UNIVERSIDADE FEDERAL DE ALFENAS

FARLEY SILVA FLAUSINO

NOVA FORMULAÇÃO DE FERRAMENTAS DE ESTATÍSTICA MULTIVARIADA COM INCERTEZAS EXPERIMENTAIS

Poços de Caldas/MG
2018

FARLEY SILVA FLAUSINO

NOVA FORMULAÇÃO DE FERRAMENTAS DE ESTATÍSTICA MULTIVARIADA COM INCERTEZAS EXPERIMENTAIS

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Física pelo Programa de Pós-Graduação em Física da Universidade Federal de Alfenas. Área de concentração: Física de Partículas e Campos. Orientador: Prof. Dr. Cássius Anderson Miquele de Melo.

Poços de Caldas/MG
2018

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal de Alfenas
Biblioteca campus Poços de Caldas

F587n Flausino, Farley Silva.
Nova formulação de ferramentas de estatística multivariada com incertezas experimentais / Farley Silva Flausino. -- Poços de Caldas/MG, 2018.
110 f. –

Orientador(a): Cássius Anderson Miquele de Melo.
Dissertação (Mestrado em em Física) – Universidade Federal de Alfenas, campus Poços de Caldas, 2018.
Bibliografia.

1. Incerteza experimental. 2. Análise multivariada. 3. Análise discriminante. 4. Análise de componentes principais. 5. Correlação canônica (Estatística). I. Melo, Cássius Anderson Miquele de. II. Título.

CDD – 519.5

Ficha Catalográfica elaborada por Thaís Aparecida de Lima
Bibliotecária-Documentalista CRB6/3032

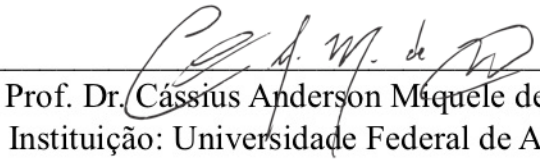
FARLEY SILVA FLAUSINO

NOVA FORMULAÇÃO DE FERRAMENTAS DE ESTATÍSTICA MUTLIVARIADA COM INCERTEZAS EXPERIMENTAIS

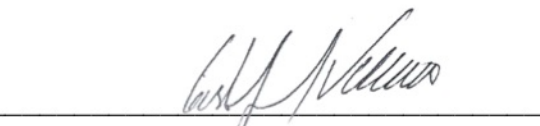
A banca examinadora abaixo-assinada, aprova a Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Física, pelo Programa de Pós-Graduação em Física da Universidade Federal de Alfenas.

Área de Concentração: Física de Partículas e Campos.

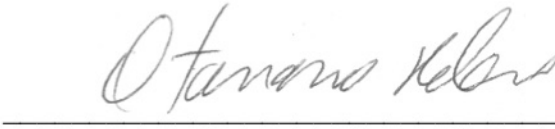
Aprovada em: 16 de março de 2018.



Prof. Dr. Cássius Anderson Miquele de Melo
Instituição: Universidade Federal de Alfenas



Prof. Dr. Gustavo do Amaral Valdiviesso
Instituição: Universidade Federal de Alfenas



Prof. Dr. Otaviano Helene
Instituição: Universidade de São Paulo

Dedico este trabalho ao meu pai Jamilson, a minha mãe Leila e a minha companheira Tati.

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, ao meu pai, Jamilson, que sempre me apoiou, acreditou em mim, nos meus sonhos e não mediu esforços para que eu pudesse realizá-los, incluindo todas as etapas da minha vida acadêmica. Também à minha mãe, Leila, que sempre esteve ao meu lado e me apoiou em todas as minhas decisões. Sem meus pais, eu jamais teria chegado até aqui e não há palavras para agradece-los por tudo que fizeram e fazem por mim.

Meus agradecimentos também vão à minha companheira, Tati, por ser uma pessoa maravilhosa, estar sempre ao meu lado em cada decisão, em todos os momentos, principalmente os mais difíceis, por ter me incentivado a iniciar esta pós-graduação e ter revisado comigo incontáveis vezes todo o meu trabalho. Sua ajuda sempre foi essencial para que eu realizasse meus objetivos e principalmente me tornasse uma pessoa melhor a cada dia.

Agradeço ao meu orientador, Prof. Dr. Cássius Anderson Miquele de Melo, por toda paciência e dedicação em transmitir seus conhecimentos e ter sido fundamental em minha formação desde o período da graduação. Se hoje estou concluindo mais esta etapa é graças aos seus conselhos e orientações que me guiaram e me ajudaram a descobrir quais caminhos eu gostaria de seguir em minha carreira profissional.

Meus sinceros agradecimentos também vão aos membros do Programa de Pós-Graduação em Física, em especial aos docentes Prof. Dr. Fernando Gonçalves Gardim, Prof. Dr. Gustavo do Amaral Valdiviesso, Prof. Dr. Cássius Anderson Miquele de Melo e Prof. Dr. Alencar José de Faria pelas aulas ministradas e todo esforço para que seus alunos pudessem evoluir. Também agradeço a todos os meus amigos que de alguma maneira fizeram parte de toda minha evolução, tanto aos amigos que fiz durante esses dois anos, na pós-graduação, quanto aqueles que já estavam ao meu lado há mais tempo em minha jornada.

Por fim, agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

"Prefiro morrer do que perder a vida."
(CHESPERITO, 1976)

RESUMO

Quando se deseja analisar um conjunto de dados medidos, assumindo a aleatóriadade das medições, levando em conta os erros estatísticos e instrumentais envolvidos no processo, as incertezas experimentais exercem um papel fundamental nos resultados de algumas análises estatísticas. Entretanto, muitas ferramentas estatísticas não as levam em conta em seus cálculos e, por isso, este estudo tem como objetivo inseri-las nos cálculos das análises de Componentes Principais, Discriminante Linear de Fisher e de Correlação Canônica, bem como analisar o impacto no resultado final destas técnicas. Como as três análises têm em comum o fato de seus resultados estarem ligados à matriz de covariância dos dados, o procedimento metodológico deste estudo consistiu em utilizar a média ponderada das variáveis, por suas incertezas experimentais, para construir as matrizes de covariância. Já para propagar esses erros para os resultados das três análises, optou-se por utilizar um método numérico *a la* Monte Carlo, através de algoritmos desenvolvidos para gerar resultados aleatórios a partir da flutuação da média ponderada dos dados. A fim de demonstrar a aplicabilidade do novo modelo de componentes principais, foram refeitas as análises de componentes principais, das variáveis que caracterizam o meio interestelar difuso, realizadas por Ensor et al. (2017) e comparados os resultados com a abordagem tradicional que não leva em conta as incertezas experimentais. Este novo modelo de componentes principais propiciou uma forma alternativa de escolher o número de componentes a ser utilizado, através dos valores obtidos para as incertezas relativas às proporções explicativas acumuladas. Já para as outras duas análises foram realizadas simulações para avaliar a aplicabilidade do método em exemplos desenvolvidos pelo autor. A análise discriminante foi a única ferramenta que apresentou uma mudança na sua interpretação, fornecendo como resposta a probabilidade de novas observações pertencerem a cada um dos grupos e não uma classificação determinística. Já a análise de correlações canônicas permitiu uma avaliação dos dados mais próximo da realidade do experimento, uma vez que tanto as variáveis canônicas e vetores de transformação quanto as correlações canônicas, possuem incertezas. Portanto, pôde-se concluir que nas três análises a inserção das incertezas experimentais possibilitou ao pesquisador uma interpretação dos resultados mais condizente com a realidade do experimento, podendo evitar uma super ou subestimação de parâmetros na análise dos dados.

Palavras-chave: Incerteza experimental. Estatística Multivariada. Ánalise Discriminante. Componentes Principais. Correlação Canônica.

ABSTRACT

When a researcher wants to analyze a set of data, assuming the randomness of the measurements, taking into account the statistical and instrumental errors involved in the process, experimental errors have a key role in the results of some statistical analysis. However, many statistical tools do not take them into account in their calculations and, therefore, this study proposes new formulations for the Principal Components, Fisher Linear Discriminant and Canonical Correlation analysis which take the experimental errors into account, and also proposes to evaluate the impact on the results of these new techniques. Since the three analysis have in common the fact that their results are tied to the data covariance matrix, the methodological procedure of this study consisted of using the weighted average of the variables by their experimental errors, in order to construct the covariance matrices. For purposes of propagating these errors to the results of the three analysis, it was chosen to use a numerical method similar to Monte Carlo, through algorithms developed to generate random results from the fluctuation of the data weighted average. In order to demonstrate the applicability of the new principal components model, it was reconstructed the principal component analysis of the variables for the diffuse interstellar band performed by Ensor et al. (2017) and the results were compared with the traditional approach that does not take into account the experimental errors. This new model of principal components provided an alternative way to choose the number of components to be used, through the values obtained for the relative errors concerning to the accumulated proportion of variance explained. For the other two analysis, simulations were performed to evaluate the applicability of the method in examples developed by the author. The discriminant analysis was the only technique that presented a change in its interpretation, providing as answer the probability of new observations belonging to each group and not a deterministic classification. The analysis of canonical correlations allowed for an evaluation of the data closer to the reality of the experiment, once both canonical variables and transformation vectors as well as canonical correlations have available now error bars. Therefore, it was possible to conclude that in the three analysis the insertion of the experimental errors enabled the researcher an interpretation of the results faithful to the real experiment, which may avoid a super or underestimation of parameters in the data analysis.

Keywords: Experimental error. Multivariate Statistics. Discriminating Analysis. Principal Components. Canonical Correlation.

LISTA DE FIGURAS

Figura 1 – Gráfico de tensão por corrente, onde $p0$ e $p1$ representam o coeficiente angular e linear, respectivamente, da reta ajustada. 19

Figura 2 – Gráfico de tensão por corrente, onde $p0$ e $p1$ representam o coeficiente angular e linear, respectivamente, da reta ajustada. 20

Figura 3 – Representação gráfica de uma função densidade de probabilidade normal bivariada. 23

Figura 4 – Um *Scree Plot*, gráfico de autovalor por componente principal. 36

Figura 5 – Uma representação gráfica da LDA aplicada em duas populações com duas variáveis discriminantes, onde $\bar{\boldsymbol{x}}$ equivale ao centroide $\hat{\boldsymbol{\mu}}$ de cada população e $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ é a metade da distância de Mahalanobis m. . . 40

Figura 6 – Tabela de dados com oito variáveis de comprimento de onda (todas em $10^{-13}m$), onde cada linha representa uma observação (estrela) e cada coluna uma variável. 61

Figura 7 – Tabela de dados com oito variáveis de comprimento de onda (todas em $10^{-13}m$), onde cada linha representa uma observação (estrela) e cada coluna uma variável. 62

Figura 8 – Tabela de dados com seis variáveis, onde cada linha representa uma observação (estrela) e cada coluna uma variável. 63

Figura 9 – Regressão linear para o conjunto de dados das variáveis $E(B-V)$, em mag, e $N(H)$, em cm^{-2}, onde $p0$ representa o coeficiente angular da reta, $p1$ o coeficiente linear e χ^2/ndf é o Chi-quadrado dividido pelo número de graus de liberdade. 66

Figura 10 – Histogramas com as distribuições dos autovalores referentes as $PC1$(vermelho) e $PC2$(azul) da PCA aplicada ás variáveis $E(B-V)$ e $N(H)$, onde *Entries*, *Mean* e *RMS* representam o número de dados colocados nos histogramas, a média aritmética deles e a raíz do valor quadrático médio respectivamente. 67

Figura 11 – Histogramas com as distribuições das proporções explicativas da $PC1$(vermelho) e $PC2$(azul). 67

Figura 12 – Ajuste a partir da equação (6.25) com $PC2 = 0$ (reta azul), ajuste obtido por Ensor et al. (2017) pela equação (6.28) (reta verde claro), regressão linear (reta verde escuro), dada pela equação (6.27), em que $p0$ e $p1$ são os coeficientes angular e linear, respectivamente, e função proposta por Bohlin, Savage e Drake (1978) dada pela equação (6.30), (reta vermelha), onde $[N(H)] = [cm^{-2}]$ e $[E(B-V)] = [mag]$. 71

