportion of observation in each group and $f_j(x)$ be the probabilistic distribution of x of group j . A new observation x_0 will be attributed to group C_j where

$$P(C_j|x) = \frac{p_j f_j(x)}{\sum_{j=1}^r p_j f_j(x)} \quad (3.19)$$

is the highest. For any new observation, the denominator will always be equal to 1, so the affection can be deduced as maximizing

$$p_j f_j(x). \quad (3.20)$$

In case the data can be assumed to be from a multivariate Gaussian model, its probability density function will be given by

$$f_j(x) = \frac{1}{(2\pi)^{\frac{p_j}{2}} \sqrt{\det(\Sigma_j)}} \exp \left[-\frac{1}{2} {}^t(x - \mu_j) \Sigma_j^{-1} (x - \mu_j) \right]. \quad (3.21)$$

Maximizing $f_j(x)$ is equivalent to maximizing its logarithm, as the logarithmic function is a strictly increasing function. Therefore, applying the logarithm to the maximization of equation (3.20) will be equivalent to

$$\text{Min} \left[{}^t(x - \mu_j) \Sigma_j^{-1} (x - \mu_j) - 2 \ln p_j + \ln(\det(\Sigma_j)) \right]. \quad (3.22)$$

If there is equality between the variance and covariance matrices, the decision rule is linear. From equation (3.22), $\ln(\det(\Sigma_j))$ becomes constant and ${}^t(x - \mu_j) \Sigma^{-1} (x - \mu_j)$ is the Mahalanobis distance between x and μ_j , which can be decomposed in

$${}^t x \Sigma^{-1} x - 2 {}^t x \Sigma^{-1} \mu_j + {}^t \mu_j \Sigma^{-1} \mu_j. \quad (3.23)$$

The term ${}^t x \Sigma^{-1} x$ of the equation above does not depend on group j , therefore the maximization of $\ln(p_j f_j(x))$ can be written as

$$\text{Max} \left[{}^t x \Sigma^{-1} \mu_j - \frac{1}{2} {}^t \mu_j \Sigma^{-1} \mu_j + \ln p_j \right] \quad (3.24)$$

Let x be a new observation. To attribute x to group i or j , the linear discriminant scores are defined as

$$W_{ij} = {}^t x S^{-1} (\bar{x}_i - \bar{x}_j) - \frac{1}{2} {}^t (\bar{x}_i - \bar{x}_j) S^{-1} (\bar{x}_i - \bar{x}_j) \quad (3.25)$$

where $\bar{x}_i$ and $\bar{x}_j$ are unbiased estimators of the mean of groups i and j , respectively and S is the unbiased estimator of the variance-covariance matrix Σ defined by each group, calculated by

$$S = \frac{1}{n-r} \sum_{i=1}^r (n_i - 1) S_i \quad (3.26)$$

where S_i is the empirical variance matrix of group i ($i = 1, \dots, r$).² The classification rule will be assigned individual x group i if $W_{ij} > 0$, for $i \neq j$.

In the particular case where $k = 3$, the discriminant scores will be:

$$\begin{aligned} W_{12} &= {}^t\chi S^{-1}(\bar{x}_1 - \bar{x}_2) - \frac{1}{2} {}^t(\bar{x}_1 - \bar{x}_2) S^{-1}(\bar{x}_1 - \bar{x}_2) \\ W_{13} &= {}^t\chi S^{-1}(\bar{x}_1 - \bar{x}_3) - \frac{1}{2} {}^t(\bar{x}_1 - \bar{x}_3) S^{-1}(\bar{x}_1 - \bar{x}_3) \\ W_{23} &= {}^t\chi S^{-1}(\bar{x}_2 - \bar{x}_3) - \frac{1}{2} {}^t(\bar{x}_2 - \bar{x}_3) S^{-1}(\bar{x}_2 - \bar{x}_3) \end{aligned} \quad (3.27)$$

Considering that $W_{23} = W_{13} - W_{12}$, it is only required two of the equation's terms to know the final one. As a result, the classification rule is defined as

- Classify the new individual to group 1 if $W_{12} > 0$ and $W_{13} > 0$
- Classify the new individual to group 2 if $W_{12} < 0$ and $W_{23} > 0$
- Classify the new individual to group 3 if $W_{13} < 0$ and $W_{23} < 0$

If Σ_j is different among the groups under study, it will be necessary to compare k quadratic functions of x , being Σ_j estimated by $\frac{n_j}{n_j-1}S_j$, where S_j is the empirical variance and covariance matrix of group j , and μ_j estimated by the center of gravity of each group $\mathbf{g}_j$.

Box M test will be used to test the equality of the variance-covariance matrices between groups, being its hypothesis

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_r$$

$$vs$$

$$H_1: \exists(r, j), \text{ with } r \neq j \text{ and } j < r: \Sigma_r \neq \Sigma_j$$

where r is the total number of groups under study (see Box M Test in the Annexes). The sensitivity of the Box M test to the lack of normality in the variable vector emphasizes the potential use of a significance level of 0.01 or lower (Olson, 1974).

If the equality between the variance-covariance matrices cannot be assumed, quadratic discriminant analysis overcomes this problem, although it will be needed to estimate each Σ_r , which will complexify the problem. Another problem associated with the use of quadratic discriminant analysis is when the sample sizes are small. It affects the robustness of the DFs obtained, so it may be better to use LDA either way (Saporta, 2011).

3.3.CONTRIBUTION OF VARIABLES

For the contribution of each variable to each DF, the coefficients are analyzed in absolute value, so that it is known if the variable of matter is or is not important to the problematic context. As for

² If Σ is estimated by S (so $\frac{n}{n-p}W$ where W is the within-group variance matrix), the Bayesian rule will correspond to the geometric rule under the equality of the probabilities *a priori*, so the geometric rule is then optimal.

the interpretation of the DFs, the signs of the coefficients are important to determine the context of the variables under analysis (Rencher, 1995). Three different approaches are suggested to analyze the contribution of each variable to discriminate the groups, namely the standardized discriminant function coefficients, the correlation between variables, and the DFs and perform partial F-tests.

The approach of analyzing the contribution of each variable to discriminate between groups may not provide the most accurate results since it does not consider the potential impact of other variables, which could lead to misleading conclusions (Rencher, 1995).

The standardized discriminant function coefficients enable the possibility of comparing each variable with one another, as the coefficients become-scale free. As a result, the coefficients will showcase the exact contribution of each variable to the corresponding DF (see Standardized Coefficients in the Annexes).

Finally, the partial F-test is a statistical test whose objective is to show the relevance of each variable to the contribution of group separation (see Partial F-Values in the Annexes). However, when there is more than one DF, the partial F-values cannot be associated with a DF, but with the overall contribution of the variable to the group discrimination. In opposition, if the weight of any eigenvalue is large enough, then most of the separability is accounted for the DF associated with such eigenvalue, and the variables, ordered according to their partial F-values, may produce similar results to those obtained by the standardized discriminant function coefficients (Rencher, 1995).

3.4. STEPWISE DISCRIMINANT ANALYSIS

Multicollinearity can influence the selection of discriminator variables. The absence of a specific variable does not necessarily mean the variables lack importance for the model. The omitted variable may be indeed capable of discriminating the groups, but its correlation with other variables does not allow it to be included in the model (Sharma, 1995). Recurring to Variance Inflation Factors (*VIF*) it is possible to determine whether or not a variable is related to another (see Variance Inflation Factors in the Annexes).

Stepwise Discriminant Analysis combines both forward and backward approaches. In the forward approach, the model selects the variable with the highest partial F-statistic based on Wilk's lambda at each step. As the analysis proceeds, the previously included variables are reevaluated to determine if any variable that has entered the model has become redundant due to newly included variables. This joint procedure continues until the largest partial F value among the included variables surpasses a predefined threshold. In the backward approach, the initial model includes all variables and, at each step, the variable with the lowest partial F values, in other words, the one that least contributes to the model, is removed from the model (Rencher, 1995). Once the subset of discriminant variables is established, it will be useful to calculate the DFs and evaluate the performance of this discriminant procedure, namely, evaluating the percentage of misclassified observations using the sample test. This procedure could be referred to as Stepwise MANOVA (Multivariate Analysis of Variance). In this approach, no DFs are calculated at each step. After the selection of variables is achieved, the DFs and the misclassified rate matrix will be calculated. The practical terms are examined in the section Stepwise Discriminant Analysis in the Annexes.

To evaluate the effectiveness and reliability of each model, they will be evaluated through a testing set. Using testing sets is a reliable way of analyzing the overall performance of the model, as it is a process that helps ensure that the model performs reliably and accurately when applied to unseen data.

4. RESULTS AND DISCUSSION

The first steps of data processing, which were simulating TND values to create the samples, were done with the software RStudio, which uses R as the programming language. With the same software, it was possible to create boxplots, perform the Henze-Zirkler's test, and Box's M-test. JMP Pro 17 and SPSS were used to perform the discriminant analysis techniques.

4.1. EXPLORATORY STATISTICAL ANALYSIS

Given that the values of each variable were simulated from a TND with a lower limit of 0 and an upper limit of 4, the range of each variable will be, at most, the same.

For the preliminary data analysis, the variables will be compared within the same group, as the objective is to discriminate the three groups. As a result, the most important variables to analyze will be the ones whose distribution can be distinguished when comparing the same variable for the three separate groups. As a follow-up, the use of parallel boxplots was found to be a good method for the statistical distribution comparison of the pseudo-random variables under study. The main purpose of analyzing parallel boxplots is to find the variables whose boxplots of each group are the furthest from one another. From Figure A.1 to Figure A.30, it can be noted that it will be easier to separate healthy individuals from non-healthy individuals (Nervous and Psychotic groups), as the healthy group, represented by dark grey, is frequently differentiated from the other two groups in several variables.

The interpretation of all boxplots led to the following interpretation: *Fatigue's*, *Cramps'*, and *Satiety's* boxplots indicate that the difference between each group is either non-existing or almost null, which suggests these variables will not seem to be good variables for this study. The variables whose groups all seemed to differ from each other were *Nightmares*, *Muscle twitching*, *Tension*, *Heartburn*, *Horripilation*, *Headaches*, *Indifference*, *Memory*, and *Vague anxiety*. In particular, it is noted that:

- Variables *Nightmares* and *Memory* have bigger differences between the healthy and non-healthy groups, while the difference within the non-healthy groups is not as relevant.
- Variable *Headaches* has bigger differences between the healthy and nervous groups, while the difference within the non-healthy groups is not as relevant.

Variables *Tremors*, *Knotted throat*, *Palpitations*, *Headaches*, *Dizziness*, *Pulse*, *Fainting*, *Eye floaters*, *Oppression*, *Indifference*, *Attention*, *Memory*, *Indecision*, *Vague Anxiety*, *Fear of the future*, *Apprehension of the worst*, *Fear of loneliness*, *Other fears*, and *Fear of crowds* were the ones who demonstrated the most relevant difference between the healthy group and the non-healthy groups. From this set, it seems that variables *Headaches*, *Indifference*, *Memory*, and *Vague Anxiety* have a small difference within the non-healthy groups, but not enough to be considered different amongst all groups. The variables whose distributions' visualization revealed a difference between the group of nervous individuals and the other two groups were variables *Diarrhea* and *Tingling*. The only variable that showed any differences between the Psychotic group and the other two groups, while there were no apparent differences between the other two groups, was *Muscle Pain*, and it is not as evident as the difference other variables may present.

To summarize, the boxplot's interpretation can be resumed in five different sets:

1. All three groups' distributions do not seem to be different - variables *Fatigue, Cramps, and Satiety*;
2. All three groups' distributions seem to differ from one another – variables *Nightmares, Muscle twitching, Tension, Heartburn, Horripilation, Headaches, Indifference, Memory, and Vague anxiety*;
3. The *Healthy* group differentiates itself from the other two groups, but there is no apparent differentiation amongst the other groups – variables *Tremors, Knotted throat, Palpitations, Headaches, Dizziness, Pulse, Fainting, Eye floaters, Oppression, Indifference, Attention, Memory, Indecision, Vague Anxiety, Fear of the future, Apprehension of the worst, Fear of loneliness, Other fears, and Fear of crowds*;
4. The *Nervous* group differentiates itself from the other two groups, but there is no apparent differentiation amongst the other groups – variables *Diarrhea and Tingling*;
5. The *Psychotic* group differentiates itself from the other two groups, but there is no apparent differentiation amongst the other groups – variable *Muscle Pain*.

To enrich the exploratory statistical analysis, FDA will be performed with all thirty variables predicting the group to which an individual will be allocated, although it is noted that some variables might not be as useful to discriminate each group from one another.

To test the multivariate normality adjacent to the individuals of each group, the Henze-Zirkler's multivariate normality test was performed, in which the p-values obtained were, rounded to the thousandth, 0.079, 0.096, and 0.134, corresponding to the nervous, psychotic, and healthy groups respectively. Therefore, the multivariate normality is not rejected in any group, and so it is possible to perform the Box's M test. The test statistic obtained was 1026.4, with a p-value of 0.015, so the null hypothesis is not rejected for the significance level fixed at most 0.01. Thus, the three groups can be assumed to have homogeneous covariance matrices. With these two assumptions, FDA can be performed on the whole data set.

4.2. FACTORIAL DISCRIMINANT ANALYSIS

Firstly, to understand how all variables would behave in each group, FDA was made using the thirty variables under analysis in an exploratory statistical analysis context.

From Figure 4.1, which showcases the dispersion of a hundred and fifty individuals, with each group centroid defined by a "+" sign, when represented by all variables, it is first noted that the separation of the healthy group and the non-healthy groups is evident, as expected from the boxplot analysis. The second finding made when analyzing the figure is the separation between the psychotic and nervous groups. It can then be inferred that the first DF will aim to separate the healthy group from the non-healthy groups, whereas the second DF will aim to separate nervous individuals from psychotic individuals. For the dispersion of the data, individuals from the Healthy group are closer to its group centroid, meaning the dispersion is relatively small. On the other hand, the behavior of the individuals in the other two groups is not as uniform, which may lead to bigger misclassification in these two groups.

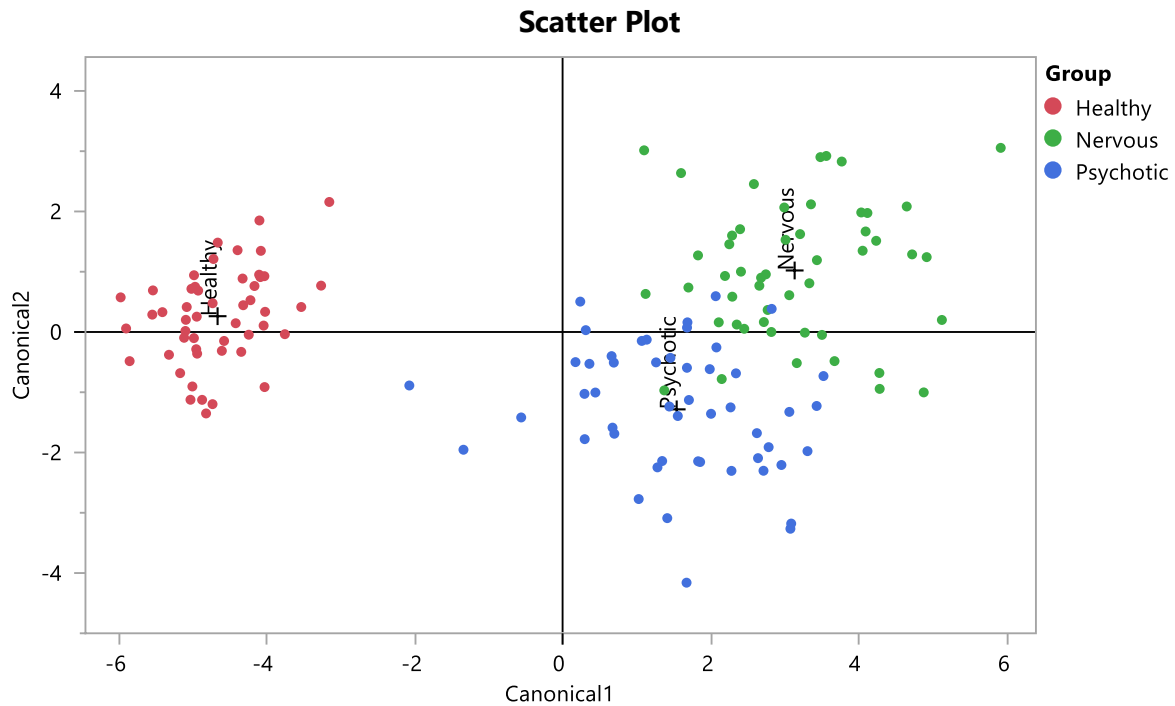


Figure 4.1 – Representation of all thirty variables in the first factorial principal plane generated by the two discriminant functions

The standardized scoring coefficients of each variable for each DF, represented in Table 4.1, indicate the contribution of the variable to the DF. Therefore, the further from zero the coefficient is, the more the variable will contribute to the discrimination of the three groups. Figure 4.2 helps the interpretation of the standardized scoring coefficients, as the length and orientation of each ray are a positive multiple of the standardized scoring coefficient of each variable. It is first noticed that, for the first DF, all the coefficients except for *Headaches*, *Muscle pain*, and *Satiety* are positive. In Figure 4.1, it can be seen that the Healthy group is represented in the second and third quadrants, whereas the other two groups are represented in the first and fourth quadrants, meaning it is expected low/negative values in the first DF to be associated with healthy individuals, while high/positive values to be associated with unhealthy groups. From the set of thirty variables, the ones who have the biggest scoring coefficient in the first DF are variables (and respective standardized scoring coefficients in parenthesis) *Palpitations* (0.401), *Heartburn* (0.4), *Apprehension of the worst* (0.397), *Oppression* (0.381), and *Fear of loneliness* (0.35). For the second DF, it can be noticed that the variables that most contribute to the discrimination of groups Nervous and Psychotic are *Palpitations* (-0.507), *Muscle pain* (0.499), *Muscle twitching* (-0.480), *Headaches* (0.429), *Tingling* (0.415), and *Tension* (0.388). *Palpitations* and *Muscle twitching* contribute to the psychotic group, while the other three variables mentioned contribute to the nervous group. All eleven variables mentioned seem to be ones whose standardized coefficients are the biggest, which means their association with the respective DFs is the biggest. All other variables unmentioned need to be further analyzed, as their standardized scoring coefficients do not seem to be relevant for group discrimination. As suggested by the exploratory data analysis, the approach of DA with a reduced number of variables will be taken into consideration later on.

Table 4.1 – Standardized scoring coefficients of the thirty variables in each discriminant function

Variable	Discriminant Function 1	Discriminant Function 2
Fatigue	0.182	0.216
Nightmares	0.289	0.112
Muscle twitching	0.212	-0.480
Cramps	0.069	0.040
Tremors	0.069	0.029
Tension	0.230	0.388
Muscle pain	-0.058	0.499
Knotted throat	0.144	0.080
Satiety	-0.059	0.235
Heartburn	0.400	0.202
Diarrhea	0.117	0.164
Horripilation	0.255	0.136
Palpitations	0.401	-0.507
Headaches	-0.030	0.429
Dizziness	0.047	0.102
Tingling	0.127	0.415
Pulse	0.230	-0.073
Fainting	0.205	-0.224
Eye floaters	0.049	0.141
Oppression	0.381	-0.090
Indifference	0.235	0.045
Attention	0.113	0.082
Memory	0.229	-0.302
Indecision	0.256	0.146
Vague anxiety	0.116	0.025
Fear of the future	0.214	-0.066
Apprehension of the worst	0.397	-0.069
Fear of loneliness	0.350	-0.080
Other fears	0.065	-0.247
Fear of crowds	0.242	0.171

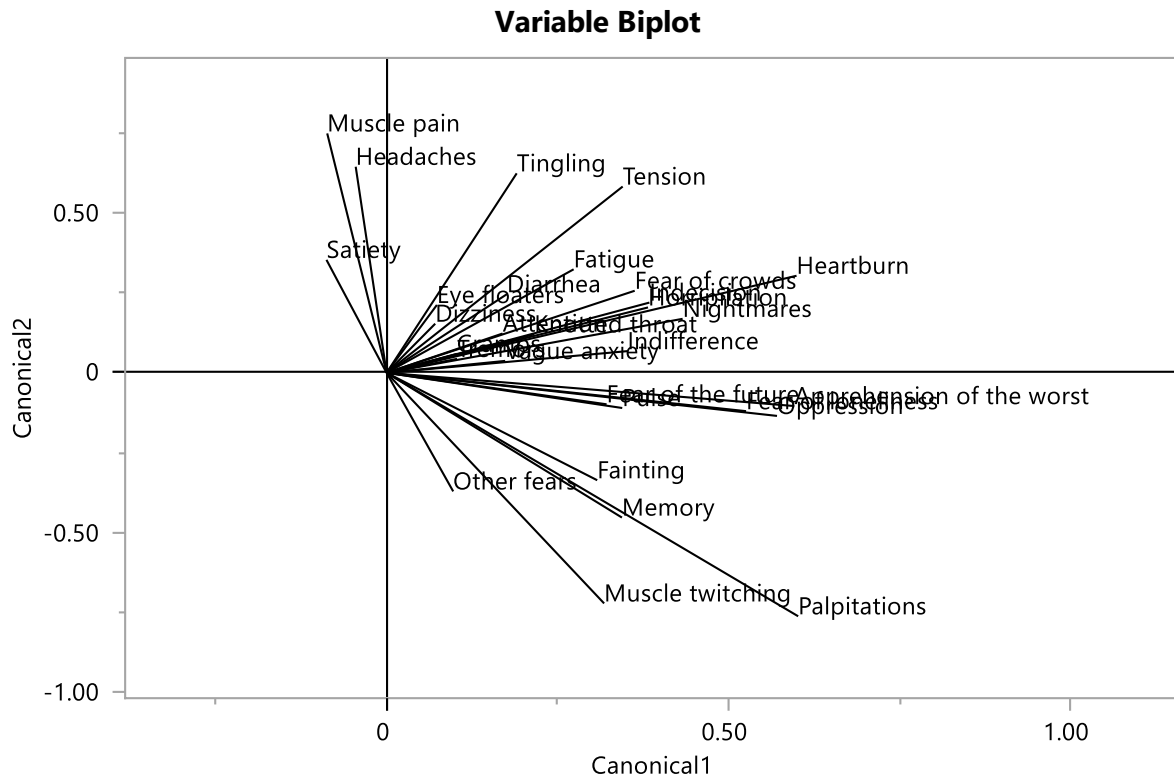


Figure 4.2 – Biplot representing a multiple of the coefficients for all standardized covariates

The two eigenvalues obtained were 11.537 and 0.940. Consequently, the discriminant power of the first DF is 92.47% and the discriminant power of the second DF is 7.53%, which comes to prove that it will be easier to separate healthy and non-healthy individuals than it will be to separate nervous individuals from psychotic individuals. The two obtained DFs associated with the model are:

$$\begin{aligned}
 C_1 = & 0.185 * x_1 + 0.377 * x_2 + 0.292 * x_3 + 0.137 * x_4 + 0.088 * x_5 + 0.265 * x_6 \\
 & - 0.077 * x_7 + 0.187 * x_8 - 0.066 * x_9 + 0.999 * x_{10} + 0.278 * x_{11} \\
 & + 0.355 * x_{12} + 0.482 * x_{13} - 0.036 * x_{14} + 0.061 * x_{15} + 0.176 \\
 & * x_{16} + 0.347 * x_{17} + 0.332 * x_{18} + 0.069 * x_{19} + 0.467 * x_{20} \\
 & + 0.277 * x_{21} + 0.121 * x_{22} + 0.241 * x_{23} + 0.251 * x_{24} + 0.114 \\
 & * x_{25} + 0.243 * x_{26} + 0.427 * x_{27} + 0.433 * x_{28} + 0.065 * x_{29} \\
 & + 0.299 * x_{30} - 8.791
 \end{aligned} \tag{4.1}$$

$$\begin{aligned}
 C_1 = & 0.218 * x_1 + 0.147 * x_2 - 0.662 * x_3 + 0.079 * x_4 + 0.037 * x_5 + 0.447 * x_6 \\
 & + 0.668 * x_7 + 0.104 * x_8 + 0.264 * x_9 + 0.505 * x_{10} + 0.388 * x_{11} \\
 & + 0.19 * x_{12} - 0.609 * x_{13} + 0.51 * x_{14} + 0.13 * x_{15} + 0.578 * x_{16} \\
 & - 0.111 * x_{17} - 0.362 * x_{18} + 0.199 * x_{19} - 0.11 * x_{20} + 0.053 * x_{21} \\
 & + 0.088 * x_{22} - 0.317 * x_{23} + 0.143 * x_{24} + 0.024 * x_{25} - 0.074 \\
 & * x_{26} - 0.074 * x_{27} - 0.1 * x_{28} - 0.248 * x_{29} + 0.212 * x_{30} - 2.631
 \end{aligned} \tag{4.2}$$

where x_1 = Fatigue, x_2 = Nightmares, x_3 = Muscle twitching, x_4 = Cramps, x_5 = Tremors, x_6 = Tension, x_7 = Muscle pain, x_8 = Knotted throat, x_9 = Satiety, x_{10} = Heartburn, x_{11} = Diarrhea, x_{12} = Horripilation, x_{13} = Palpitations, x_{14} = Headaches, x_{15} = Dizziness, x_{16} = Tingling, x_{17} = Pulse, x_{18} = Fainting, x_{19} = Eye floaters, x_{20} = Oppression, x_{21} = Indifference, x_{22} = Attention, x_{23} = Memory, x_{24} = Indecision, x_{25} = Vague anxiety, x_{26} = Fear of the future, x_{27} = Apprehension of the worst, x_{28} = Fear of loneliness, x_{29} = Other fears, and, x_{30} = Fear of crowds. When testing the

likelihood ratio test, the p-value obtained for each DF, C_1 and C_2 , was below 0.01, justifying the use of both DF to discriminate the three groups under study.

Representing the classifications attained from the training set in the first FDA performed, Table 4.2 shows that the healthy individuals are not for once mistaken with individuals from any other group, leading to a perfect classification of the Healthy group, and only one individual (from the Psychotic group) is misclassified for healthy. In addition, the percentage of individuals that were misclassified is 4.(6)% (3 individuals from the nervous group and four individuals from the psychotic group, out of the total 150 individuals).

Table 4.2 – Classification count in the training set with thirty variables

Classification Results						
		Predicted Classification Group				Total
Group		Nervous	Psychotic	Healthy		
Original	Count	Nervous	47	3	0	50
		Psychotic	3	46	1	50
		Healthy	0	0	50	50

To validate the models' accuracy in discriminating the three groups, the model was submitted to a testing set. The testing set contained seventy-five new individuals, evenly distributed by the three groups, generated from the same TND as the training sample. A misclassification rate of 10.(6)% was obtained and, in Table 4.3, it can be seen that the model was able to perfectly classify healthy individuals, while the psychotic individuals were only once mistakenly classified as healthy individuals. As a result, it can be said that the model was able to discriminate the three groups predominantly correctly.