Figura 13 – Da esquerda para a direita estão as distribuições de frequências do 23º ao 17º autovalor. 72
Figura 14 – Da esquerda para a direita estão as distribuições de frequência do 16º ao 6º autovalor. 72
Figura 15 – Da esquerda para a direita estão as distribuições de frequência do 5º ao 2º autovalor. 73
Figura 16 – Distribuição de frequência do autovalor correspondente à PC1. 73
Figura 17 – Distribuição de frequência do 15º (azul) e 16º (marrom) autovalor. . . 74
Figura 18 – Distribuição de frequência do 18º (vermelho claro) e 19º (azul claro) autovalor. 74
Figura 19 – Distribuição de frequência do 21º (laranja) e 22º (preta) autovalor. . . 74
Figura 20 – Gráfico das incertezas relativas as proporções acumuladas de cada componente principal. 77
Figura 21 – Gráfico *Scree Plot* de autovalores por número da componente principal. 78
Figura 22 – Componentes principais 1(vermelho) e 2(azul) referentes à estrela $HD15137$. 79
Figura 23 – Diagrama com a representação da nova função discriminante aplicada ao caso de dois grupos e duas variáveis discriminantes. 82
Figura 24 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 1%. 84
Figura 25 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 5%. 85
Figura 26 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 15%. 85
Figura 27 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 25%. 86
Figura 28 – Distribuições de frequência, individuais, dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 1%. 86
Figura 29 – Distribuições de frequência, individuais, dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 15%. 87
Figura 30 – Distribuições de frequência dos *scores* discriminantes das observações 1 (vermelho) e 2 (azul). 88
Figura 31 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 3%. 92

Figura 32 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 5%. 93
Figura 33 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 10%. 93
Figura 34 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 15%. 94
Figura 35 – Distribuições de frequência do primeiro par de variáveis canônicas (U_1 e V_1) das variáveis X_1 (vermelho) e Y_1 (azul). 97

LISTA DE TABELAS

Tabela 1 – Matriz de classificação onde as linhas representam as populações reais e as colunas as populações classificadas. 42

Tabela 2 – Resultados da PCA para as variáveis $E(B-V)$ e $N(H)$, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores, as proporções explicativas, as proporções acumuladas, as incertezas relativas do percentual acumulado e os autovetores, respectivamente, juntamente com suas incertezas. 68

Tabela 3 – Resultados da PCA para as variáveis $E(B-V)$ e $N(H)$, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores, as proporções explicativas, as proporções acumuladas e os autovetores. 68

Tabela 4 – Resultados da PCA para as 23 variáveis, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores e suas incertezas, as proporções explicativas e suas incertezas, as proporções acumuladas e suas incertezas e as incertezas relativas. 75

Tabela 5 – Resultados da PCA para as 23 variáveis, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores, as proporções explicativas e as proporções acumuladas. 76

Tabela 6 – Resultados da análise com amostra de treinamento. Da esquerda para a direita estão os coeficientes discriminantes , suas incertezas, o centroide da população 1, sua incertezas, o centroide da população 2 e sua incerteza. 87

Tabela 7 – *Scores* discriminantes das observações 1 (coluna 1) e 2 (colunas 2). . . 88

Tabela 8 – Resultados das probabilidades de pertencer a cada um dos grupos, onde cada linha representa uma nova observação e as colunas dois e três representam as probabilidades. 89

Tabela 9 – Dados dos vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$, onde cada linhar representa uma observação. 90

Tabela 10 – Resultados obtidos para as médias ponderadas (linha 2) e suas incertezas (linha 3) para cada uma das variáveis dos vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$ (linha 1). 91

Tabela 11 – Resultados dos autovalores, suas incertezas (colunas 2 e 3), correlações canônicas e suas incertezas (colunas 4 e 5). 94

Tabela 12 – Resultados dos autovetores de transformação linear e suas incertezas. . 95

Tabela 13 – Pares de variáveis canônicas das 20 observações da análise realizada com dados com incertezas relativas de 10%. 96

Tabela 14 – Correlações entre as variáveis canônicas e cada uma das variáveis originais dos dois grupos de dados $\boldsymbol{X}$ e $\boldsymbol{Y}$, bem como suas incertezas estimadas. 97
Tabela 15 – Proporções explicativas em termos de variância total, para cada variável canônica individual. 98

SUMÁRIO

1 INTRODUÇÃO 17

2 CONCEITOS DE ESTATÍSTICA MULTIVARIADA 21
2.1 Amostras Aleatórias Multidimensionais 21
2.2 Distribuição Normal Multivariada 22
2.3 Estimadores de Máxima Verossimilhança do vetor médio e da matriz de covariância 24
2.4 Teste de Normalidade Multivariada 27
2.5 Propagação de incertezas 28

3 ANÁLISE DE COMPONENTES PRINCIPAIS 30
3.1 Análise de Componentes Principais por matriz de Covariância . . . 31
3.2 Análise de Componentes Principais por Matriz de Correlação . . . 33
3.3 Escolha das Componentes Principais 34

4 ANÁLISE DISCRIMINANTE 37
4.1 Análise Discriminante Linear de Fisher para duas populações . . . 37
4.2 Estimativa das Probabilidades de Classificação Incorreta 41
4.3 Análise Discriminante Linear de Fisher para mais de duas populações 43

5 ANÁLISE DE CORRELAÇÃO CANÔNICA 47
5.1 Análise de Correlações Canônicas por variáveis não padronizadas 48
5.2 Análise de Correlações Canônicas por variáveis padronizadas . . . 52
5.3 Número de pares de variáveis canônicas e a qualidade de um modelo reduzido 54

6 DESENVOLVIMENTO DAS NOVAS FORMULAÇÕES 58
6.1 Análise de Componentes Principais 60
6.2 Análise Discriminante 79
6.3 Análise de Correlação Canônica 90

7 CONSIDERAÇÕES FINAIS 99

REFERÊNCIAS 101

APÊNDICES 105

APÊNDICE A – PCS DAS 29 ESTRELAS PARA A ANÁLISE COM DUAS VARIÁVEIS 106

APÊNDICE B – PCS DAS 29 ESTRELAS PARA A ANÁLISE COM 23 VARIÁVEIS 107

PREFÁCIO

A ideia de realizar este trabalho nasceu em um projeto de iniciação científica, quando eu ainda estava na graduação de Bacharelado Interdisciplinar em Ciência e Tecnologia. Ao trabalhar com algumas análises de estatística multivariada as seguintes questões foram levantadas: por que utilizamos as incertezas experimentais em algumas análises como regressão linear simples e não as utilizamos em outras análises como discriminante, correlação canônica e outras análises multivariadas? E como poderíamo inserí-las nestas análises?

Buscar responder a segunda questão ficou, então, como proposta do meu orientador, Prof. Dr. Cássius Anderson Miquele de Melo, para um possível trabalho de mestrado caso eu decidisse seguir este caminho. Foi um trabalho desafiador nesses últimos dois anos e com certeza há muito o que trabalhar. Entretanto, a medida que eu ia estudando, o que me motivava a continuar a pesquisar era descobrir as diversas aplicações destas ferramentas estatísticas nas mais diferentes áreas da ciência, além da Física.

Portanto, ao ler este trabalho, o leitor irá perceber que os cálculos de cada método está, na medida do possível, bem detalhado. Ao desenvolver esta dissertação eu viso como público não só pessoas da Física, Estatísitica ou das ciências exatas em geral, mas também pesquisadores de outras áreas que queiram conhecer um pouco mais dos assuntos abordados neste trabalho. Assim, espero que este texto sirva, além das constribuições dos resultados das pesquisas desenvolvidas, também, como fonte de estudo para aqueles que se interessarem pelo assunto.

1 INTRODUÇÃO

A Estatística é indispensável e fundamental para análise e interpretações de dados em todos os ramos da ciência, desde estudos sociais até as mais diversas áreas das ciências exatas, e em particular na Física (MAGALHÃES; LIMA, 2008). No entanto, esta não é uma área de estudo nova, segundo Castro (1970), a estatística é tão antiga quanto os primeiros homens, uma vez que a necessidade de enumerar e organizar seus objetos sempre esteve presente na natureza humana.

A origem da palavra estatística, que deriva da palavra "estado", em latim, se deu em torno de 1730, e foi proposta por Schmeitzel na Universidade de Iena, no entanto, até hoje há uma controvérsia se a palavra se referia ao Estado como organização política ou ao estado como um modo de ser (CASTRO, 1970).

Há registros de recenseamentos realizados por grupos sumérios entre 5000 e 2000 anos antes de Cristo, contudo, foi em meados do século XVII que houve grandes progressos no estudo de técnicas estatísticas devido à necessidade dos líderes de estado da época quantificar e explicar fenômenos sociais e econômicos (ANDRIOTTI, 2003).

A estatística pode ser dividida em três importantes períodos: o primeiro, marcado pela necessidade do Estado de organizar e registrar informações de forma sistemática, ocorreu desde a época dos senhores feudais até meados do século XVII; o segundo período, onde a estatística começou a ser vista como um ramo da ciência de forma autônoma foi quando a Universidade de Iena inaugurou o primeiro curso de Estatística, em 1708. Foi também nesse período que Schmeitzel propôs o uso da palavra Estatística; o terceiro período é o período contemporâneo da estatística, teve início a partir de 1853 com o primeiro Congresso de Estatística e se mantem até os dias de hoje (CASTRO, 1970).

A estatística começou a se ramificar na medida que novos problemas eram levantados. Houve um grande interesse da parte de algumas pessoas em desenvolver teorias que explicassem os jogos de azar e com isso surgiu o que se conhece por teoria das probabilidades. Leonardo Pisano, conhecido como Fibonacci, viveu do século XII até meados do século XIII e foi o primeiro, que se tem registros, a estudar jogos de azar (VIALI, 2008). Blaise Pascal (1623-1662) e Pièrre de Fermat (1601-1665) trocaram cartas sobre problemas levantados por Pascal, como por exemplo qual seria o momento ideal para se interromper um jogo e a melhor maneira de fazer a divisão do dinheiro apostado (ANDRIOTTI, 2003).

Para Viali (2008), Laplace (1749 - 1827) publicou no início do século XIX uma clássica obra, considerada fundamental para sua época, intitulada Théorie Analytique des Probabilités, e abordou temas como as funções geratrizes, a regra de Thomas Bayes (1702 - 1761) sobre probabilidade inversa, probabilidade de eventos compostos, a teoria dos mínimos quadrados, e em edições posteriores também destacou a probabilidade de erros de observações, na determinação das massas dos maiores planetas, Júpiter e Saturno,

e também problemas de Geodésia.

Segundo Boyer e Merzbach (2012), Laplace foi um dos maiores contribuintes para o avanço da teoria das probabilidades. Mas para Viali (2008), Jacques Bernoulli teve grande influência ao dar início nos estudos sobre a Lei dos Grandes Números da qual se derivou, segundo Magalhães e Lima (2008), o Teorema do Limite Central em que mostra que as tomadas de valores aleatórios tendem a se distribuir igualmente em torno de um valor central. Esse Teorema do Limite Central levou ao que se conhece, hoje, por distribuição normal de probabilidade ou distribuição gaussiana descrita por uma equação matemática que leva o nome de equação de Laplace-Gauss (VUOLO, 1996).

A estatística contemporânea visa, através de diferentes ferramentas, estabelecer uma relação de causa e efeito de diversos fenômenos buscando prevê-los dentro de um específico intervalo de incerteza (CASTRO, 1970). Estas ferramentas foram e são desenvolvidas de acordo com a necessidade da análise em questão. Assim, de acordo com Hair et al. (2009), as análises de problemas que envolvem apenas uma variável são denominadas univariadas e problemas que possuem duas variáveis, dependentes e/ou independentes são classificados como bivariados e possuem como algumas ferramentas as análises de distribuições de probabilidade, correlação e regressão linear simples, e outras mais.