Table 4.3 – Classification count in the testing set with thirty variables

Classification Results						
		Predicted Classification Group				Total
Group		Nervous	Psychotic	Healthy		
Original	Count	Nervous	18	7	0	25
		Psychotic	0	24	1	25
		Healthy	0	0	25	25

To further analyze the three groups on behalf of this paper's main objective, three different approaches are referred to analyze the behavior of each variable to the DFs, from which standardized scoring coefficients have already been analyzed in combination with the variable biplot.

From Table 4.4, with the test statistic F and p-values associated with the test statistic obtained for each variable, it can be seen that variables *Fatigue*, *Cramps*, and *Satiety* are not significantly

contributing to the discriminant model created, as their p-value is superior to 0.28, which goes according to the exploratory analysis made.

Table 4.4 – Partial F test' statistic and p-value performed on all thirty variables

Variable	Test Statistic - F	P-value
Fatigue	1.256	0.288
Nightmares	19.448	0.000
Muscle twitching	14.245	0.000
Cramps	0.017	0.983
Tremors	28.381	0.000
Tension	30.398	0.000
Muscle pain	7.036	0.001
Knotted throat	31.163	0.000
Satiety	1.004	0.369
Heartburn	37.491	0.000
Diarrhea	12.382	0.000
Horripilation	33.683	0.000
Palpitations	57.591	0.000
Headaches	13.652	0.000
Dizziness	8.485	0.000
Tingling	11.315	0.000
Pulse	31.812	0.000
Fainting	46.508	0.000
Eye floaters	13.969	0.000
Oppression	47.004	0.000
Indifference	27.133	0.000
Attention	15.362	0.000
Memory	17.020	0.000
Indecision	9.216	0.000
Vague anxiety	15.771	0.000
Fear of the future	28.811	0.000
Apprehension of the worst	45.353	0.000
Fear of loneliness	54.338	0.000
Other fears	7.905	0.001
Fear of crowds	58.842	0.000

From Table 4.5, it is observed that the correlations between the variables and the two DFs are bigger in the first DF than it is in the second one. One of the main causes is the percentage of the variance of the first DF being 92.47%, while the second DF covers the remaining 7.53%, as previously stated. For this reason, the weight of the correlation between variables and the first DF will be bigger than the correlation between variables and the second DF. The correlation between the variables to the first DF goes as high as 0.693 (*Fear of crowds*) to as low as -0.096 (*Muscle pain*). For the second DF, the correlation goes as high as 0.391 (*Muscle twitching*) to as low as -0.437 (*Tingling*). When analyzing the correlation with the first DF, only two variables are negatively (and weakly) correlated. It confirms that the first DF's purpose is to separate healthy individuals from non-healthy individuals. It is expected that healthy individuals score low on almost all variables, as they are so well separated from the other two groups as shown in Figure 4.1. In opposition to the first DF, the correlation between variables and the second DF are essentially bidirectional. Although there is more variability in the coefficients of the correlations, they are not as strong as in the first DF as previously mentioned. Having rays pointing at different quadrants when analyzing the second DF translates to having variables that describe better the nervous group and variables that describe better the psychotic group. It is noted that variables *Tingling*, *Muscle pain*, and *Tension* with correlations 0.437, 0.403, and 0.37, in the order mentioned, best describe the psychotic group, while the variables that best describe the nervous group are *Muscle twitching*, *Palpitations*, and *Other fears* with correlations -0.391, -0.277, and -0.234 in the respective order.

Table 4.5 – Correlation of each of the thirty variables to each discriminant function

Variable	Discriminant Function 1	Discriminant Function 2
Fatigue	0.131	0.044
Nightmares	0.467	0.136
Muscle twitching	0.310	-0.391
Cramps	-0.014	-0.011
Tremors	0.550	0.038
Tension	0.496	0.370
Muscle pain	-0.096	0.403
Knotted throat	0.568	0.042
Satiety	0.059	0.146
Heartburn	0.562	0.311
Diarrhea	0.379	0.159
Horripilation	0.561	0.225
Palpitations	0.661	-0.277
Headaches	0.369	0.255
Dizziness	0.318	-0.146
Tingling	0.211	0.437
Pulse	0.569	-0.097
Fainting	0.643	-0.117

Variable	Discriminant Function 1	Discriminant Function 2
Eye floaters	0.416	-0.038
Oppression	0.651	0.006
Indifference	0.530	0.152
Attention	0.430	-0.070
Memory	0.441	0.134
Indecision	0.345	-0.061
Vague anxiety	0.418	0.183
Fear of the future	0.549	-0.096
Apprehension of the worst	0.635	-0.146
Fear of loneliness	0.659	-0.230
Other fears	0.277	-0.234
Fear of crowds	0.693	0.079

As previously stated and accordingly to one of the pre-established papers' objectives, the use of all variables to reach the main purpose of the study may prove not to be ideal, as many variables can show the same results, or even not be good to discriminate one group from another. The separation of healthy and unhealthy individuals is notably easier than the separation within unhealthy groups. For this reason and based on the analysis conducted on the contribution of each variable to FDA, the variables that indicate a good separation amongst the three groups are *Muscle twitching*, *Muscle pain*, *Tension*, *Heartburn*, *Horripilation*, *Palpitations*, *Headaches*, *Tingling*, *Pulse*, *Fainting*, *Oppression*, *Indifference*, *Apprehension of the worst*, *Fear of loneliness*, and *Fear of crowds*. Firstly, it is noted that all variables are statistically significant at a significance level fixed at most 0.01, except for *Muscle pain*. Despite this, *Muscle pain* has a standardized coefficient of 0.499 with the second DF and a correlation with the same DF of 0.403. For this reason, the variable was chosen to be included in the model, as it is an important variable to discriminate both unhealthy groups from each other. Variables *Headaches* and *Tingling* were also chosen to take part in the model as their correlation to the second DF and standardized scoring coefficients were high when compared to other variables. Although the partial-F values were not the highest, the variables had to take part in the model to give more weight to the discrimination of Nervous and Psychotic groups, otherwise, the model's focus would be to only discriminate healthy and unhealthy groups. Still, variables *Fear of crowds*, *Fainting*, *Oppression*, *Apprehension of the worst*, *Indifference*, *Pulse*, and *Fear of loneliness* were chosen as their behavior was very focused on separating the Healthy group from the two others. Variables *Muscle twitching*, *Tension*, *Heartburn*, *Horripilation*, *Palpitations*, and *Fainting* are variables that have a remarkable performance to discriminate the groups under study, the reason being choosing them to take part in the new model.

4.3. REDUCED MODEL

Since the set of predictive variables is different, it is required to remake the multivariate normality test and Box's M test to elaborate the final FDA. It is not required to remake the exploratory data analysis, as each variable's analysis was made independent of other variables.

For Henze-Zirkler's multivariate normality test, the test statistics attained were 0.997, 0.995, and 0.995, with p-values of 0.14, 0.26, and 0.23, respectively for the Nervous, Psychotic, and Healthy groups. The hypothesis of multivariate normality adjacent to the individuals of each group is not rejected. For the Box's M test, the test statistics obtained was 280.88, with a p-value of 0.04, from which the null hypothesis is not rejected. So, the homogeneity of covariance matrices amongst the three groups can be assumed.

From the FDA elaborated with the fifteen predictive variables, Figure 4.3 shows the dispersion of a hundred and fifty individuals, with each group centroid represented by a "+" sign when illustrated by the respective fifteen variables. It is clear that the dispersion of the three groups has increased, although it is still noticeable that the dispersion of the healthy group is smaller than the other two groups. It is also visible that there is a greater mix of individuals in the nervous and psychotic groups, which indicates a bigger misclassification rate in this group.

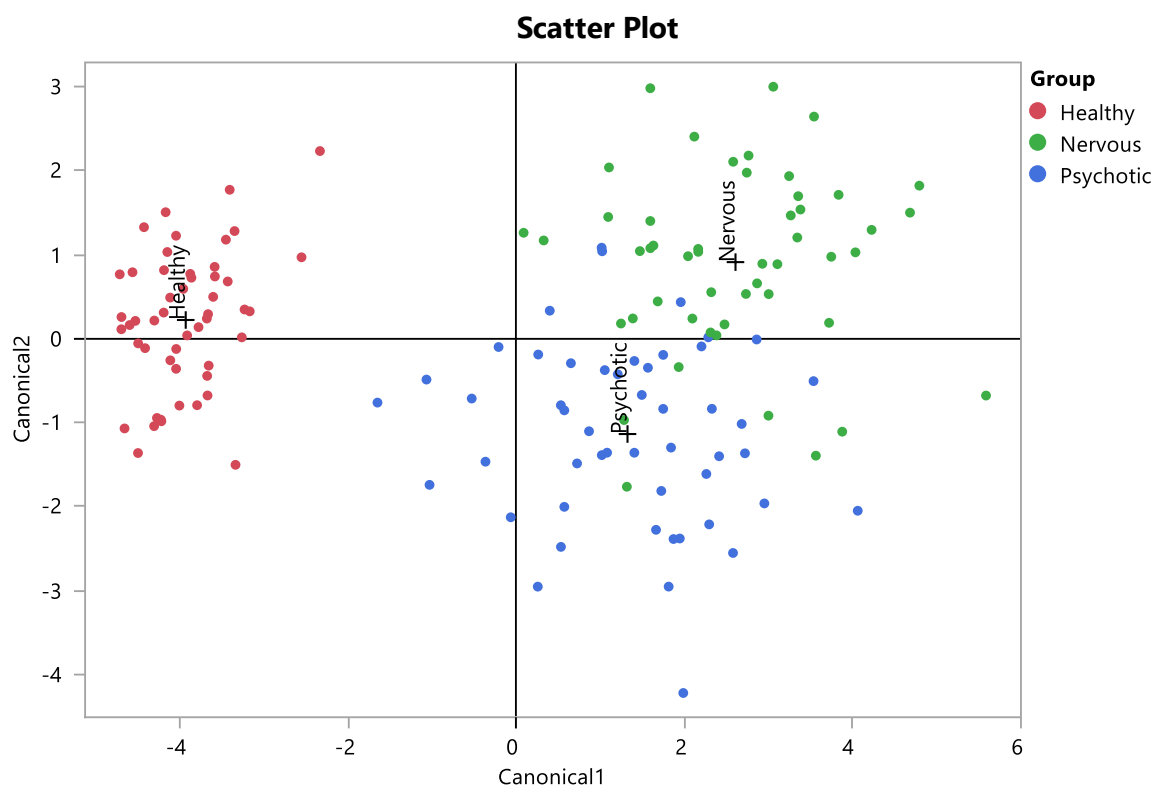


Figure 4.3 – Representation of the proposed variable selection (15 variables) in the first factorial principal plane generated by the two discriminant functions

The standardized scoring coefficients are represented in Table 4.6, and as expected, all variables have a high standardized coefficient. In either one of the two functions, each variable shows how it behaves in accordance with the function. For instance, despite variable *Headaches* having a near-

zero standardized scoring coefficient with the first DF, it has one of the highest standardized scoring coefficients in the second DF. Comparing Figure 4.2 to Figure 4.4, where it is illustrated a positive multiple of the scoring coefficients of the corresponding covariates, it is seen that the rays' orientations do not differ from each other, as the variables chosen were the ones that had the biggest weight when discriminating the three groups.

Table 4.6 – Standardized scoring coefficients of the fifteen variables in each discriminant function

Variable	Discriminant Function 1	Discriminant Function 2
Muscle twitching	0.141	-0.479
Muscle pain	-0.118	0.459
Tension	0.214	0.405
Heartburn	0.443	0.197
Horripilation	0.321	0.099
Palpitations	0.492	-0.330
Headaches	-0.002	0.422
Tingling	0.183	0.402
Pulse	0.375	-0.032
Fainting	0.216	-0.247
Oppression	0.389	-0.038
Indifference	0.335	0.106
Apprehension of the worst	0.355	-0.033
Fear of loneliness	0.400	-0.150
Fear of crowds	0.208	0.163

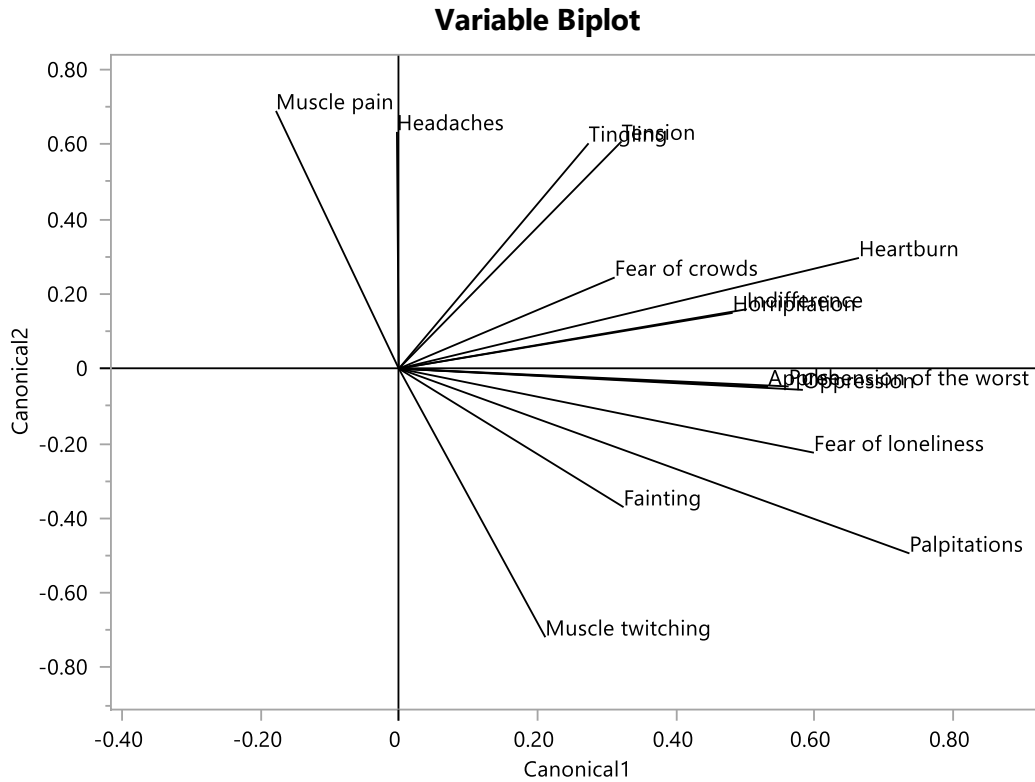


Figure 4.4 – Biplot representing a multiple of the coefficients for the fifteen standardized covariates

From this FDA produced, the two new eigenvalues associated with the first and second DFs, respectively, were 8.158 and 0.739, which corresponds to a discriminant power of 91.69% and 8.31%. The two DFs subjunct to the model were the following:

$$C_1 = 0.194 * x_1 - 0.158 * x_2 + 0.247 * x_3 + 1.107 * x_4 + 0.447 * x_5 + 0.59 * x_6 - 0.002 * x_7 + 0.254 * x_8 + 0.567 * x_9 + 0.35 * x_{10} + 0.478 * x_{11} + 0.396 * x_{12} + 0.382 * x_{13} + 0.494 * x_{14} + 0.257 * x_{15} - 6.113 \quad (4.3)$$

$$C_2 = -0.66 * x_1 + 0.616 * x_2 + 0.467 * x_3 + 0.493 * x_4 + 0.138 * x_5 - 0.396 * x_6 + 0.501 * x_7 + 0.559 * x_8 - 0.048 * x_9 - 0.4 * x_{10} - 0.046 * x_{11} + 0.125 * x_{12} - 0.035 * x_{13} - 0.185 * x_{14} + 0.201 * x_{15} - 1.797 \quad (4.4)$$

where $x_1 = \text{Muscle twitching}$, $x_2 = \text{Muscle pain}$, $x_3 = \text{Tension}$, $x_4 = \text{Heartburn}$, $x_5 = \text{Horripilation}$, $x_6 = \text{Palpitations}$, $x_7 = \text{Headaches}$, $x_8 = \text{Tingling}$, $x_9 = \text{Pulse}$, $x_{10} = \text{Fainting}$, $x_{11} = \text{Oppression}$, $x_{12} = \text{Indifference}$, $x_{13} = \text{Apprehension of the worst}$, $x_{14} = \text{Fear of loneliness}$, and, $x_{15} = \text{Fear of crowds}$. The p-value obtained for each DF, C_1 and C_2 , in the likelihood ratio test was below 0.01, justifying the use of both DF to discriminate the three groups under study.

Visualizing Table 4.7, representing the count and percentage of individuals classified in one of the three groups, it is first noticed that the misclassification rate has increased. The misclassification rate obtained for the training set of this model was 10%, where it was registered seven nervous individuals that were classified as psychotic, and eight psychotic individuals were misclassified, from which seven were classified as nervous and one as healthy.

Table 4.7 – Classification count in the training set with fifteen variables

Classification Results						
		Predicted Classification Group				Total
Group		Nervous	Psychotic	Healthy		
Original	Count	Nervous	43	7	0	50
		Psychotic	7	42	1	50
		Healthy	0	0	50	50

When submitted to the testing set previously created, the percentage of misclassification obtained was 14.(6)%, where it is important to emphasize that the model was still able to perfectly classify healthy individuals into their group, visible in Table 4.8. The biggest rate of misclassification is present in the classification of nervous individuals to psychotic individuals, and vice versa. As expected, reducing the number of variables to fifteen implies a bigger misclassification rate. On the other hand, fifteen fewer variables are needed to maintain an expressive classification rate.

Table 4.8 – Classification count in the testing set with fifteen variables

Classification Results						
		Predicted Classification Group				Total
Group		Nervous	Psychotic	Healthy		
Original	Count	Nervous	18	7	0	25
		Psychotic	3	21	1	25
		Healthy	0	0	25	25

4.4. STEPWISE DISCRIMINATION

Firstly, to detect the presence of multicollinearity in the dataset, it can be observed back in Figure 4.2 that the rays of each variable, representing the length and orientation of a positive multiple of the standardized scoring coefficient, do not follow the same orientation, which is a preliminary result that multicollinearity is not present in this dataset. To further analyze its presence, Table 4.9 reveals that multicollinearity will not be a problem among the variables. Almost all variables have *VIF* values between one and two, indicating few signs of multicollinearity. For those values whose *VIF* exceeds two, the Tolerance values are superior to 0.4, and remembering that multicollinearity is assumed when tolerance is below 0.1, then multicollinearity is not considered in the variables under study.

Table 4.9 – Tolerance and VIF values for each of the thirty variables

Variable	Tolerance	VIF
Fatigue	0.803	1.246

Variable	Tolerance	VIF
Nightmares	0.683	1.465
Muscle twitching	0.747	1.338
Cramps	0.794	1.259
Tremors	0.558	1.791
Tension	0.582	1.719
Muscle pain	0.669	1.495
Knotted throat	0.577	1.734
Satiety	0.846	1.183
Heartburn	0.569	1.758
Diarrhea	0.666	1.502
Horripilation	0.589	1.699
Palpitations	0.455	2.200
Headaches	0.672	1.487
Dizziness	0.666	1.503
Tingling	0.772	1.296
Pulse	0.603	1.658
Fainting	0.491	2.037
Eye floaters	0.699	1.430
Oppression	0.582	1.717
Indifference	0.614	1.629
Attention	0.597	1.674
Memory	0.581	1.720
Indecision	0.730	1.369
Vague anxiety	0.660	1.516
Fear of the future	0.612	1.634
Apprehension of the worst	0.443	2.259
Fear of loneliness	0.477	2.094
Other fears	0.756	1.323
Fear of crowds	0.428	2.335

The stepwise method took an optimal eighteen steps, and the following eighteen variables were chosen to take part in the stepwise model: *Fear of crowds, Fainting, Fear of loneliness, Horripilation, Nightmares, Heartburn, Palpitations, Pulse, Apprehension of the worst, Tension, Muscle twitching, Oppression, Indecision, Muscle pain, Memory, Headaches, Tingling, and Indifference*. Each variable is enumerated in the order it entered the model.

The interpretation of the data dispersion does not fall far from the other interpretations made, as there is a clear separation of the healthy individuals and the unhealthy individuals, which can be seen in Figure 4.5. The dispersion of individuals in each group is also maintained, where it is clear that the healthy individuals are closer to their group centroid when compared to either the nervous or psychotic individuals, which are more spread out.

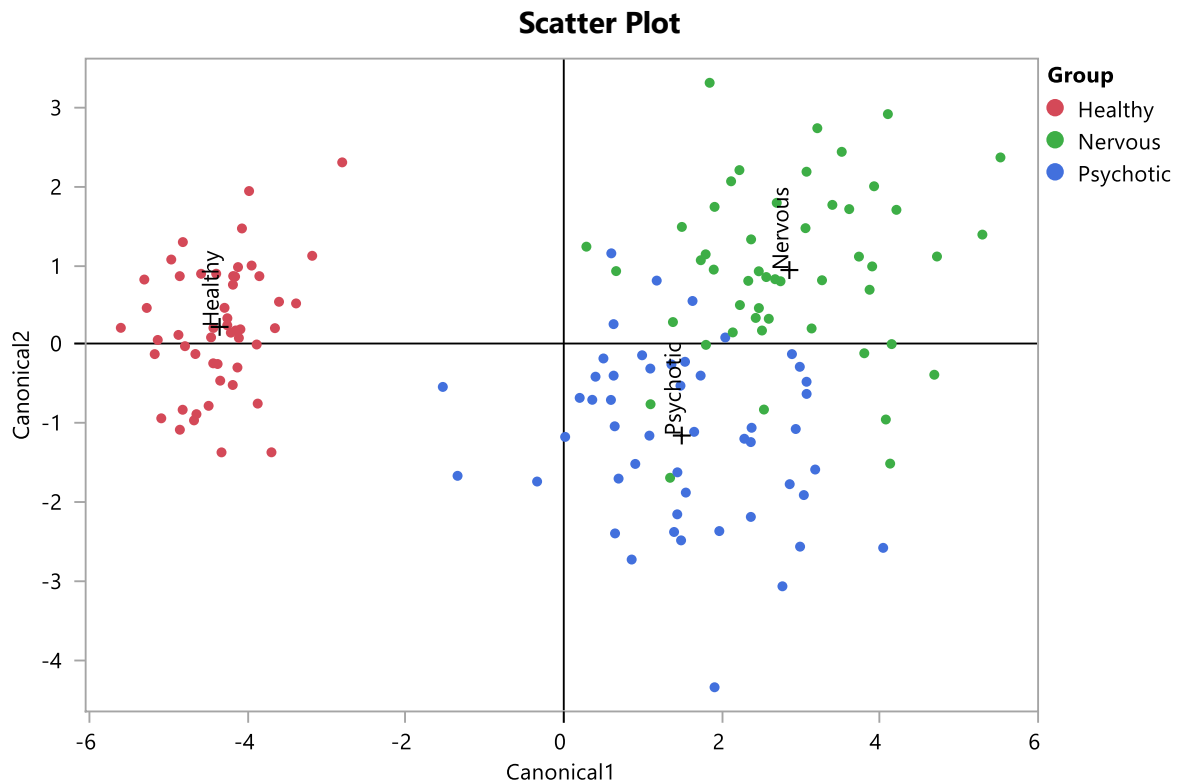


Figure 4.5 – Representation of the eighteen variables obtained from stepwise discrimination, represented in an orthogonal axis generated by the two discriminant functions

In Figure 4.6, it is observed that the orientation of the rays does not considerably change when compared to the variable biplot created when performing FDA with the thirty variables. Recording the interpretation of Figure 4.2, when analyzing how each variable contributes to the first DF, it is noticed that only the variables *Headaches* and *Muscle pain* are negatively correlated to the first DF. For the second DF, there is a bigger variation in the contributions of the variables to either the nervous or the psychotic classification. Such results are expected, as the training set did not change, only the set of pseudo-random variables used to classify individuals was altered. The length and orientation of the rays changed, but there will not be severe adjustments that may cause different variables' interpretation than the ones that have been done so far.

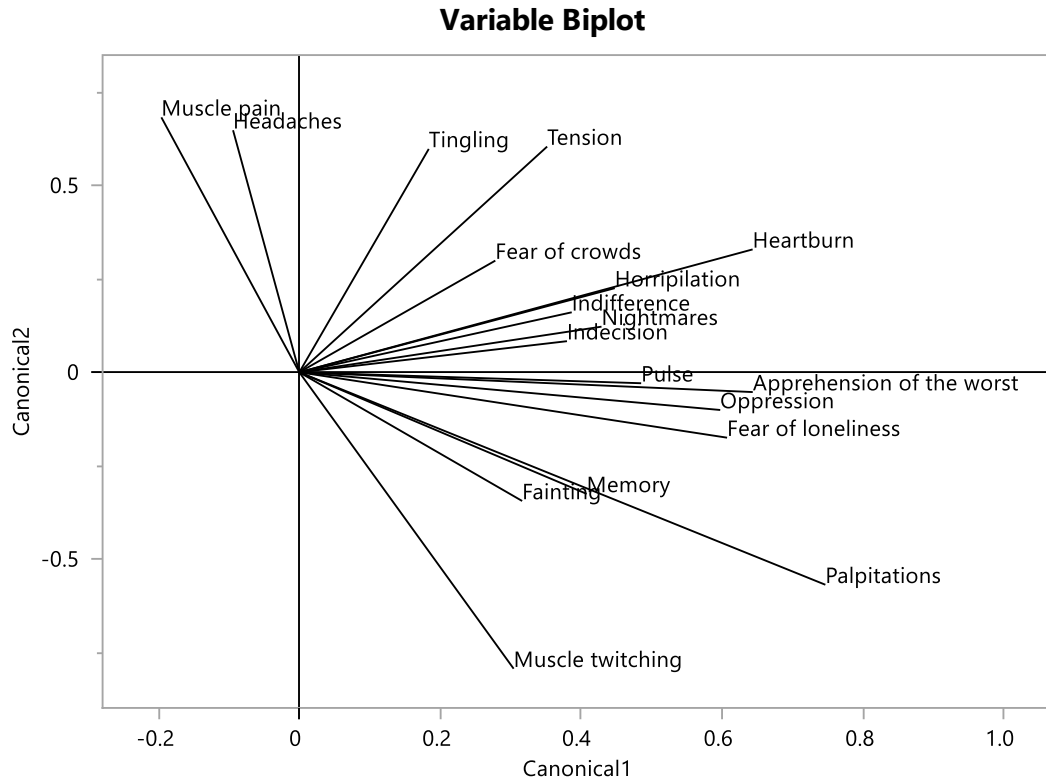


Figure 4.6 – Biplot representing a multiple of the coefficients for the eighteen standardized covariates