Já para casos em que se tem múltiplas variáveis dependentes e independentes as análises são classificadas como multivariadas, em outras palavras, esta denominação se dá para todas as ferramentas estatísticas que abrangem mais de duas variáveis independentes que possam explicar uma ou mais variável dependente (HAIR et al., 2009). Alguns métodos de análise univariada e bivariadas tiveram seus conceitos generalizados para análises multivariadas, como por exemplo a análise de distribuição de uma única variável, regressão linear, análise de correlação entre outros. Assim, algumas ferramentas multivariadas são as análises de regressão múltipla, correlação canônica, componentes principais, discriminante. (FERREIRA, 2008).

Segundo Ferreira (2008), um conjunto de dados apresenta uma característica intrínseca a ele denominada variabilidade e é devido a esta característica que se torna necessário a utilização de ferramentas estatísticas para melhor interpretação dos resultados. Esta variabilidade, além do desvio estatístico, deve-se, também, ao fato de que é inevitável realizar uma medida isenta de erro, ou seja, com um valor absoluto sem nenhuma incerteza experimental (VUOLO, 1996).

Assim, para Taylor (2012), o conhecimento da incerteza experimental é imprescindível para uma boa análise e interpretação dos resultados, pois dados que apresentam menores incertezas oferecem maior credibilidade e, portanto, apresentam um maior peso, comparado a dados com maiores incertezas, para a análise estatística que estiver sendo aplicada. Isso quer dizer que um conjunto de dados, cujas incertezas experimentais são desconhecidas, pode ser completamente inutilizado devido à falta de confiança em seus valores.

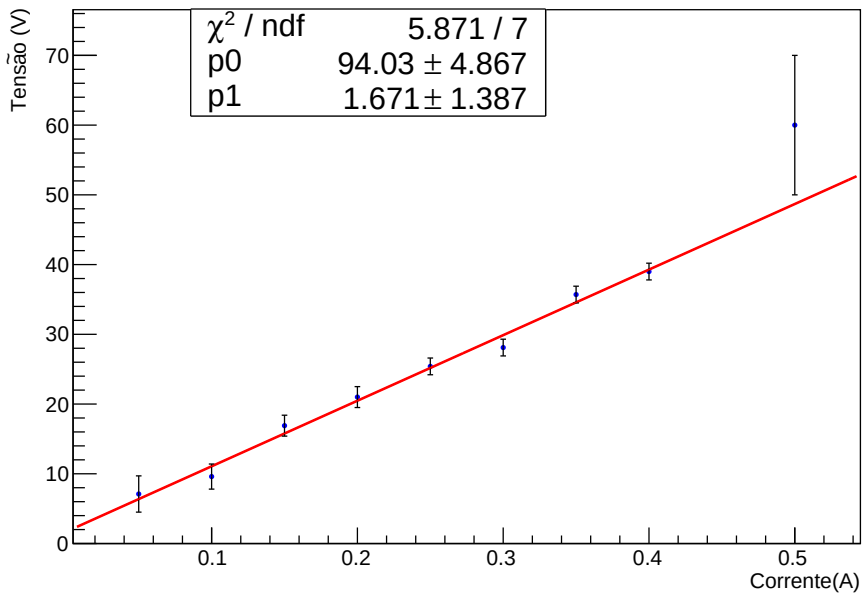
Para isso, a ferramenta estatística em uso deve ser capaz de assimilar esta incerteza experimental. Assim, devido aos pontos citados, este trabalho tem como objetivo inserir os valores de incertezas experimentais, de um conjunto de dados, nas análises Discriminante Linear de Fisher, de Componentes Principais e de Correlação Canônica, e também avaliar o impacto desta modificação nos resultados finais destas técnicas de estatística multivariada. Para isso, cada modelo será reformulado de modo que cada erro tenha um peso nos resultados finais de cada análise.

Quando se realiza uma análise estatística de um determinado conjunto de dados, levar em consideração as incertezas experimentais de cada elemento do conjunto se torna fundamental para a interpretação e análise dos resultados,pois os erros intrínsecos à tomada de dados exercem um peso na análise de forma que dados com incertezas pequenas são mais importantes que aqueles com incertezas maiores.

Isso pode ser observado, por exemplo, em uma análise de regressão linear simples em que se adota a distribuição normal de probabilidade para as incertezas experimentais. De uma maneira mais grosseira, o peso que cada dado exerce sobre o ajuste é inversamente proporcional ao quadrado da sua incerteza de modo que quanto maior o erro menos influência o ponto terá sobre o ajuste dos parâmetros da reta.

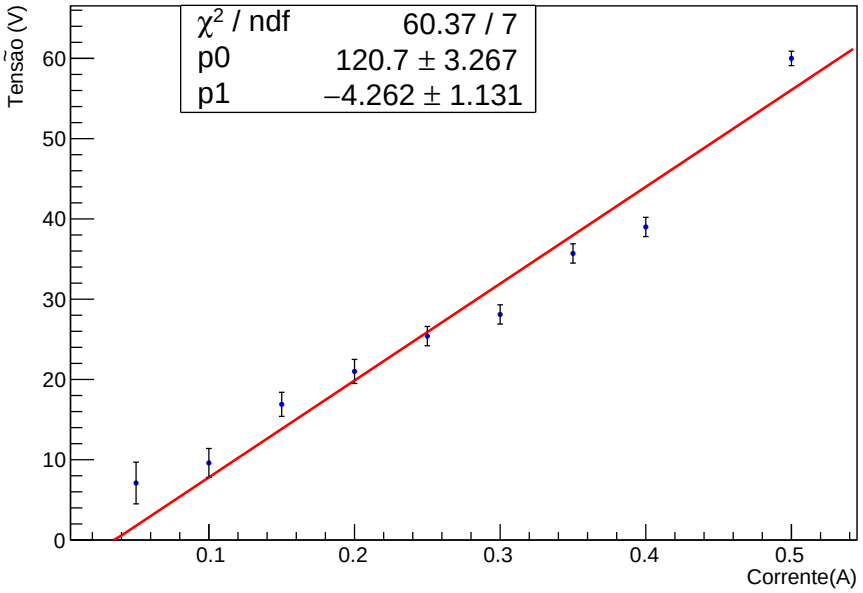
As duas imagens a seguir exemplificam bem esta situação. Como exemplo, assumiu-se valores fictícios para uma tomada de dados de tensão e corrente elétrica, com as incertezas da corrente já rebatidas para tensão.

Figura 1 – Gráfico de tensão por corrente, onde $p0$ e $p1$ representam o coeficiente angular e linear, respectivamente, da reta ajustada.



Fonte: Adaptado de Vuolo (1996).

Figura 2 – Gráfico de tensão por corrente, onde $p0$ e $p1$ representam o coeficiente angular e linear, respectivamente, da reta ajustada.



Fonte: Adaptado de Vuolo (1996).

Como pode ser observado no primeiro caso, Figura 1, o último ponto apresentou uma incerteza muito maior que os outros dados e isso fez com que os parâmetros fossem ajustados de modo com que a reta ficasse mais rente aos pontos com barras de erro menores. Já para o segundo caso, Figura 2, onde o último ponto tem uma incerteza menor, os parâmetros são tais que a reta se afasta de alguns pontos e se aproxima do último devido à sua maior importância para o ajuste.

Assim, essa grande importância que as incertezas experimentais exercem tanto no resultado da análise quanto em sua qualidade, justifica este trabalho de buscar inserir os valores de erros experimentais em modelos de análises estatísticas multivariadas. Portanto, no Capítulo 2 serão abordados alguns conceitos básicos de estatística multivariada que irão fundamentar o desenvolvimento. Nos Capítulos 3, 4 e 5 os conceitos clássicos das análises de Componentes Principais, Discriminante Linear de Fisher e de Correlações Canônicas irão ser discutidos, respectivamente. Já no Capítulo 6 será demonstrado e discutido um novo modelo desenvolvido neste trabalho para cada análise e por fim, no Capítulo 7 será realizada uma conclusão evidenciando as peculiaridades dos modelos desenvolvidos.

2 CONCEITOS DE ESTATÍSTICA MULTIVARIADA

A utilização de técnicas de análise multivariada de dados cresceu tanto que se tornou complicado mensurar as diferentes formas de aplicação para seus métodos (JOHNSON; WICHERN, 2007). Para Han, Pan e Yang (2016), o rápido desenvolvimento tecnológico demanda análises e previsões estatísticas em dados multidimensionais que estãos presentes não só na Física mas nos mais diferentes ramos da Ciência, como Egenharia, desenvolvimento de softwares, processamento de imagens dentre outros. Já para Aktekin, Polson e Soyer (2017), modelos de análise multivariada de dados têm importantes aplicações no desenvolvimento de ferramentes para *websites* como o número de *clicks*, contagem de acessos em múltiplas páginas da *web* e muitos outros.

Pode-se perceber, dado os exemplos citados acima, que há um número ilimitado de aplicações para ferramentes de estatística multivariada, e também existe a necessidade de melhoria dos modelos já existentes além da possibilidade de desenvolvimento de novos modelos. Como temos como objetivo reformular alguns modelos já existentes, é importante introduzir alguns conceitos e ferramentas matemáticas necessárias para compreensão e desenvolvimento deste trabalho. Por isso, neste capítulo, serão demonstradas algumas ferramentas imprescindíveis para o cumprimento deste estudo.

2.1 Amostras Aleatórias Multidimensionais

Quando se realiza um estudo científico, é imprescindível que exista uma etapa em que os conceitos teóricos sejam confrontados com resultados práticos e, para isso, é necessário que se realize experimentos e coleta de dados. Estes dados são conhecidos como observações de uma dada variável aleatória e quando há apenas uma, todo o estudo estatístico se resume ao caso univariado. Já para uma situação em que existem multiplas variáveis aleatórias, o conhecimento de estatística multivariada deve ser empregado (FERREIRA, 2008).

Segundo Johnson e Wichern (2007), no caso de n observações aleatórias, cada uma pode ser representada por um vetor aleatório de dimensão p, dado por

$$\boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{pmatrix} \tag{2.1}$$

onde $j = 1, 2, \cdots, n$. Outra forma de representar essa variável é através do seu vetor transposto $\boldsymbol{x}_j = [x_{1j}, x_{2j}, \cdots, x_{pj}]^T$. Já para Härdle e Simar (2015), todo o conjunto de variáveis aleatórias, também conhecido como amostra aleatória, pode ser expresso de

forma matricial, onde se tem

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \tag{2.2}$$

e cada coluna da matriz $\boldsymbol{X}$ é um vetor aleatório de dimensão p. Estas observações podem ser consideradas aleatórias devido ao processo de coleta dos dados e são distribuídas conforme uma determinada função densidade de probabilidade com m parâmetros descrita por $f(x_j; \theta_{j1}, \cdots, \theta_{jm})$, onde a probabilidade total das amostras ocorrerem é dada pela distribuição conjunta das n observações aleatórias. Assim, a probabilidade conjunta pode ser dada por

$$F(\theta) = \prod_{j=1}^{n} f(\boldsymbol{x}_j; \theta_{j1}, \cdots, \theta_{jm}). \tag{2.3}$$