The eigenvalues obtained for the performed model were 9.967 and 0.772, respectively for the first and second DF, which can be interpreted as the first DF having a discriminant power of 92.81% and the second DF having a discriminant power of 7.19%. The respective first and second discriminant obtained were

$$C_1 = 0.229 * x_1 + 0.341 * x_2 + 0.5 * x_3 + 0.415 * x_4 + 0.373 * x_5 + 1.071 * x_6 + 0.597 * x_7 + 0.488 * x_8 + 0.46 * x_9 + 0.27 * x_{10} + 0.27 * x_{11} + 0.489 * x_{12} + 0.248 * x_{13} - 0.175 * x_{14} + 0.286 * x_{15} - 0.074 * x_{16} + 0.17 * x_{17} + 0.304 * x_{18} - 7.219 \quad (4.5)$$

$$C_2 = 0.247 * x_1 - 0.371 * x_2 - 0.144 * x_3 + 0.209 * x_4 + 0.106 * x_5 + 0.548 * x_6 - 0.455 * x_7 - 0.03 * x_8 - 0.038 * x_9 + 0.463 * x_{10} - 0.729 * x_{11} - 0.083 * x_{12} + 0.054 * x_{13} + 0.609 * x_{14} - 0.228 * x_{15} + 0.513 * x_{16} + 0.555 * x_{17} + 0.126 * x_{18} - 1.668 \quad (4.6)$$

where $x_1 = \text{Fear of crowds}$, $x_2 = \text{Fainting}$, $x_3 = \text{Fear of loneliness}$, $x_4 = \text{Horripilation}$, $x_5 = \text{Nightmares}$, $x_6 = \text{Heartburn}$, $x_7 = \text{Palpitations}$, $x_8 = \text{Pulse}$, $x_9 = \text{Apprehension of the worst}$, $x_{10} = \text{Tension}$, $x_{11} = \text{Muscle twitching}$, $x_{12} = \text{Oppression}$, $x_{13} = \text{Indecision}$, $x_{14} = \text{Muscle pain}$, $x_{15} = \text{Memory}$, $x_{16} = \text{Headaches}$, $x_{17} = \text{Tingling}$, and, $x_{18} = \text{Indifference}$. From the likelihood ratio test, the p-value obtained for each DF, C_1 and C_2 , was below 0.01, justifying the use of both DF to discriminate the three groups under study.

In the training set of the model, a total of fifteen individuals out of the total one hundred and fifty were misclassified, visible in Table 4.10, which corresponds to a misclassification percentage of 10%. Of the fifteen misclassified individuals, six belong to the nervous group and were classified as

psychotic, while eight psychotic individuals were classified as nervous and one psychotic as healthy. On the other hand, the individuals who belong to the healthy group were not, for once, misclassified.

Table 4.10 – Classification count in the training set with eighteen variables

Classification Results						
Group		Predicted Classification Group				Total
			Nervous	Psychotic	Healthy	
Original	Count	Nervous	44	6	0	50
		Psychotic	8	41	1	50
		Healthy	0	0	50	50

When submitted to the testing set, the model obtained a misclassification apparent of 16%. The misclassifications were all prevenient from the nervous and psychotic groups, which can be seen in Table 4.11. Equivalently to the two other models analyzed, healthy individuals are accurately classified, which is a demonstration of how the set of variables can correctly recognize the patterns and characteristics of healthy individuals.

Table 4.11 – Classification count in the testing set with eighteen variables

Classification Results						
Group		Predicted Classification Group				Total
			Nervous	Psychotic	Healthy	
Original	Count	Nervous	17	8	0	25
		Psychotic	3	21	1	25
		Healthy	0	0	25	25

In summary, the stepwise model has shown itself to be an effective classification model. On one hand, when comparing this model with the one with thirty variables, the misclassification percentage is higher in both the training and the testing set. However, the increase in the misclassification percentage is justified by the use of fewer variables.

4.5.EVALUATING MODELS

When comparing the three models with one another, it is noted that the model created with thirty predictive variables obtained the best results. The misclassification apparent rate was 4.(6)% and 10.(6)%, respectively for the training and testing set. This result was proven to be counter-productive, as the exploratory data analysis provided useful information on the variables that could be excluded from the analysis to reach the study's objective. Additionally, the interpretation of the correlation between variables and the DFs, partial F-tests, and the standardized discriminant function coefficients corroborated the suspicions raised by boxplot analysis. These four methods working together resulted in a set of fifteen variables. This new set was then used to perform FDA and create

a new, reduced model, resulting in a misclassification rate of 10% in the training set and 14.(6)% in the testing set.

To further validate if the use of the fifteen variables provided a reliable result, the model was compared to a stepwise discrimination model with eighteen variables. This stepwise model obtained a misclassification apparent rate of 10% in the training set and 16% in the testing set. As a result, the reduced model provides a more reliable model to classify individuals into one of the three groups. The misclassification rate obtained in the training set was the same, but the reduced model had a better performance in the testing set.

When comparing the reduced model to the full model, it was verified that the full model is a better classifier both in the training and in the testing set, as expected, but the complexity and accuracy of the reduced model compensate for choosing the reduced model over the full model. As the main objective of this study is to help early diagnosis of mental illnesses, such as nervousness or psychosis, the larger misclassification rate in both training and testing sets is acceptable.

5. CONCLUSIONS AND FUTURE WORKS

This research has proven once again the effectiveness of DA, more precisely FDA, to affect individuals to pre-established groups. In all three models created, it is noteworthy that the characteristics and patterns of the Healthy group were consistently recognized by each model, leading to a misclassification apparent rate of 0% for healthy individuals. While the accuracy for nervous and psychotic individuals was not as high, the model still managed to get promising results that can contribute to the diagnostic process.

For the established objectives of this paper, the first model created demonstrated a consistent ability to predict to which one of the three groups, healthy, nervous, or psychotic groups, a new individual was mostly likely considered to be. Out of the thirty initial variables used to perform FDA, it was possible to reduce half of the variables while maintaining the expected performance. In addition, it was necessary to use two DFs to effectively separate the three groups under study. The first DF's objective was to separate healthy from unhealthy individuals, while the second DF's purpose was to separate both unhealthy groups. The error rate obtained from the training set is biased, and, as a result, it was not used as a measure to validate the DFs. Nowadays, several approaches are available to face this question. A complementary validation study using different approaches to validate the DFs is being prepared. Namely, the holdout method, and the bootstrap validation will be used, which will result in an unbiased estimate of the classification rate.

Although the use of only fifteen parameters has proved to be an option for diagnosing an individual's mental state, the model should not be used exclusively to classify patients. MDs are highly delicate issues, and it is essential that supplementary studies and assessments are carried out on a case-by-case basis before any conclusions are drawn.

This study focused its methodology on FDA, but a non-parametric approach could be used, as the borderline between the assumption of multinormality subjacent to data dispersion and the equality of variance-covariance was almost crossed. For future works or projects, it could be interesting to recreate the same study without being restricted to parametric discriminant analysis assumptions. This would allow larger flexibility on the dataset used to conduct the study. The behavior of individuals may change, leading to heterogeneity of data dispersion. Investigators could try approaching data classification by ranking observations, or even in distribution-free techniques.

BIBLIOGRAPHICAL REFERENCES

- Ali, L., Zhu, C., Zhang, Z., & Liu, Y. (2019). Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network. *IEEE Journal of Translational Engineering in Health and Medicine*, 7, 1–10. <https://doi.org/10.1109/JTEHM.2019.2940900>
- Anxiety Care UK. (2017). The Biological Effects and Consequences of Anxiety. *Anxiety Care UK*. <https://anxietycare.org.uk/the-biological-effects-and-consequences-of-anxiety/>
- Batelaan, N. M., Seldenrijk, A., Bot, M., Balkom, A. J. L. M. van, & Penninx, B. W. J. H. (2016). Anxiety and new onset of cardiovascular disease: Critical review and meta-analysis. *The British Journal of Psychiatry*, 208(3), 223–231. <https://doi.org/10.1192/bjp.bp.114.156554>
- Baxter, A. J., Scott, K. M., Vos, T., & Whiteford, H. A. (2013). Global prevalence of anxiety disorders: A systematic review and meta-regression. *Psychological Medicine*, 43(5), 897–910. <https://doi.org/10.1017/S003329171200147X>
- Chen, S., Chen, G., Li, Y., Yue, Y., Zhu, Z., Li, L., Jiang, W., Shen, Z., Wang, T., Hou, Z., Xu, Z., Shen, X., & Yuan, Y. (2022). Predicting the diagnosis of various mental disorders in a mixed cohort using blood-based multi-protein model: A machine learning approach. *European Archives of Psychiatry and Clinical Neuroscience*. <https://doi.org/10.1007/s00406-022-01540-3>
- Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, 949(1), 012009. <https://doi.org/10.1088/1742-6596/949/1/012009>
- Dugas, M. J., Gagnon, F., Ladouceur, R., & Freeston, M. H. (1998). Generalized anxiety disorder: A preliminary test of a conceptual model. *Behaviour Research and Therapy*, 36(2), 215–226. [https://doi.org/10.1016/S0005-7967\(97\)00070-3](https://doi.org/10.1016/S0005-7967(97)00070-3)
- Ebner, B., & Henze, N. (2020). Tests for multivariate normality—A critical review with emphasis on weighted χ^2 -statistics. *TEST*, 29(4), 845–892. <https://doi.org/10.1007/s11749-020-00740-0>

- Esteki, M., Shahsavari, Z., & Simal-Gandara, J. (2018). Use of spectroscopic methods in combination with linear discriminant analysis for authentication of food products. *Food Control*, 91, 100–112. <https://doi.org/10.1016/j.foodcont.2018.03.031>
- Gissler, M., Laursen, T. M., Ösby, U., Nordentoft, M., & Wahlbeck, K. (2013). Patterns in mortality among people with severe mental disorders across birth cohorts: A register-based study of Denmark and Finland in 1982–2006. *BMC Public Health*, 13(1), 834. <https://doi.org/10.1186/1471-2458-13-834>
- Gomes, P. (1993). *Análise de Dados*. Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa.
- Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & Ustun, T. B. (2007). Age of onset of mental disorders: A review of recent literature. *Current Opinion in Psychiatry*, 20(4), 359–364. <https://doi.org/10.1097/YCO.0b013e32816ebc8c>
- Ma, G., Zhang, Y., Zhang, J., Wang, G., Chen, L., Zhang, M., Liu, T., Liu, X., & Lu, C. (2016). Determining the geographical origin of Chinese green tea by linear discriminant analysis of trace metals and rare earth elements: Taking Dongting Biluochun as an example. *Food Control*, 59, 714–720. <https://doi.org/10.1016/j.foodcont.2015.06.037>
- Milberger, S., Biederman, J., Faraone, S. V., Murphy, J., & Tsuang, M. T. (1995). Attention deficit hyperactivity disorder and comorbid disorder: Issues of overlapping symptoms. *The American Journal of Psychiatry*, 152, 1793–1799. <https://doi.org/10.1176/ajp.152.12.1793>
- Nasrallah, H. A. (2015). Consequences of misdiagnosis: Inaccurate treatment and poor patient outcomes in bipolar disorder. *The Journal of Clinical Psychiatry*, 76(10), e1328. <https://doi.org/10.4088/JCP.14016tx2c>
- National Institute of Mental Health. (2022, April). *Anxiety Disorders*. National Institute of Mental Health (NIMH). <https://www.nimh.nih.gov/health/topics/anxiety-disorders>

- Olatunji, B. O., Cisler, J. M., & Tolin, D. F. (2007). Quality of life in the anxiety disorders: A meta-analytic review. *Clinical Psychology Review*, 27(5), 572–581. <https://doi.org/10.1016/j.cpr.2007.01.015>
- Olson, C. L. (1974). Comparative Robustness of Six Tests in Multivariate Analysis of Variance. *Journal of the American Statistical Association*, 69(348), 894–908. <https://doi.org/10.2307/2286159>
- Remes, O., Brayne, C., van der Linde, R., & Lafortune, L. (2016). A systematic review of reviews on the prevalence of anxiety disorders in adult populations. *Brain and Behavior*, 6(7), e00497. <https://doi.org/10.1002/brb3.497>
- Renard, S. B., Huntjens, R. J. C., Lysaker, P. H., Moskowitz, A., Aleman, A., & Pijnenborg, G. H. M. (2017). Unique and Overlapping Symptoms in Schizophrenia Spectrum and Dissociative Disorders in Relation to Models of Psychopathology: A Systematic Review. *Schizophrenia Bulletin*, 43(1), 108–121. <https://doi.org/10.1093/schbul/sbw063>
- Rencher, A. C. (1995). *Methods of multivariate analysis*. New York : Wiley.
- Romed, J.-M. (1973). *Méthodes et programmes d'analyse discriminante*. Dunod.
- Saporta, G. (2011). *Probabilités, analyse des données et statistique*. Editions TECHNIP.
- Sharma, S. (1995). *Applied Multivariate Techniques*. Wiley.
- Shear, M. K. (1997). Anxiety Disorders in Women: Gender-Related Modulation of Neurobiology and Behavior. *Seminars in Reproductive Endocrinology*, 15(1), 69–76. <https://doi.org/10.1055/s-2008-1067969>
- Walker, E. R., McGee, R. E., & Druss, B. G. (2015). Mortality in Mental Disorders and Global Disease Burden Implications: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, 72(4), 334–341. <https://doi.org/10.1001/jamapsychiatry.2014.2502>
- World Health Organization. (2022, June 8). *Mental disorders*. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>

APPENDIX A

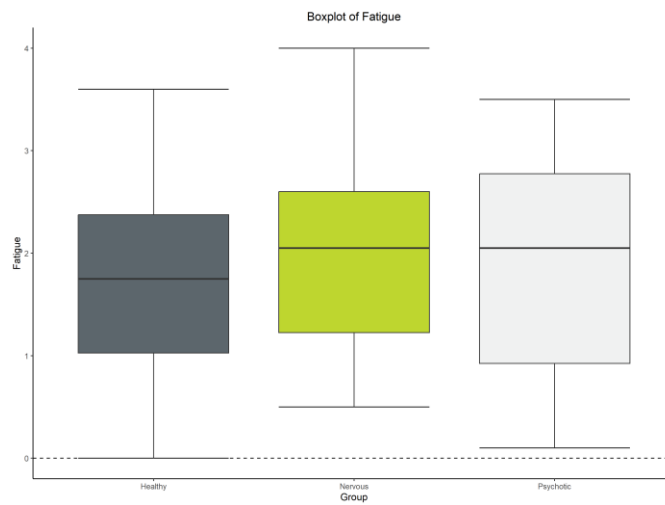


Figure A.1 – Parallel boxplot of variable Fatigue

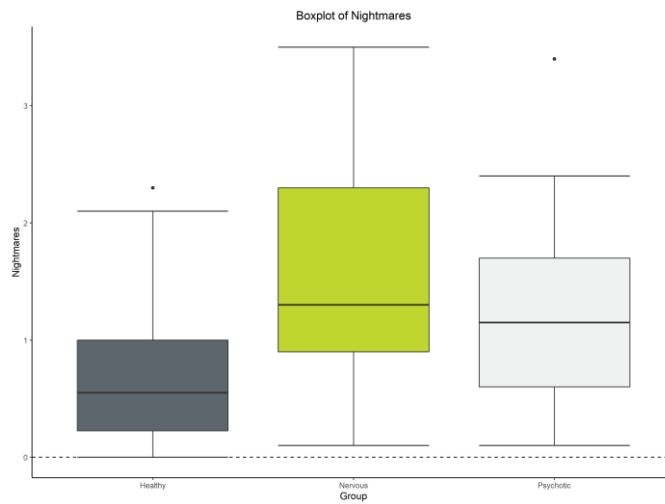


Figure A.2 – Parallel boxplot of variable Nightmares

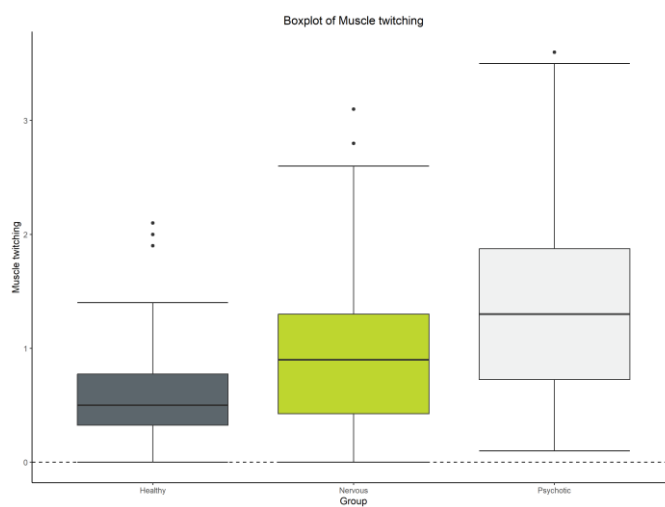


Figure A.3 – Parallel boxplot of variable Muscle twitching

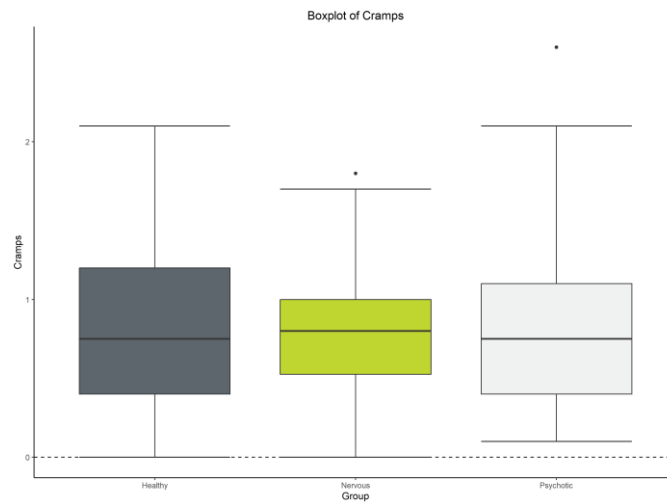


Figure A.4 – Parallel boxplot of variable Cramps

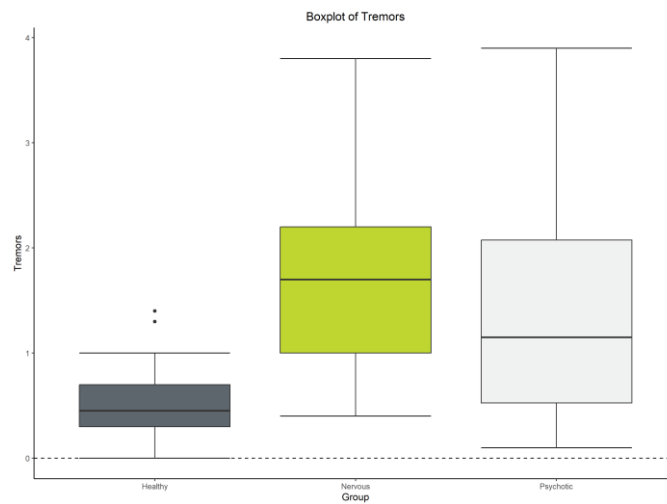


Figure A.5 – Parallel boxplot of variable Tremors

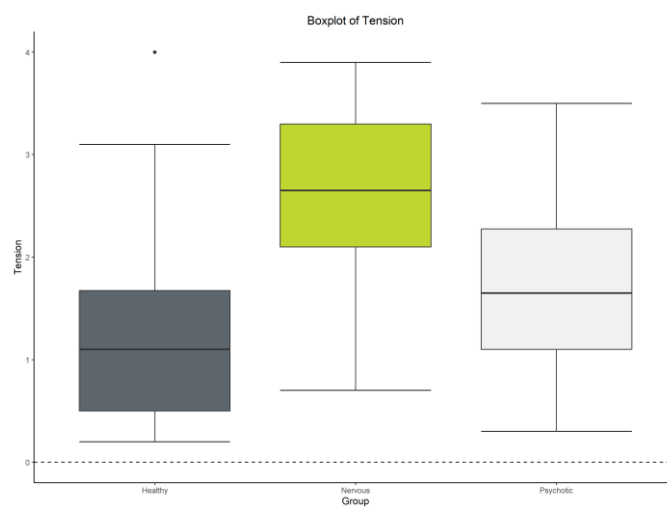


Figure A.6 – Parallel boxplot of variable Tension

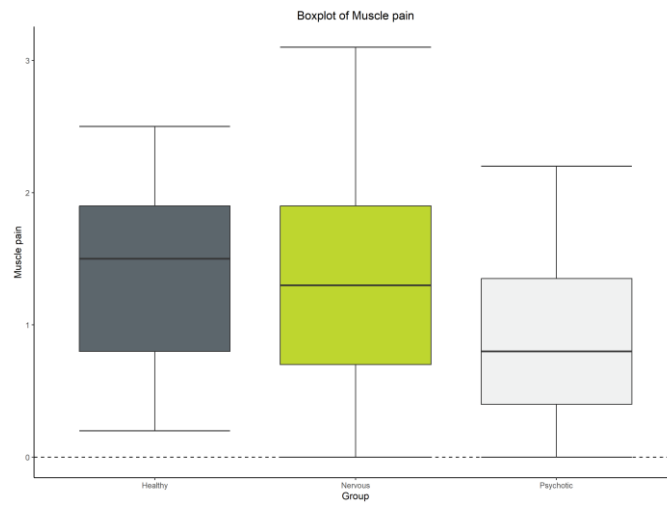


Figure A.7 – Parallel boxplot of variable Muscle pain

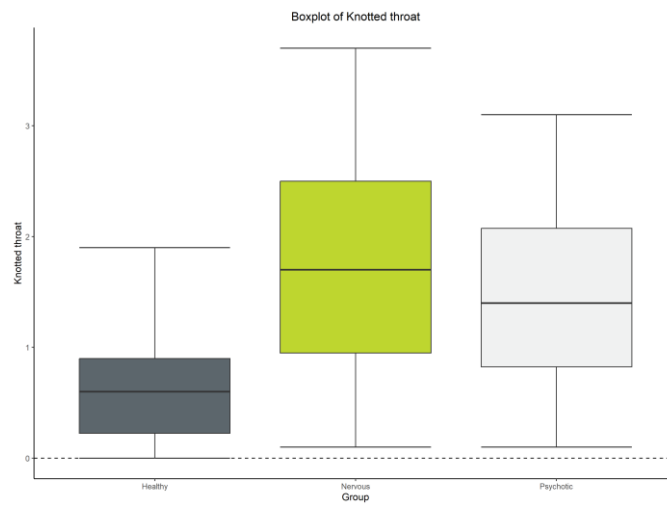


Figure A.8 – Parallel boxplot of variable Knotted throat

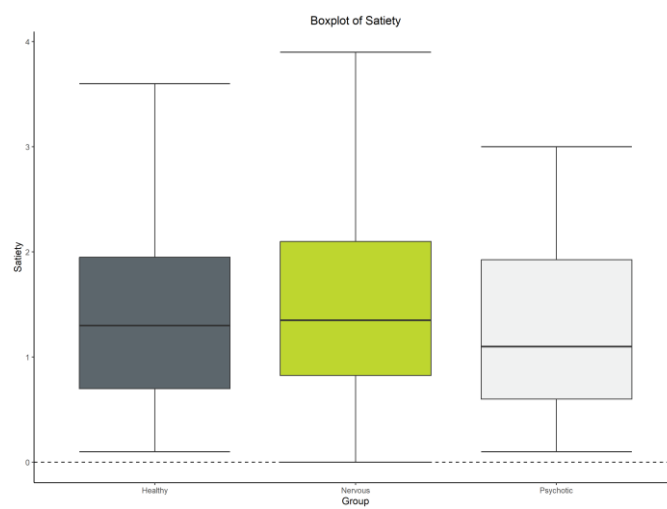


Figure A.9 – Parallel boxplot of variable Satiety

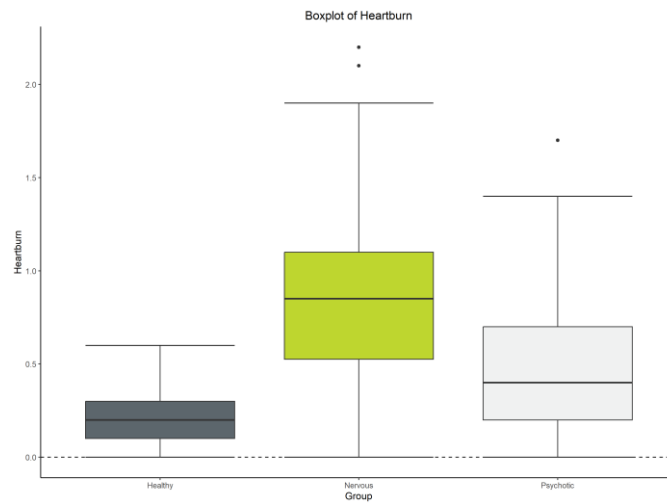


Figure A.10 – Parallel boxplot of variable Heartburn

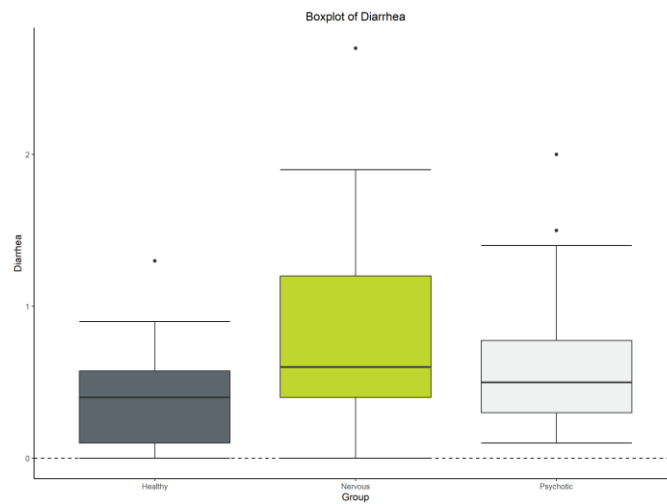


Figure A.11 – Parallel boxplot of variable Diarrhea

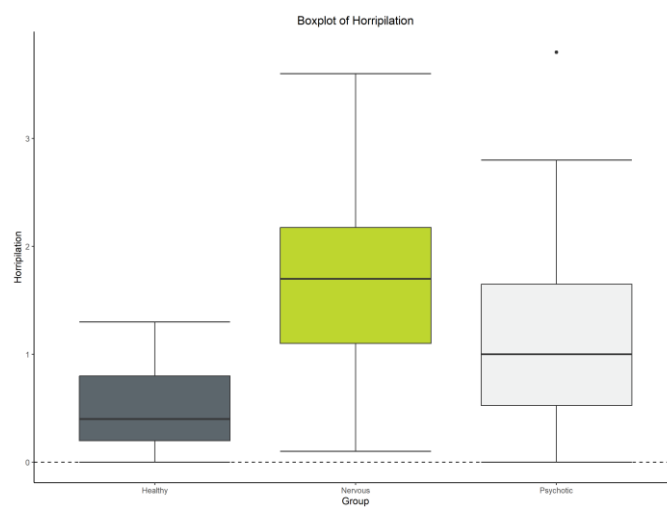


Figure A.12 – Parallel boxplot of variable Horripilation

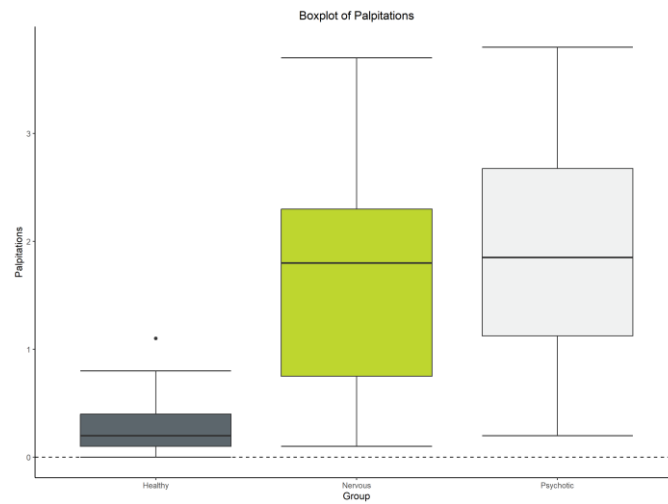


Figure A.13 – Parallel boxplot of variable Palpitations

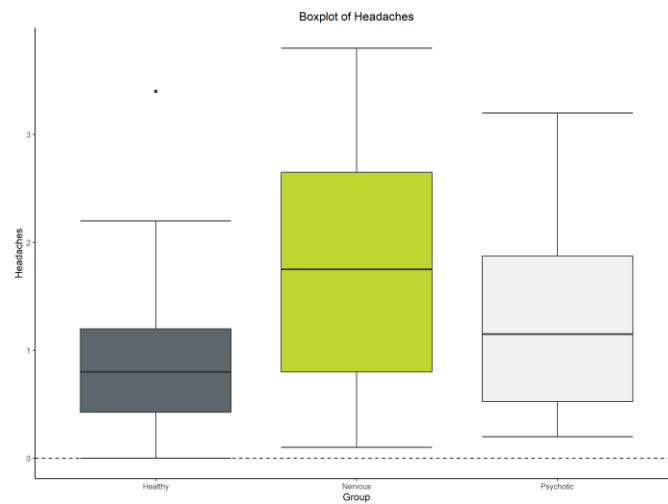


Figure A.14 – Parallel boxplot of variable Headaches

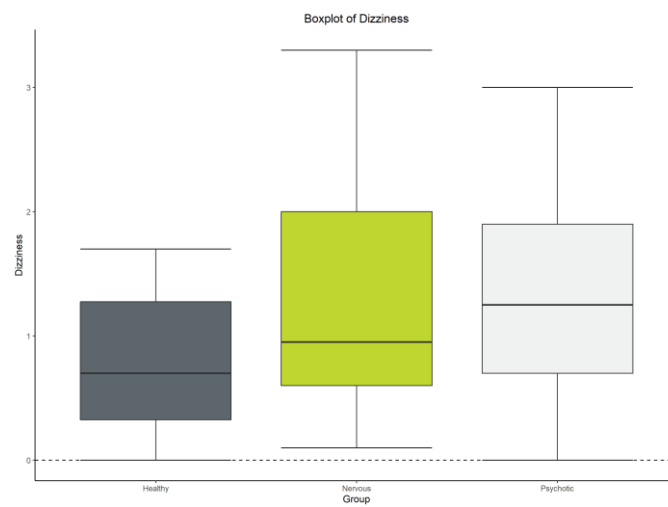


Figure A.15 – Parallel boxplot of variable Dizziness

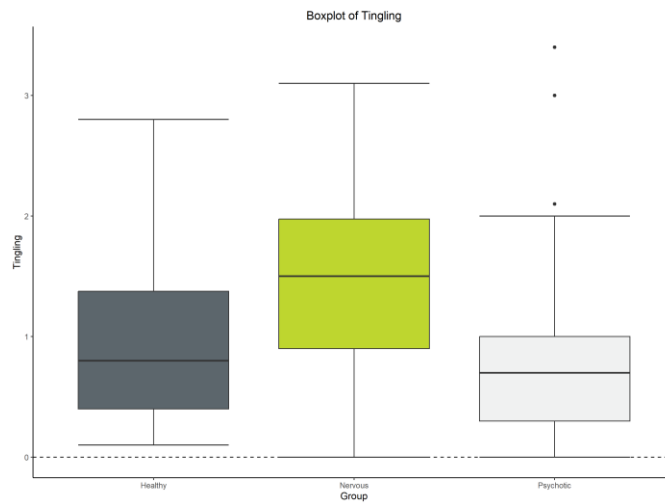


Figure A.16 – Parallel boxplot of variable Tingling

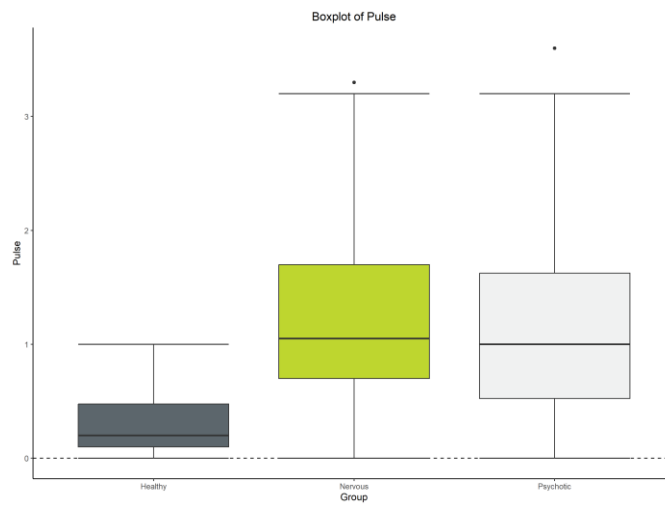


Figure A.17 – Parallel boxplot of variable Pulse

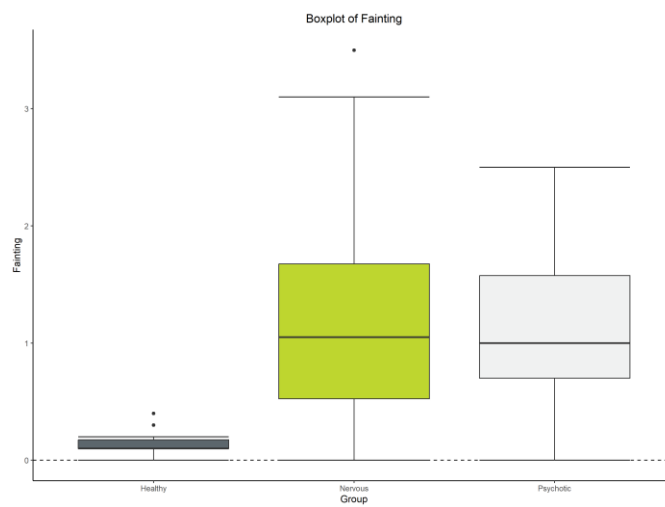


Figure A.18 – Parallel boxplot of variable Fainting

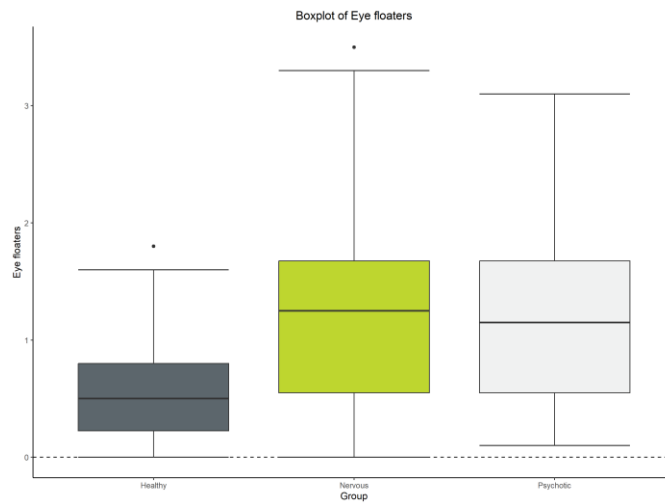


Figure A.19 – Parallel boxplot of variable Eye floaters

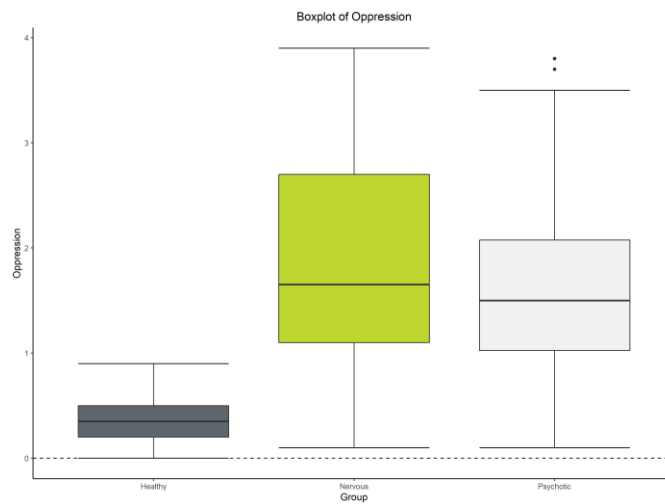


Figure A.20 – Parallel boxplot of variable Oppression

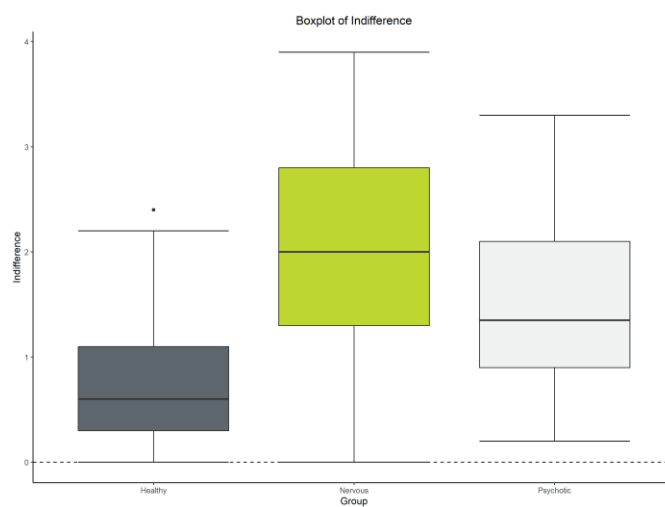


Figure A.21 – Parallel boxplot of variable Indifference

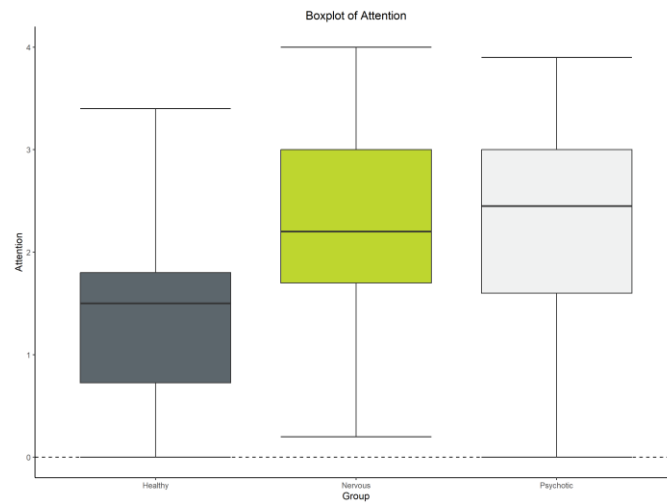


Figure A.22 – Parallel boxplot of variable Attention

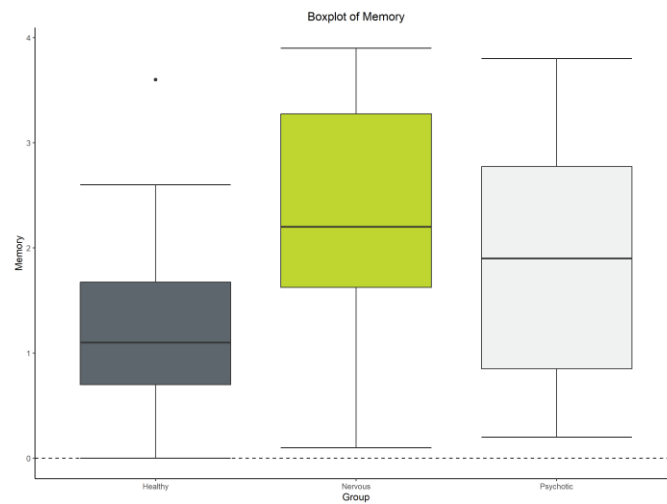


Figure A.23 – Parallel boxplot of variable Memory

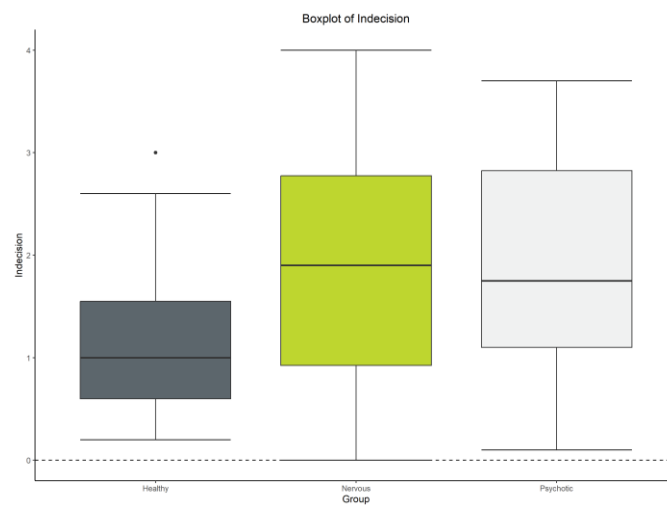


Figure A.24 – Parallel boxplot of variable Indecision

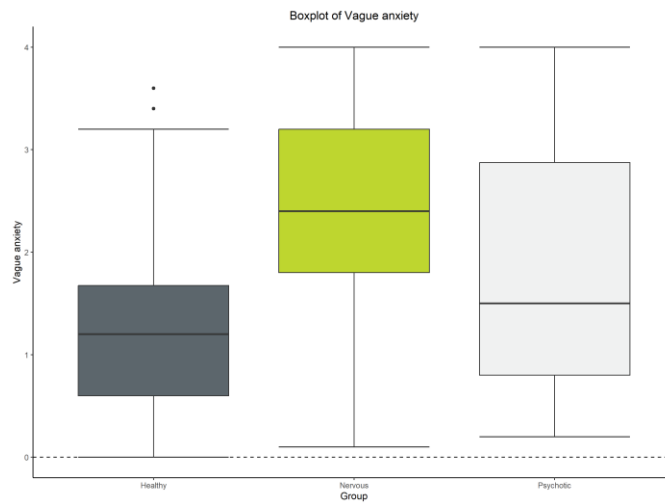


Figure A.25 – Parallel boxplot of variable Vague anxiety

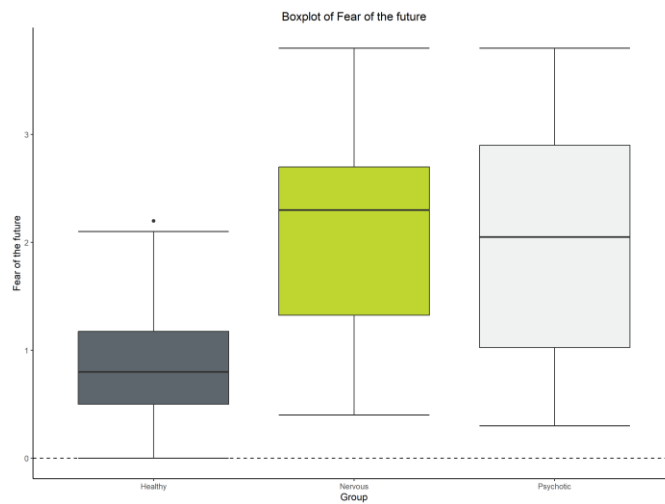


Figure A.26 – Parallel boxplot of variable Fear of the future

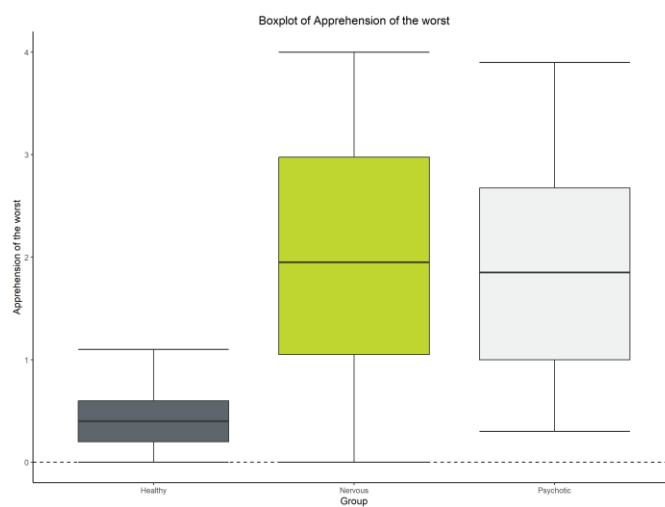


Figure A.27 – Parallel boxplot of variable Apprehension of the worst