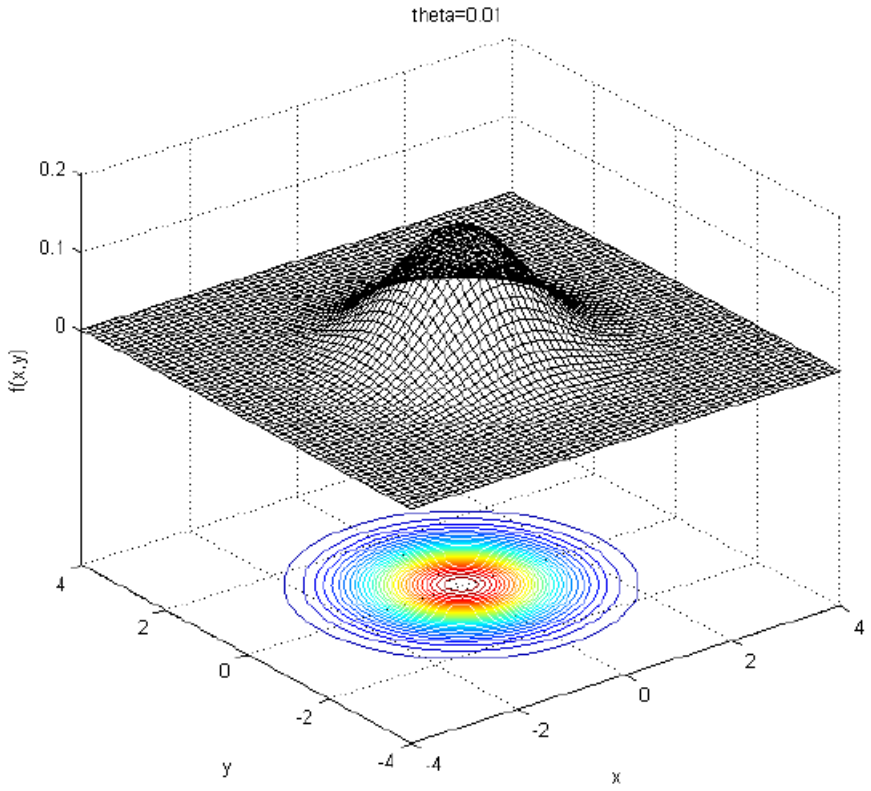
Se for assumido, previamente, que a função densidade de probabilidade, $f(x_j; \theta_{j1}, \cdots, \theta_{jm})$, é conhecida, como a distribuição normal, por exemplo, e que os dados são observados conforme essa distribuição, então essa probabilidade conjunta é denominada função de verossimilhança (FERREIRA, 2008).

2.2 Distribuição Normal Multivariada

A distribuição normal multivariada pode ser interpretada como uma densidade de probabilidade conjunta de múltiplas variáveis normalmente distribuidas, que dependa dos parâmetros valor médio e variância de cada uma delas e também das correlações entre si (MAJUMDAR; MAJUMDAR, 2016). A função densidade de probabilidade gaussiana é aplicada em teoria dos erros pois descreve o comportamento de erros experimentais (VUOLO, 1996). Para Haubold, Mathai e Thomas (2007) a distribuição normal (gaussiana) é uma família de distribuições contínuas de probabilidade e está presente em todas as áreas da estatística e teoria de probabilidade.

Tanto para Haubold, Mathai e Thomas (2007) quanto para Vuolo (1996) sua importância é fundamentada no teorema do limite central que, de maneira mais simples, mostra que se uma medida, x, é a soma de muitras outras medidas com diferentes distribuições, a distribuição de probabilidade para x tende à gaussiana à medida que o número de observações tende ao infinito. Deste modo, como qualquer coleta de dados apresenta diferentes fontes de erro, o conhecimento da distribuição normal de probabilidade se torna fundamental. Uma maneira possível de visualizá-la graficamente é através do caso particular em que a distribuição é bivariada, como mostra a Figura a seguir:

Figura 3 – Representação gráfica de uma função densidade de probabilidade normal bivariada.



Fonte: (MAHMOUDI; MAHMOODIAN, 2017).

De acordo com Balakrishnan e Lai (2009), a distribuição normal multivariada pode ser demonstrada a partir da distribuição normal univariada, quando as variáveis são independentes. Por exemplo, em um conjunto de p variáveis independentes pertencentes a um vetor aleatório $\boldsymbol{x} = [x_1, x_2, \cdots, x_p]^T$, cada variável possui uma função densidade de probabilidade dada por

$$f(x_i; \mu_i, \sigma_i) = \frac{1}{\sqrt{2\Pi\sigma_i^2}} \exp[-\frac{1}{2}\frac{(x_i - \mu_i)^2}{\sigma_i^2}] \tag{2.4}$$

onde μ_i e σ_i^2 são o valor esperado e a variância, respectivamente, para i-ésima variável. A densidade conjunta das variáveis presentes no vetor aleatório pode ser calculada como o produto de todas as probabilidades individuais, logo

$$f(x; \mu, \theta) = \prod_{i=1}^{p} f(x_i; \mu_i, \sigma_i) = \frac{1}{(2\Pi)^{\frac{p}{2}}}(\prod_{i=1}^{p}\sigma_i^2)^{-\frac{1}{2}} \exp[-\frac{1}{2}\sum_{i=1}^{p}\frac{(x_i - \mu_i)^2}{\sigma_i^2}]. \tag{2.5}$$

Essa expressão pode ser reescrita utilizando notações de vetores e matrizes, dada por

$$f(\boldsymbol{x}) = \frac{1}{(2\Pi)^{\frac{p}{2}}}|\boldsymbol{V}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})]. \tag{2.6}$$

em que $\boldsymbol{V} = diag(\sigma_{ii}^2)$ e $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_p]^T$. A matriz $\boldsymbol{V}$ representa um caso particular em que as variáveis são independentes entre si. Segundo Ambikasaran et al. (2016), para

um caso mais geral, onde as variáveis não são necessariamente independentes, pode-se utilizar, no lugar de $\boldsymbol{V}$, uma matriz de covariância dada por

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp}^2 \end{pmatrix}. \tag{2.7}$$

Uma vez que o termo $(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{V}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$ está na forma quadrática, e portanto seus elementos são não negativos, para que a inversa da matriz $\boldsymbol{\Sigma}$ exista, ela deve ser positiva definida. Assim, pode-se reescrever a equação 2.6 como

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{p}{2}}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp[-\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})]. \tag{2.8}$$

onde $\boldsymbol{x} = [x_1, x_2, \cdots, x_i, \cdots, x_p]^T$ é o vetor p-dimensional cujas componentes são as variáveis x_i, $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_i, \cdots, \mu_p]^T$ é o vetor em que suas componentes são os valores médios de cada variável e $\boldsymbol{\Sigma}$ é a matriz de covariância entre as variáveis (FERREIRA, 2008).

2.3 Estimadores de Máxima Verossimilhança do vetor médio e da matriz de covariância

Quando se tem um conjunto de dados finitos que respeitam uma determinada função densidade de probabilidade (fdp) onde pelo menos um dos parâmetros são desconhecidos, pode-se utilizar o método da máxima verossimilhança como uma técnica para estimá-los. A função de verossimilhança representa a probabilidade total de todos os valores da amostra ocorrerem (COWAN, 1998). Assim, para uma variável aleatória medida n vezes e pertencente a uma fdp $f(x;\theta_1,\cdots,\theta_m)$, a probabilidade de cada uma das medidas ocorrerem, separadamente, é $f(x_1;\theta_1,\cdots,\theta_m),\cdots,f(x_n;\theta_1,\cdots,\theta_m)$. Já a probabilidade total de todas as medidas ocorrerem segundo Cowan (1998) é

$$L(\theta) = \prod_{i=1}^{n} f(x;\theta_1,\cdots,\theta_i,\cdots,\theta_m) \tag{2.9}$$

onde $L(\theta)$ é denominada função de verossimilhança. Os parâmetros desejados serão aqueles que maximizam a função de verossimilhança e para obtê-los basta resolver as equações para cada um dos parâmetros, dadas por

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0 \; i = 1, 2, \cdots, m. \tag{2.10}$$

Como os parâmetros de qualquer fdp são puramente teóricos, e portanto inclui-se os parâmetros da distribuição normal multivariada, $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, uma maneira de estimá-los é aplicando as equações (2.9) e (2.10) em (2.8), portanto, tem-se que

$$L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{j=1}^{n} f(\boldsymbol{x}_j) \tag{2.11}$$

$$L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{j=1}^{n} \frac{1}{(2\Pi)^{\frac{p}{2}}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})] \tag{2.12}$$

$$L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp[-\frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})] \tag{2.13}$$

Para Ferreira (2008), pode-se tomar o logaritmo natural da função de verossimilhança, transformando-a em uma nova função denominada função suporte. Isto facilita os cálculos e não altera os resultados obtidos para os parâmetros desejados. Portanto, uma maneira para estimar o valor de $\boldsymbol{\mu}$ que maximiza (2.13), é calcular a função suporte aplicada à densidade de probabilidade gaussiana, derivá-la e igualá-la à zero, então determinar o valor de $\boldsymbol{\mu}$ que satisfaz a equação. Assim, a função suporte será

$$\ln[L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})] = \ln\{\frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp[-\frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})]\} \tag{2.14}$$

rearranjando obtem-se

$$\ln[L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})] = \ln(2\Pi)^{-\frac{np}{2}} + \ln|\boldsymbol{\Sigma}|^{-\frac{n}{2}} - \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}) \tag{2.15}$$

e derivando com relação a $\boldsymbol{\mu}$ e igualando a zero encontra-se

$$\frac{\partial \ln[L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})]}{\partial \boldsymbol{\mu}} = \sum_{j=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}) = 0. \tag{2.16}$$

Logo

$$\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}) = \mathbf{0} \tag{2.17}$$

$$\sum_{j=1}^{n}\boldsymbol{x}_j - \sum_{j=1}^{n}\boldsymbol{\mu} = \mathbf{0} \tag{2.18}$$

e o estimador de verossimilhança de $\boldsymbol{\mu}$, será

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{2.19}$$

que, segundo Ferreira (2008), corresponde com o valor da média amostral.

Para estimar a matriz de covariância, é mais adequado levar em consideração que o argumento da exponencial da equação (2.13) é um escalar, visto que os elementos da soma são da forma quadrática e, portanto, pode-se aplicar a seguinte propriedade: seja uma matriz simétrica e quadrada, $p \times p$, e um vetor coluna, $p \times 1$, tem-se que

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} = tr(\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}) = tr(\boldsymbol{A}\boldsymbol{v}\boldsymbol{v}^T) \tag{2.20}$$

$$tr(\boldsymbol{A}) = \sum_{i=1}^{p} \lambda_i \tag{2.21}$$

onde p e λ_i são a ordem e autovalor i da matriz $\boldsymbol{A}$, repectivamente (JOHNSON; WICHERN, 2007).

Assim, reescrevendo a equação (2.13) tem-se

$$L(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\{-\frac{1}{2} \sum_{j=1}^{n} tr[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^T]\} \tag{2.22}$$

como a somatória do traço de matrizes é o traço da somatória de matrizes, então

$$L(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\{-\frac{1}{2} tr[\boldsymbol{\Sigma}^{-1} \sum_{j=1}^{n} (\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^T]\}. \tag{2.23}$$

Somando e subtraindo dentro de cada vetor da soma o elemento $\bar{\boldsymbol{x}}$, obtém-se

$$L(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\{-\frac{1}{2} tr[\boldsymbol{\Sigma}^{-1} \sum_{j=1}^{n} (\boldsymbol{x}_j + \bar{\boldsymbol{x}} - \bar{\boldsymbol{x}} - \boldsymbol{\mu})(\boldsymbol{x}_j + \bar{\boldsymbol{x}} - \bar{\boldsymbol{x}} - \boldsymbol{\mu})^T]\}. \tag{2.24}$$

Assim, a equação anterior pode ser escrita da forma

$$L(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\{-\frac{1}{2} tr[\boldsymbol{\Sigma}^{-1} (\sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T + \sum_{j=1}^{n} (\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T)]\} \tag{2.25}$$

$$L(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\{-\frac{1}{2} tr[\boldsymbol{\Sigma}^{-1} (\sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T)]\}. \tag{2.26}$$

Lembrando que o valor esperado do vetor aleatório, $\boldsymbol{x}$, que maximiza a equação (2.13) é $\boldsymbol{\mu} = \boldsymbol{x}$, dado por (2.19), e portanto a equação acima fica

$$L(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\Pi)^{\frac{np}{2}}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\{-\frac{1}{2} tr[\boldsymbol{\Sigma}^{-1} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T]\}. \tag{2.27}$$

Já a matriz, $\boldsymbol{\Sigma}$, que maxima a função $L(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ pode ser determinada pela seguinte propriedade: seja uma matriz quadrada $p \times p$ e simétrica positiva definida, $\boldsymbol{B}$, e um escalar positivo b, então

$$\frac{1}{|\boldsymbol{\Sigma}|^b} \exp\left[-\frac{tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{B})}{2}\right] \leq \frac{1}{|\boldsymbol{B}|^b} (2b)^{-bp} \tag{2.28}$$

que apresentará uma igualdade quando $\boldsymbol{\Sigma} = \frac{1}{2b}\boldsymbol{B}$.

Portanto, utilizando a propriedade descrita por Johnson e Wichern (2007), considerando $\boldsymbol{B} = \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T$ e $b = (\frac{n}{2})$, tem-se que a matriz $\boldsymbol{\Sigma}$ que maximiza (2.13) é

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T, \tag{2.29}$$

o que, segundo Ferreira (2008) é o mesmo que

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}(n-1)\boldsymbol{S} \tag{2.30}$$

onde

$$\boldsymbol{S} = \frac{1}{(n-1)}\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T. \tag{2.31}$$

2.4 Teste de Normalidade Multivariada

Ao trabalhar com estatística multivariada para análise de dados, depara-se com métodos que só podem ser aplicados caso o conjunto de observações respeite uma distribuição normal multivariada de probabilidade. Portanto, antes de se realizar alguns métodos, é estritamente necessário saber se a probabilidade de ocorrência dos dados a serem analisados respeitam a função densidade gaussiana multivariada (THULIN, 2014). Neste tópico serão abordados dois métodos utilizados para testar normalidade, o cálculo de assimetria e de curtose da distribuição de frequência dos dados. Contudo, é importante ressaltar que há muitos outros testes capaze de realizar esta averiguação, conforme Henze e Koch (2016) tem-se, por exemplo, os testes de Anderson-Darling, Shapiro-Wilk, Epps-Pulley, dentre outros.

Ferreira (2008) mostra que pode-se analisar a normalidade de um conjunto de dados realizando um teste de hipótese, em conjunto, do coeficiente de assimetria, β_{1p}, e curtose, β_{2p}. Assim, seja a hipótese nula $H_0 : \beta_{1p} = 0; \beta_{2p} = p(p+2)$, o teste consiste em determinar o valor

$$k = n\frac{\hat{\beta}_{1p}}{6} + \frac{n}{8p(p+2)}\left[\hat{\beta}_{2p} - \frac{p(p+2)(n-1)}{(n+2)}\right]^2 \tag{2.32}$$

onde $\hat{\beta}_{1p}$ e $\hat{\beta}_{2p}$ são os estimadores de β_{1p} e β_{2p}, respectivamente, p é a dimensão do vetor de observação e n a quantidade desses vetores. deste modo, a hipótese nula deve ser rejeitada se k for maior que o valor obtido pela distribuição χ^2_ν com $\nu = 1 + p(p+1)(p+2)/6$ graus de liberdade para um dado valor de significância α. Ainda para Ferreira (2008), os estimadores dos coeficientes de assimetria e curtose podem ser obtidos calculando

$$\boldsymbol{G} = (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\cdot\mathbf{1}^T)\boldsymbol{X}\boldsymbol{S}_n^{-1}\boldsymbol{X}^T(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\cdot\mathbf{1}^T) \tag{2.33}$$

em que $\boldsymbol{I}$ é a matriz identidade de ordem n e $\mathbf{1} = [1_{(1)}, 1_{(2)}, \cdots, 1_{(n)}]^T$ é um vetor de dimensão $(n \times 1)$ onde todos os elementos são iguais a 1. Uma vez construida a matriz $G(n \times n)$, os estimadores dos coeficientes de assimetria e curtose serão

$$\hat{\beta}_{1p} = \sum_{j=1}^{n}\sum_{i=1}^{n}\frac{(G_{ij})^3}{n^2} \tag{2.34}$$

e

$$\hat{\beta}_{2p} = \sum_{j=1}^{n}\sum_{i=1}^{n}\frac{(G_{ij})^2}{n}. \tag{2.35}$$

2.5 Propagação de incertezas

Foi demonstrado no tópico anterior como estimar o valor esperado e a matriz de covariância de uma vetor aleatório $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^T$ cuja função densidade de probabilidade é dada pela função gaussiana. Neste tópico será abordado o problema de estimar a incerteza de uma variável dependente de x. Assim, a variável dependente, y, e a variável idependente, x, podem ser representadas através de uma função dada por $y(x)$.

Na prática não se conhece por completo a função densidade de probabilidade da variável x, mas se conhece os estimadores de seu valor esperado e variância e isso já é suficiente para estimar a incerteza de y, assim, pode-se estimar o valor esperado de $y(x)$ expandindo essa função em série de Taylor, até primeira ordem, em torno do valor médio de x denotado por $\boldsymbol{\mu}$ (COWAN, 1998).

Segundo Vuolo (1996), essa condição de se expandir somente até a primeira ordem só é considerada uma boa aproximação se o desvio $\boldsymbol{d}_{\boldsymbol{x}_i} = (\boldsymbol{x}_i - \boldsymbol{\mu})$ for da ordem de grandeza do desvio padrão σ_x, pois assim o termo quadrático de $\boldsymbol{d}_{\boldsymbol{x}_i} = (\boldsymbol{x}_i - \boldsymbol{\mu})$ pode ser considerado desprezível. Portanto, temos que

$$y(\boldsymbol{x}) = y(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_i - \mu_i) + \frac{1}{2}\sum_{i=1}^{n} \left[\frac{\partial^2 y}{\partial x_i^2}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_i - \mu_i)^2 + \cdots, \tag{2.36}$$

como dito

$$\frac{1}{2}\sum_{i=1}^{n} \left[\frac{\partial^2 y}{\partial x_i^2}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_i - \mu_i)^2 \approx 0 \tag{2.37}$$

quando $(x_i - \mu_i) \approx \sigma_{x_i}$. Logo, para Cowan (1998) todos os termos a partir da segunda ordem serão desprezíveis e uma aproximação para o valor de $y(x)$ é dada por

$$y(\boldsymbol{x}) \approx y(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_i - \mu_i) \tag{2.38}$$

e usando a propriedade em que a esperança da soma é a soma das esperanças, então seu valor médio será

$$\begin{aligned} E[y(\boldsymbol{x})] &\approx E[y(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_i - \mu_i)] \\ &= E[y(\boldsymbol{\mu})] + E[\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_i - \mu_i)] \end{aligned} \tag{2.39}$$

$$E[y(\boldsymbol{x})] \approx y(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} E[(x_i - \mu_i)] \tag{2.40}$$

Portanto, uma aproximação para o valor esperado de $y(\boldsymbol{x})$ será

$$E[y(\boldsymbol{x})] \approx y(\boldsymbol{\mu}) \tag{2.41}$$

desde que

$$E[x_i - \mu_i] = 0. \tag{2.42}$$

Como a variância de uma variável qualquer, A, é calculada por $Var[A] = E[A^2] - E^2[A]$, então a variância de $y(\boldsymbol{x})$ é

$$Var[y(\boldsymbol{x})] = E[y^2(\boldsymbol{x})] - E^2[y(\boldsymbol{x})] = \sigma_y^2. \tag{2.43}$$

Assim, expandindo $y^2(\boldsymbol{x})$ em série de Taylor até a primeira ordem e calculando a esperança do resultado, tem-se

$$\begin{aligned} E[y^2(\boldsymbol{x})] \approx y^2(\boldsymbol{\mu}) + 2y(\boldsymbol{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} E[(x_i - \mu_i)] + \\ + E\left[\left(\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_i - \mu_i)\right)\left(\sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} (x_j - \mu_j)\right)\right] \end{aligned} \tag{2.44}$$

$$E[y^2(\boldsymbol{x})] \approx y^2(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} E[(x_i - \mu_i)(x_j - \mu_j)] \tag{2.45}$$

$$E[y^2(\boldsymbol{x})] \approx y^2(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} V_{ij} \tag{2.46}$$

onde V_{ij} é a componente ij da matriz de covariância de $\boldsymbol{x}$. Logo

$$\sigma_y^2 = y^2(\boldsymbol{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} V_{ij} - y^2(\boldsymbol{\mu}) \tag{2.47}$$

$$\sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} V_{ij} \tag{2.48}$$

Caso haja mais de uma varável dependente, $y_1(x), y_2(x), \cdots, y_m(x)$, a matriz de covariância é obtida de forma análoga, e sua equação é dada por

$$Cov[y_k, y_l] = \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j}\right]_{\boldsymbol{x}=\boldsymbol{\mu}} V_{ij} \tag{2.49}$$

3 ANÁLISE DE COMPONENTES PRINCIPAIS

Quando se deseja conhecer as causas de uma medida ou fenômeno, o pesquisador pode, a princípio, se deparar com uma quantidade muito grande de variáveis capazes de explicar certos comportamentos observados. Isso torna o problema ainda maior, uma vez que quanto maior o número de variáveis, mais complicado se torna identificar quais delas possuem, de fato, mais peso para uma determinada análise. Nestes casos, a análise de componentes principais, ou *Principal Components Analysis* (PCA) pode exercer um papel fundamental, pois seus principais objetivos são a redução de dimensionalidade do problema e identificar padrões que não sejam triviais em um conjunto de dados (ENSOR et al., 2017).

Conforme abordado por Johnson e Wichern (2007), a PCA consiste em determinar novas variáveis (componentes principais) através de combinações lineares das variáveis originais do problema, utilizando sua matriz de covariância, de maneira que estas novas variáveis possam explicar de um modo mais sucinto a variabilidade do sistema. De acordo com Mingoti (2005), após a obtenção das componentes principais, calcula-se um valor numérico, também denominado *score*, para cada observação do conjunto de dados utilizado. Assim, pode-se empregar esta técnica para construir variáveis que possam ser analizadas através de outras ferramentas estatísticas como análise de regressão, variância e até mesmo análise discriminante.

Dentre as aplicações, Mursula e Holappa (2017) utilizaram a análise de componentes principais para estudar a atividade geomagnética a partir de dados obtidos por 26 e 40 estações magnéticas no período de 1966 à 2015 e 1980 à 2015, respectivamente. Já Sirunyan et al. (2017) utilizaram, pela primeira vez, a PCA para separar modos ortogonais da matriz de correlação de duas partículas em uma colisão de íons pesados. Em outro trabalho, ao estudarem, através de análise de imagens, os domínios ferroelétrico em amostras de $LiNbO_3$ polidas periodicamente, Esfahani, Liu e Li (2017) destacaram que com a utilização da PCA foram capazes de reduzir os níveis de ruídos do experimento e analizar, de maneira quantitativa, dados que antes não eram possíveis de ser analisados.

Como pôde ser observado, a análise de componentes principais é uma ferramenta bastante poderosa para interpretação e análise de dados multivariados. Sua aplicação, assim como qualquer outra ferramenta de estatística multivariada, se extende, além da Física, as mais diversas áreas como ciências econômicas, biológicas, da saúde e suas ramificações. Neste capítulo será abordada toda a estrutura matemática usual da PCA bem como a interpretação dos possíveis resultados.

3.1 Análise de Componentes Principais por matriz de Covariância

Como já mencionado, a análise de componentes principais tem como objetivo determinar novas variáveis que sejam combinações lineares das variáveis originais. Contudo, a análise terá um maior impacto se as observações originais forem correlacionadas. Uma consequência da análise é que as combinações lineares (componentes principais), geradas, não são correlacionadas entre si. Se o problema inicial contempla p variáveis, então a análise gerará p componentes principais, entretanto, o objetivo de quem a aplica é escolher $k \leq p$ componentes que expliquem satisfatoriamente o comportamento e variabilidade das variáveis originais (MINGOTI, 2005).