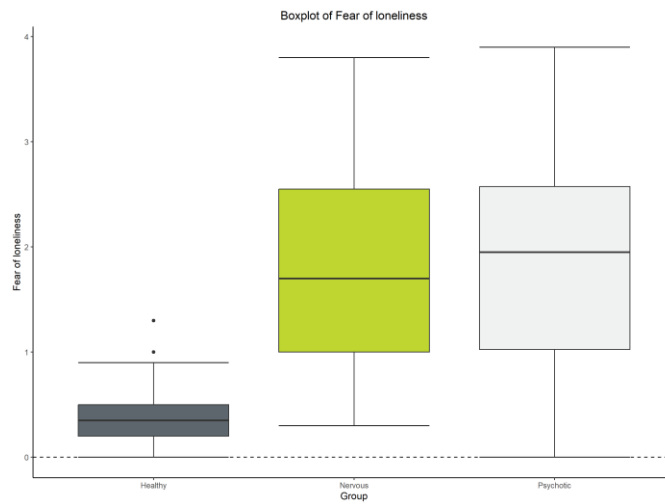


Figure A.28 – Parallel boxplot of variable Fear of loneliness

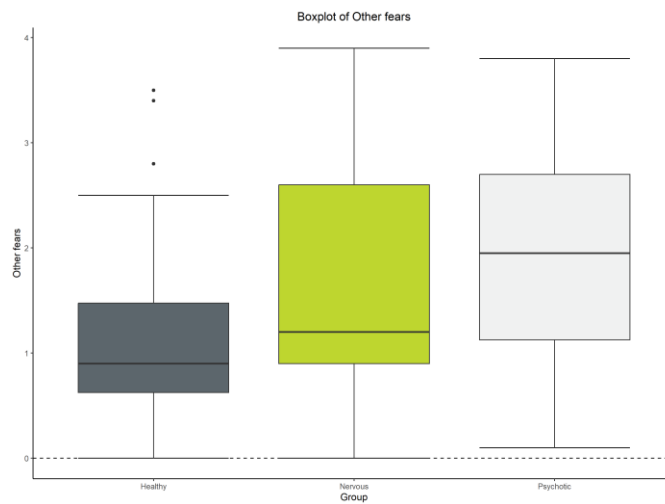


Figure A.29 – Parallel boxplot of variable Other fears

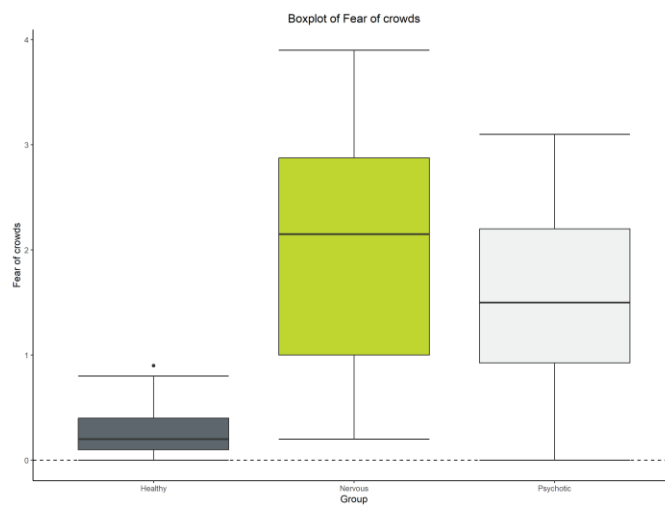


Figure A.30 – Parallel boxplot of variable Fear of crowds

ANNEXES

A) HENZE-ZIRKLER TEST

The Henze-Zirkler's statistical test has the hypothesis

$$H_0: \text{The sample can be assumed to be from a multivariate Gaussian model} \\ vs$$

$$H_1: \text{The sample cannot be assumed to be from a multivariate Gaussian model,}$$

and its test statistic "may be written in the form

$$\text{BHEP}_{n,\beta} = (2\pi)^{\frac{d}{2}} \beta^{-d} \int_{\mathbb{R}^d} \left(g_{n,\beta}(x) - \frac{1}{(2\pi\tau^2)^{\frac{d}{2}}} \exp\left(-\frac{\|x\|^2}{2\tau^2}\right) \right)^2 dx, \quad (\text{A.1})$$

where $\tau^2 = \frac{2\beta^2+1}{2\beta^2}$ and $g_{n,\beta}(x) = \frac{1}{nh^d} \sum_{j=1}^n \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x-Y_{n,j}\|^2}{2h^2}\right)$, where $h^2 = \frac{1}{2\beta^2}$. This function $g_{n,\beta}$ is convenient from a nonparametric kernel density estimator with Gaussian kernel w_1 and bandwidth h , applied to $Y_{n,1}, \dots, Y_{n,n}$ " (Ebner & Henze, 2020, pp. 859–860).

B) LIKELIHOOD RATIO TEST

The likelihood ratio test of

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r \\ vs \\ H_1: \exists(r,j), \text{ with } r \neq j \text{ and } j < r: \mu_r \neq \mu_j$$

has, under the null hypothesis, the test statistic

$$\Lambda = \frac{|W|}{|V|}, \quad (\text{B.1})$$

where W is the within-group variance and V is the total variance. The null hypothesis is rejected if Λ is inferior to the tabulated critical values of Wilks Λ (Rencher, 1995).

C) BOX M TEST

Box M test is used to test the equality of the variance-covariance matrices of groups, being its hypothesis

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_r \\ vs \\ H_1: \exists(r,j), \text{ with } r \neq j \text{ and } j < r: \Sigma_r \neq \Sigma_j$$

where r is the total number of groups under study. From Saporta (2011), under the null hypothesis, the test statistic is

$$M = C \left[\left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n - k} \right) (n - k) \det \left(\frac{n}{n - k} W \right) - \sum_{i=1}^r (n_i - 1) \det \left(\frac{n_i}{n_i - k} V_i \right) \right], \text{ where } C = \left(1 - \frac{(2p^2 + 3p - 1)(r + 1)}{6(p + 1)rn} \right) \quad (\text{C.1})$$

and follows approximately a χ^2 with $\frac{(r-1)*p*(p+1)}{2}$ degrees of freedom.

D) PARTIAL F-VALUES

In the case of several groups under study, the partial λ of function x_r adapted is given by