A princípio, as observações mensuradas experimentalmente não precisam respeitar uma distribuição de probabilidade específica, porém, neste trabalho será admitido que os dados possuem uma função densidade de probabilidade gaussiana. Assim, seja um vetor aleatório $\boldsymbol{x}_i = [x_1, x_2, \cdots, x_p]^T$, de um conjunto de dados com n medidas onde $i = 1, 2, \cdots, n$, com média $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_p]^T$ e matriz de covariância $\boldsymbol{\Sigma}$ com dimensões $(p \times p)$, então a j - ésima componente principal será um escalar dado por

$$Y_j = \boldsymbol{e}_j \cdot \boldsymbol{x}, \tag{3.1}$$

onde $\boldsymbol{e}_j$ é o j - ésimo vetor desconhecido que se deseja determinar sob um critério de maximização de modo que $j = 1, 2, \cdots, p$ (FERREIRA, 2008).

Para Johnson e Wichern (2007), o que a PCA faz, na realidade, é rotacionar os sistema de eixos coordenados de modo que os novos eixos estejam na direção de maior variabilidade dos dados. Portanto, a análise consiste em determinar os p vetores $\boldsymbol{e}_j$ que maximizam a variância das componentes principais. Seja a variância da j-ésima componente principal, Y_j, dada por

$$Var(Y_j) = Var(\boldsymbol{e}_j \boldsymbol{x}) = \boldsymbol{e}_j^T Var(\boldsymbol{x}) \boldsymbol{e}_j, \tag{3.2}$$

então

$$Var(Y_j) = \boldsymbol{e}_j^T \boldsymbol{\Sigma} \boldsymbol{e}_j \tag{3.3}$$

e a covariância entre Y_j e Y_k, com $j \neq k$ igual a

$$Cov(Y_j, Y_k) = Cov(\boldsymbol{e}_j^T \boldsymbol{x}, \boldsymbol{e}_k^T \boldsymbol{x}) = \boldsymbol{e}_j^T Var(\boldsymbol{x}) \boldsymbol{e}_k \tag{3.4}$$

ou

$$Cov(Y_j, Y_k) = \boldsymbol{e}_j^T \boldsymbol{\Sigma} \boldsymbol{e}_k. \tag{3.5}$$

Tanto para Johnson e Wichern (2007) quanto para Mingoti (2005) a maximização da variância deve ocorrer sob as restrições tais que $\boldsymbol{e}_j^T \boldsymbol{e}_j = 1$ e $\boldsymbol{e}_j^T \boldsymbol{e}_k = 0$. Assim o máximo de (3.3) será

$$\lambda_j = \max_{\boldsymbol{e}_j} \boldsymbol{e}_j^T \boldsymbol{\Sigma} \boldsymbol{e}_j \tag{3.6}$$

que dado a restrição é o mesmo que

$$\lambda_j = \max_{\boldsymbol{e}_j} \frac{\boldsymbol{e}_j^T \boldsymbol{\Sigma} \boldsymbol{e}_j}{\boldsymbol{e}_j^T \boldsymbol{e}_j}. \tag{3.7}$$

Segundo Härdle e Simar (2015) a maior variância que uma componente principal pode ter corresponde ao maior autovalor, λ_j, da matriz $\boldsymbol{\Sigma}$ de modo que este máximo é alcançado com o autovetor, $\boldsymbol{e}_j$, correspondente a este autovalor. De forma análoga, o autovetor gerado pelo menor autovalor de $\boldsymbol{\Sigma}$ resultará na menor variância possível de uma componente principal.

Assim sendo, a componente principal de maior relevância (PC1) será aquela estabelecida pelo autovetor correspondente ao maior autovalor da matriz $\boldsymbol{\Sigma}$, já a PC2 corresponderá ao segundo maior autovalor, e assim por diante até a última componente principal (PC) (AL-SAYED, 2015; GRATIER et al., 2017). Cada autovetor, $\boldsymbol{e}_j$, da matriz $\boldsymbol{\Sigma}$ será um eixo do novo sistema de coordenadas e considerando que $\boldsymbol{e}_j^T \boldsymbol{e}_k = 0$ então os eixos são ortognais e consequentemente as componentes principais são independentes entre si. Conforme Hamill et al. (2018) isso pode ser demonstrado, uma vez que a determinação dos p diferentes Y_j consiste, no fundo, em solucionar o seguinte sistema de equações,

$$(\boldsymbol{\Sigma} - \lambda_j \boldsymbol{I})\boldsymbol{e}_j = \boldsymbol{0},$$

então

$$\boldsymbol{\Sigma} \boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j,$$

e, ao substituir a igualdade acima na equação (3.3), tem-se

$$Var(Y_j) = \boldsymbol{e}_j^T \lambda_j \boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j^T \boldsymbol{e}_j = \lambda_j. \tag{3.8}$$

Analogamente,

$$Cov(Y_j) = \boldsymbol{e}_j^T \lambda_k \boldsymbol{e}_k = \lambda_k \boldsymbol{e}_j^T \boldsymbol{e}_k = 0. \tag{3.9}$$

Uma característica desta análise é que os coeficientes das combinações lineares, ou seja, as componentes de cada autovetor da matriz $\boldsymbol{\Sigma}$, determinam a importância de cada variável. Assim, a variável mais importante para cada PC será sempre aquela que possuir o maior coeficiente (MINGOTI, 2005). Outra propriedade importante da PCA, é que segundo o teorema da decomposição espectral, a matriz de covariância pode ser escrita como $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^T$, em que $\boldsymbol{V}$ é uma matriz diagonal cujos elementos da diagonal principal são os autovalores da matriz $\boldsymbol{\Sigma}$ e $\boldsymbol{U}$ é a matriz com os correspondentes autovetores em cada coluna. Portanto,

$$tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^T) = tr(\boldsymbol{V}\boldsymbol{U}^T\boldsymbol{U}) = tr(\boldsymbol{V}\boldsymbol{I}) \tag{3.10}$$

e

$$tr(\boldsymbol{\Sigma}) = \sum_{j=1}^{p} \lambda_j. \tag{3.11}$$

Porém, como o traço de uma matriz é a soma dos elementos da sua diagonal principal, então pode-se afirmar que a soma das variâncias das variâveis originais é igual a soma das variâncias das componentes principais, ou seja, a variabilidade total das medidas é mantida (HÄRDLE; SIMAR, 2015; MINGOTI, 2005).

Na prática, não se conhece os valores médios dos dados amostrais. Por isso, é necessário estimar o vetor de média, $\boldsymbol{\mu}$, e a matriz de covariância, $\boldsymbol{\Sigma}$, através das equações (2.19) e (2.31), respectivamente. Assim, na ausência de incertezas experimentais, para um número n de vetores aleatórios coletados experimentalmente, seu vetor médio e matriz de covariância serão

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{3.12}$$

e

$$\boldsymbol{S} = \frac{1}{(n-1)}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T. \tag{3.13}$$

Desta forma, o j-ésimo autovalor e seu correspondente autovetor, da matriz $\boldsymbol{S}$, serão $\hat{\lambda}_j$ e $\hat{\boldsymbol{e}}_j$, respectivamente (JOHNSON; WICHERN, 2007).

3.2 Análise de Componentes Principais por Matriz de Correlação

Quando se utiliza a matriz de covariância dos dados originais para realizar a análise de componentes principais, as PCs sofrem bastante influência das variáveis cujas variâncias são muito discrepantes das restantes. Em outras palavras, em cada combinação linear o maior coeficiente será, geralmente, daquela variável que possuir a maior variância. Uma das causas dessa discrepância pode ser a diferença de escala das variáveis mensuradas (MINGOTI, 2005).

Esta situação pode ser resolvida ou suavizada se os autovalores e autovetores forem extraídos da matriz de correlação, $\boldsymbol{\rho}$, dos dados originais. Isso pode ser mostrado ao realizar uma transformação nas variáveis originais, de modo que

$$\boldsymbol{z}_i = \boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{3.14}$$

onde $\boldsymbol{\Gamma}^{-\frac{1}{2}}$ é uma matriz diagonal cuja diagonal principal é a diagonal principal da matriz $\boldsymbol{\Sigma}$, e $\boldsymbol{z}_i$ é denominada variável padronizada. Assim, as componentes principais serão combinações lineares do tipo

$$Y_j = \boldsymbol{e}_j^T\boldsymbol{z}, \tag{3.15}$$

que para manter a simplicidade será utilizado neste caso a mesma notação de autovalores λ, autovetores $\boldsymbol{e}$, e componentes principais Y, daquela vista no tópico anterior. Entretanto, é importante ressaltar que seus valores numéricos são diferentes (FERREIRA, 2008).

Johnson e Wichern (2007) discutem que o processo de obtenção das componentes principais através das variáveis padronizadas é o mesmo, pois as restrições $\boldsymbol{e}_j^T\boldsymbol{e}_j = 1$

e $\boldsymbol{e}_j^T \boldsymbol{e}_k = 0$ continuam as mesmas e o conceito de maximização da variância também, porém, agora as PCs não são afetadas pela mudança de escala das variáveis originais. Assim, tem-se que

$$Var(Y_j) = Var(\boldsymbol{e}_j^T \boldsymbol{z}) = \boldsymbol{e}_j^T Var(\boldsymbol{z}) \boldsymbol{e}_j \tag{3.16}$$

ou

$$Var(Y_j) = \boldsymbol{e}_j^T \boldsymbol{\rho} \boldsymbol{e}_j \tag{3.17}$$

já que a matriz de covariância das variáveis padronizadas é igual a matriz de correlação das variáveis originais. Portanto, encontram-se os mesmos resultados obtidos no tópico anterior, visto que

$$\boldsymbol{\rho} \boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j. \tag{3.18}$$

Logo, fica demonstrado que as PC das variáveis padronizadas são os autovalores e autovetores da matriz de correlação dos dados originais. Com isso, a variabilidade total pode ser descrita como

$$tr(\boldsymbol{\rho}) = \sum_{j=1}^{p} 1 = \sum_{j=1}^{p} \lambda_j = p. \tag{3.19}$$

Na prática, este método pode ser aplicado estimando-se as variáveis padronizadas,

$$\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{\Gamma}}^{-\frac{1}{2}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{3.20}$$

e, a partir delas, sua matriz de covariância dada pela equação (2.31). Isso é equivalente à calcular diretamente a matriz de correlação dos dados originais,

$$\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\Gamma}}^{-\frac{1}{2}} \boldsymbol{S} \hat{\boldsymbol{\Gamma}}^{-\frac{1}{2}} \tag{3.21}$$

onde $\hat{\boldsymbol{\Gamma}}^{-\frac{1}{2}} = diag(\frac{1}{\sqrt{S_{ii}}})$. Assim, a j-ésima componente principal pode ser determinada a partir dos j-ésimo autovalor $\hat{\lambda}_j$, e autovetor $\hat{\boldsymbol{e}}_j$, da matriz $\hat{\boldsymbol{\rho}}$ (HÄRDLE; SIMAR, 2015).

3.3 Escolha das Componentes Principais

Um dos principais objetivos da análise de componentes principais é reduzir o número de variáveis com as quais o pesquisador irá trabalhar, como já citado anteriormente. Como o número de componentes principais geradas pela PCA é igual ao número de variáveis originais do problema, se torna necessário escolher $k \leq p$ PCs que possam explicar satisfatoriamente a variabilidade do problema. Para isso, é possível calcular a variabilidade individual de cada componente principal com relação a variabilidade total dos dados originais. Dado que a variância da componente Y_j é o autovalor λ_j da matriz $\boldsymbol{\Sigma}$ ou $\boldsymbol{\rho}$, Ferreira (2008) aborda que seu peso, γ_j, será

$$\gamma_j = \frac{\lambda_j}{tr(\boldsymbol{\Sigma})} = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda j}. \tag{3.22}$$

e que o peso total das k componentes escolhidas pode ser calculado por

$$\gamma_T = \sum_{j}^{k} \gamma_j. \tag{3.23}$$

Assim, através da equação (3.22) é possível identificar qual componente principal explica a maior parte da variabilidade dos dados originais. Conforme discutido por Mingoti (2005), o que se busca geralmente com a PCA, é reduzir o número de variáveis p para $k < p$, sem que se perca muita informação da variabilidade das variáveis originais. Assim, o pesquisador pode analisar e tirar conclusões a partir de um número k muito menor de variáveis comparado à p, o que torna mais simples o seu trabalho. Não há, entretanto, uma regra que determina o número k de componentes principais de devam ser selecionadas. Para Ensor et al. (2017), há casos em que são utilizadas as componentes cuja soma de seus pesos γ_j seja superior a $0,9$. Mingoti (2005) também discute que a escolha de um peso de corte cabe ao pesquisador e pode variar de acordo com o problema.

Outra forma de estimar quantas componentes principais deverão ser utilizadas no problema é utilizar, como referência, a média aritimética ou média geométrica dos autovalores tirados tanto da matriz de covariância dos dados originais quanto dos dados padronizados. Assim, os autovalores médios podem ser calculados por

$$\bar{\lambda} = \frac{1}{p} \sum_{j=0}^{p} \lambda_j \tag{3.24}$$

e

$$\bar{\lambda}_G = \left(\prod_{j=0}^{p} \lambda_j \right)^{\frac{1}{p}} \tag{3.25}$$

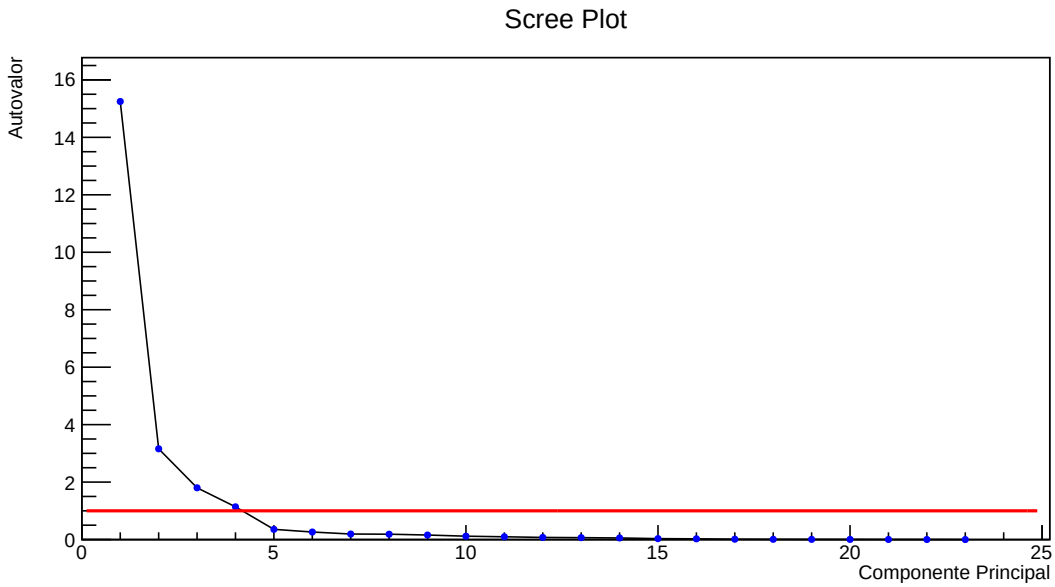
onde $\bar{\lambda}$ e $\bar{\lambda_G}$ são os valores da média aritmética e geométrica respectivamente. A ideia deste método é que uma PC com variância menor que a média das variâncias comportará menos informação que qualquer uma das variáveis originais e portanto pode ser desprezada (FERREIRA, 2008; MINGOTI, 2005). Ao se utilizar a matriz de correlação, a média simples será $\bar{\lambda} = 1$ e portanto as PC cujos autovalores forem menores que 1 devem ser desconsideradas (ENSOR et al., 2017).

Já nos casos em que as variâncias das componentes principais são muito discrepantes, Ferreira (2008) argumenta que a média geométrica, por apresentar um valor menor que a média aritmética, pode ser mais conservadora e acabar retendo alguma componente cuja variância é maior que a variabilidade da maioria dos dados originais. Entretanto, essa regra não deve ser aplicada sem uma interpretação do pesquisador, pois é possível que uma componente que possua um autovalor abaixo da média contenha informações importantes para o problema e cabe a quem estiver utilizando esta análise tomar esta decisão.

Segundo Johnson e Wichern (2007), outro método de escolha das componentes principais é através de uma representação gráfica, denominada *screeplot* (Figura 4), onde

no eixo das ordenadas está o valor numérico ou relativo de cada autovalor e no eixo das abscissas está a ordem, decrescente, dos autovalores. Em um *scree plot*, é possível observar que a nuvem de pontos se assemelha a um cotovelo e o método consiste em reter as PCs que estiverem acima do ponto onde os autovalores passam a ter, aproximadamente, o mesmo valor ou se tornem suficientemente pequenos.

Figura 4 – Um *Scree Plot*, gráfico de autovalor por componente principal.



Fonte: Adaptado de Ensor et al. (2017).

Conforme pode ser observado na Figura 4, ao analisar dados do meio interestelar, Ensor et al. (2017) utilizaram uma combinação de métodos para escolher as PCs. Foi averiguado a média dos autovalores da matriz de correlação e usado este resultado para determinar o ponto que forma o "cotovelo"do *scree plot*. Deste modo, foram escolhidas apenas as quatro primeiras componentes principais. Por fim, é importante destacar que há vários estudos sobre como escolher as melhores componentes principais, mas não há uma regra específica de modo que cabe ao pesquisador as escolhas para cada problema. Há, também, outros métodos como o logarítmo do autovalor e o teste de hipótese de que os $p - k$ autovalores são iguais, que não foram descritos neste trabalho (FERREIRA, 2008; MINGOTI, 2005).

4 ANÁLISE DISCRIMINANTE

A análise discriminante é uma ferramenta de estatística multivariada cujo objetivo é determinar uma regra para classificação ou para separação de um conjunto de dados previamente observados em diferentes grupos para que futuramente essa regra possa classificar novas observações realizadas (HUBERT; DRIESSEN, 2004). Já para Tang e Li (2016), a análise discriminante, também conhecida como Análise Discriminante Linear de Fisher ou *Linear Discriminant Analysis* (LDA), é um método bastante utilizado para reduzir a dimensão de um conjunto de dados e que busca uma combinação linear entre diferentes características que possam separar dois ou mais objetos ou eventos.

Segundo Aerts e Wilms (2017), a LDA é uma das técnicas de classificação mais aplicadas em casos sujeitos à normalidade, tanto Tang e Li (2016) quanto Zheng et al. (2017) citam sua aplicação na área da saúde em diagnosticar doenças em pacientes, estudo de genética e imagens cerebrais, por se tratarem de efeitos com múltiplas causas, já Ferrer (2017) prôpos um novo modelo de reconhecimento de voz utilizando análise discriminante probabilística. Para Ferreira (2008), bancos podem utilizar LDA para classificar seus clientes em bons ou maus pagadores e posteriormente classificar novos clientes nestas categorias. Enfim, quando se trata de métodos de classificação em que se tem diversas variáveis capazes de explicar a segregação de diferentes grupos de observações, a análise discriminante é amplamente utilizada devido à sua simplicidade, eficiência e por ser de fácil interpretação (ZHENG et al., 2017).

Neste capítulo será introduzido toda a estrutura de cálculo da Análise Discriminante Linear de Fisher. Será abordada uma regra de classificação para os casos em que as observações devem ser discriminadas em apenas dois grupos, para um caso generalizado com $k \geq 2$ grupos e, também, uma avaliação da regra de classificação utilizada. É importante ressaltar que uma observação terá valores para as p variáveis discriminantes, $\boldsymbol{x}^T = [x_1, x_2, \cdots, x_p]$, e pode, a *priori*, ser classificada em k diferentes populações com uma função densidade de probabilidade qualquer, $f_i(x)$, com $i = 1, 2, \cdots, k$, se diferenciando apenas pelos parâmetros da distribuição, de modo que todas as populações respeitem a mesma fdp. Contudo, neste trabalho, será considerado que as k populações respeitam uma distribuição normal de probabilidade.

4.1 Análise Discriminante Linear de Fisher para duas populações

A proposta dada por Fisher de uma função discriminante capaz de classificar observações em uma de duas populações, foi a de transformar as variáveis aleatórias multidimensionais $\boldsymbol{x}$ em variáveis unidimensionais y de modo que estas variáveis fossem uma

combinação linear dada por

$$y = \boldsymbol{a}^T \cdot \boldsymbol{x}, \tag{4.1}$$

onde $\boldsymbol{a}^T$ é o vetor de transformação.

Assim, é possível obter valores $y_{11}, y_{12}, \cdots, y_{1n_1}$ e $y_{21}, y_{22}, \cdots, y_{2n_2}$, onde o primeiro índice refere-se á população e o segundo á observação, pertencentes às populações Π_1 e Π_2 respectivamente, que são denominados *scores* discriminantes. A princípio, a técnica de Fisher não exige que as populações sejam distribuídas de acordo com alguma função densidade de probabilidade específica, contudo é necessário que o conjunto de dados possua uma matriz comum de covariância. A diferença entre as duas populações se dá através da distância entre os valores médios de cada população de modo que a transformação linear, dada pela equação (4.1), deve maximizar essa distância (JOHNSON; WICHERN, 2007). Segundo Ferreira (2008), a combinação linear deve ser tal que maximize a distância estatística quadrática entre as médias dos *scores* discriminantes ou *score* discriminante médio, pois abrangeria tanto a distância euclidiana entre as médias das duas populações quanto a variância de y. As esperanças do conjunto de dados que pertencem às populações Π_1 e Π_2 também são conhecidas como centroide do grupo Π_1 e Π_2. Portanto, se esses centroides podem ser calculados por, respectivamente, $E(\boldsymbol{x}|\boldsymbol{x} \in \Pi_1) = \boldsymbol{\mu}_1$ e $E(\boldsymbol{x}|\boldsymbol{x} \in \Pi_2) = \boldsymbol{\mu}_2$, então as médias da variável escalar y para ambas as populações são

$$E(y|\Pi_1) = E(\boldsymbol{a}^T\boldsymbol{x}|\Pi_1) = \boldsymbol{a}^T E(\boldsymbol{x}|\Pi_1) = \boldsymbol{a}^T\boldsymbol{\mu}_1 \tag{4.2}$$

e

$$E(y|\Pi_2) = E(\boldsymbol{a}^T\boldsymbol{x}|\Pi_2) = \boldsymbol{a}^T E(\boldsymbol{x}|\Pi_2) = \boldsymbol{a}^T\boldsymbol{\mu}_2 \tag{4.3}$$

e a variância de y, independentemente da população de origem, é

$$Var(\boldsymbol{a}^T\boldsymbol{x}) = \boldsymbol{a}^T Var(\boldsymbol{x})\boldsymbol{a} = \boldsymbol{a}\boldsymbol{\Sigma}\boldsymbol{a}. \tag{4.4}$$