$$\lambda(x_r | x_1, \dots, x_{r-1}, x_{r+1}, \dots, x_p) = \frac{\lambda_p}{\lambda_{p-1}} \quad (\text{D.1})$$

being λ_{p-1} the Wilks' λ for all variables except x_r and λ_p the Wilks' for all variables. The partial F-value is

$$F = \frac{1 - \lambda}{\lambda} * \frac{v_E - p + 1}{v_H}, \text{ where } v_E = N - k \text{ and } v_H = k - 1 \quad (\text{D.2})$$

and follows a distribution of $F_{v_H, v_E - p + 1}$ (Rencher, 1995).

E) STANDARDIZED COEFFICIENTS

The standardized variables are calculated by

$$z_r = \alpha_r^* \frac{y_{1ir} - \bar{y}_{1r}}{s_r}, \text{ where } i = 1, 2, \dots, n \text{ and } r = 1, 2, \dots, p \quad (\text{E.1})$$

and α_r^* is the eigenvector of the matrix $W^{-1}B$, so the influence of each variable with the remaining ones is taken into account (Rencher, 1995).

F) STEPWISE DISCRIMINANT ANALYSIS

In practical terms, the first step is to identify the variable which maximizes the univariate F, meaning the first variable is the best variable that separates the three groups by itself. In the second step, it is calculated the partial Λ ,

$$\Lambda(y_i | y_1) = \frac{\Lambda(y_1, y_i)}{\Lambda(y_1)} \quad (\text{F.1})$$

and the underlying partial F

$$F = \frac{1 - \Lambda(y_i, y_1)}{\Lambda(y_i, y_1)} * \frac{v_E - 1}{v_H}, \text{ where } v_E = N - k \text{ and } v_H = k - 1 \quad (\text{F.2})$$

where y_1 represents the variable selected in the first step, and y_i represents the remaining variables under study. The variable y_i which maximizes F will be the second selected variable to enter the

model. Meanwhile, it is verified if the first variable is still significant, calculating the partial Λ and F for y_1 adjusted for y_2 . This procedure will be iterated until the stabilization of the optimal process.

G) VARIANCE INFLATION FACTORS

VIF is a method used to calculate multicollinearity in regression analysis. It is an indicator of how much the variance of the predictor's coefficients is inflated due to multicollinearity. It can be calculated by

$$VIF = \frac{1}{1 - R_i^2} \quad (G.1)$$

where R_i^2 is the coefficient of determination obtained from the regression of the i^{th} variable against all others. *VIF* can assume one of the following three scenarios: it can either be 1, indicating there is no correlation amongst variables; it can be between 1 and 5, indicating the variables are moderately correlated; or it can be bigger than 5, indicating a high correlation between variables. Tolerance, $1 - R_i^2$, can also be a good value to interpret multicollinearity, where values below 0.1 indicate collinearity (Daoud, 2017).