Assim, essa distância quadrática pode ser descrita por

$$D^2 = \frac{[E(y|\Pi_1) - E(y|\Pi_2)]^2}{Var(\boldsymbol{a}^T\boldsymbol{x})}, \tag{4.5}$$

ou

$$D^2 = \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_1 - \boldsymbol{a}^T\boldsymbol{\mu}_2)^2}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} \tag{4.6}$$

$$D^2 = \frac{[\boldsymbol{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} \tag{4.7}$$

onde o $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ é a diferença entre os centroides das populações Π_1 e Π_2. Logo o objetivo da análise de Fisher é determinar o vetor $\boldsymbol{a}$ que maximiza esta distância. Para isso, Johnson e Wichern (2007) sugerem utilizar a desigualdade extendida de Cauchy-Schwarz em que dado dois vetores $\boldsymbol{u}(p \times 1)$ e $\boldsymbol{v}(p \times 1)$ e uma matriz $\boldsymbol{B}(p \times p)$ positiva definida tem-se que $(\boldsymbol{u}^T\boldsymbol{v})^2 \leq (\boldsymbol{u}^T\boldsymbol{B}\boldsymbol{u})(\boldsymbol{v}^T\boldsymbol{B}^{-1}\boldsymbol{v})$ onde a igualdade ocorre quando $\boldsymbol{u} = c\boldsymbol{B}^{-1}\boldsymbol{v}$

ou $\boldsymbol{v} = c\boldsymbol{B}\boldsymbol{u}$ para alguma constante c. Assim, se for considerado, na equação (4.7) que $\boldsymbol{a} = \boldsymbol{u}$, $\boldsymbol{v} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ e $\boldsymbol{\Sigma} = \boldsymbol{B}$, então a desigualdade extendida de Cauchy-Schwarz se torna

$$[\boldsymbol{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2 \leq (\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)], \tag{4.8}$$

que ao dividir os dois lados da desigualdade por $(\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})$ obtém-se

$$\frac{[\boldsymbol{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{(\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})} \leq [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \tag{4.9}$$

cujo máximo ocorrerá quando a igualdade for satisfeita, ou seja, quando $\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Assim, conclui-se que a distância quadrática máxima visada pela análise de Fisher é

$$D^2 = [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \tag{4.10}$$

onde D^2 é conhecido como distância de Mahalanobis e representa a maior distância entre as médias das populações. Segundo Bodnar et al. (2017), $\boldsymbol{a}^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}$ são os coeficientes discriminantes da função discriminante linear de Fisher que, de acordo com Johnson e Wichern (2007), Ferreira (2008) e Härdle e Simar (2015), é dada pela seguinte combinação linear

$$y = \boldsymbol{a}^T\boldsymbol{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}. \tag{4.11}$$

Uma vez determinada a função discriminante, deve-se obter uma regra de classificação. Para duas populações a maneira mais tradicional é encontrar o valor médio m da distância máxima entre as médias das funções discriminantes aplicadas em cada centroide. Essa distância, é a própria distância de Mahalanobis e portanto uma observação $\boldsymbol{x}$ será classificada na população Π_1 se seu *score* discriminate, $\boldsymbol{a}^T\boldsymbol{x}$, estiver mais próximo de seu *score* médio, $\boldsymbol{a}^T\boldsymbol{\mu}_1$, e em Π_2 caso contrário (JOHNSON; WICHERN, 2007). Assim, o ponto médio será

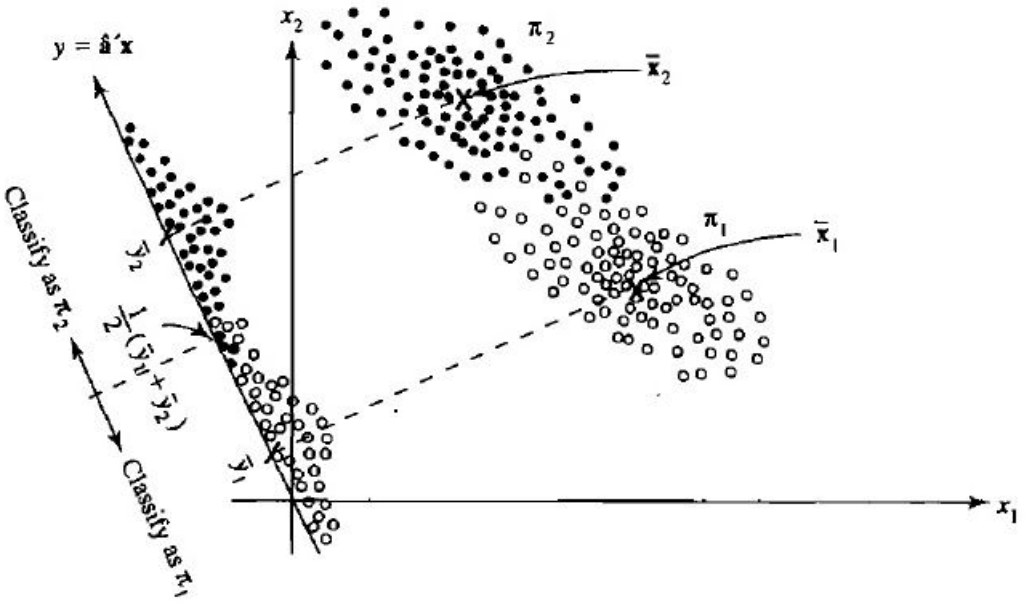
$$m = \frac{1}{2}\{E[y|\Pi_1] - E[y|\Pi_2]\} = \frac{1}{2}(\boldsymbol{a}^T\boldsymbol{\mu}_1 - \boldsymbol{a}^T\boldsymbol{\mu}_2) \tag{4.12}$$

que pode ser escrito como

$$m = \frac{1}{2}\boldsymbol{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{4.13}$$

e lembrando que $\boldsymbol{a}^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}$, então pode-se verificar que $m = (\frac{1}{2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ é de fato a metade da distância de Mahalanobis entre os centroides. Toda esta dedução pode ser melhor interpretada pela seguinte Figura

Figura 5 – Uma representação gráfica da LDA aplicada em duas populações com duas variáveis discriminantes, onde $\bar{\boldsymbol{x}}$ equivale ao centroide $\hat{\boldsymbol{\mu}}$ de cada população e $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ é a metade da distância de Mahalanobis m.



Fonte (JOHNSON; WICHERN, 2007).

É importante destacar que os valores de média e covariância utilizados nos cálculos são considerados puramente teóricos e desta forma é necessário utlizar seus estimadores que, na ausência de incertezas experimentais, são

$$\hat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{x}_{1i} \tag{4.14}$$

e

$$\hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{x}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{x}_{2i} \tag{4.15}$$

Já a matriz de covariância comum, é dada por

$$\boldsymbol{S}_c = \sum_{j=1}^{k}(n_j - 1)\left(\frac{\boldsymbol{S}_j}{n-k}\right) \tag{4.16}$$

onde k é o número de populações e $\boldsymbol{S}_j$ é a matriz de variância e covariância amostral dada pela equação (2.31) (BODNAR et al., 2017). Desta forma, a matriz de covariância comum para duas populações será

$$\boldsymbol{S}_c = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2} \tag{4.17}$$

e Johnson e Wichern (2007) destaca que o tamanho das amostras, $n_1 + n_2$, deve ser maior ou igual ao número de variáveis discriminantes, p, pois em caso contrário a matriz $\boldsymbol{S}_c$ será

singular e sua inversa não existirá. Em situações reais e na ausência de informação sobre as incertezas experimentais, a função discrminante linear de Fisher bem como o ponto médio de classificação, será respectivamente

$$\hat{y} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \boldsymbol{S}_c^{-1} \boldsymbol{x} \tag{4.18}$$

$$\hat{m} = \frac{1}{2}\hat{D}^2 \tag{4.19}$$

onde

$$\hat{D}^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \boldsymbol{S}_c^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \tag{4.20}$$

é um estimador da distância real de Mahalanobis entre duas populações (FERREIRA, 2008).

Conforme Johnson e Wichern (2007), essa classificação só fará sentido se houver uma diferença entre as médias das duas populações baseada em um teste de significância. Portanto, para duas populações com distribuição normal multivariada e com matriz de covariância comum, pode-se realizar um teste de hipótese onde a hipótese nula é de que as médias são iguais e a hipótese alternativa é de que são diferentes. Assim, utilizando a distribuição F, pode-se rejeitar a hipótese nula para um dado nível de significância, α, quando

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\hat{D}^2$$

for maior que o quantil superior da distribuição F_{α,ν_1,ν_2}, onde $\nu_1 = p$ e $\nu_2 = n_1 + n_2 - 1 - p$ são graus de liberdade. Assim, se a hipótese nula for rejeitada, conclui-se que a separação entre as duas populações é significativa. Ainda para Johnson e Wichern (2007), mesmo que a hipótese nula seja rejeitada não há garantias de que a regra de classificação será de boa qualidade e para isso é necessário estimar a probabilidade de se classificar uma determinada observação em uma população incorreta. Desta forma, é de suma importância descrever um método para estimar tais probabilidades, conforme será visto a seguir.

4.2 Estimativa das Probabilidades de Classificação Incorreta

Foi abordado no tópico anterior como encontrar a função discriminante linear de Fisher para discriminar observações em duas diferentes populações. No entanto, uma vez determinada uma regra de classificação é necessário avaliar seu poder discriminante. Portanto, neste tópico serão abordados dois métodos de estimar as probabilidades de se classificar uma observação em um grupo inadequado, o Método de Amostra de Validação e o Método de Validação Cruzada. Em ambos os métodos, deve-se conhecer previamente a classificação de todas as observações na amostra de teste (FERREIRA, 2008).

O primeiro deles consiste em dividir as observações em duas partes, uma para análise e outra para teste. Os dados do grupo de análise serão utilizados para realizar o cálculo das funções discriminantes, que uma vez ajustadas, servirão para classificar os dados do grupo de teste. A importância de se conhecer previamente em qual grupo ou população os dados pertencem é que após feita a classificação do grupo de teste é possível verificar quais observações foram alocadas incorretamente (HAIR et al., 2009).

Para Ferreira (2008) este método possui como ponto fraco não ter validade em amostras pequenas e também a perda de informações importantes para a determinação das funções discriminantes uma vez que os dados foram divididos em duas partes. Hair et al. (2009) sugerem que uma forma de aumentar a confiança deste método é realizar diversas vezes, e aleatoriamente, a escolha dos dados que pertencerão aos grupos de análise e teste. Assim, seria possível estimar uma probabilidade de classificação incorreta diferentes vezes e ao final obter um valor médio para essas probablidades. Lembrando que em cada escolha, novas funções discriminantes deverão ser ajustadas.

O segundo método, o de validação cruzada, é comparado ao método de *Jackknife* e consiste em realizar o ajuste das funções discriminantes deixando uma observação de fora. Contudo, apesar desta observação não ter sido usada para o cálculo das funções discriminantes ela é utilizada para avaliá-las, pois como já se conhece o grupo ao qual essa observação pertence é possível saber se as funções ajustadas a classificaram no grupo correto (FERREIRA, 2008). Este método também apresenta fragilidades em amostras pequenas sendo que é aconselhado utilizá-lo quando o tamanho da menor população for no mínimo de três a cinco vezes maior que o número de variáveis discriminantes (HAIR et al., 2009).

Tanto Ferreira (2008) quanto Hair et al. (2009) sugerem que após contabilizar o número de observações classificadas nas populações corretas e incorretas pode-se construir uma matriz de classificação em que os elementos da diagonal principal, σ_{ii}, correspondem ao número de acertos e o restante, $\sigma_{ij}, i \neq j$, representam o número de erros. Assim, a matriz de classificação será da forma

Tabela 1 – Matriz de classificação onde as linhas representam as populações reais e as colunas as populações classificadas.

Real/Classificada	Π_1	Π_2	$\cdots$	Π_k	Total
Π_1	σ_{11}	σ_{12}	$\cdots$	σ_{1k}	n_1
Π_2	σ_{21}	σ_{22}	$\cdots$	σ_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Π_k	σ_{k1}	σ_{k2}	$\cdots$	σ_{kk}	n_k

Fonte: Do autor.

Desta forma, Ferreira (2008) afirma que as probabilidades de se classificar uma

observação em Π_i sendo que ela pertence à população Π_j com $i \neq j$, será

$$p(i|j) = \frac{n_{ji}}{n_j} \tag{4.21}$$

e a probabilidade total, denominada taxa de erro, será

$$P_{Total} = \frac{\sum_{i \neq j=1}^{k} \sigma_{ij}}{\sum_{i=1}^{k} n_i} \tag{4.22}$$

No caso particular de duas populações, $k = 2$, a equação (4.21) gera

$$p(2|1) = \frac{n_{12}}{n_1} \tag{4.23}$$

e

$$p(1|2) = \frac{n_{21}}{n_2}. \tag{4.24}$$

Já a equação (4.22) gera

$$P_{Total} = \frac{n_{12} + n_{21}}{n_1 + n_2}. \tag{4.25}$$

4.3 Análise Discriminante Linear de Fisher para mais de duas populações

Neste tópico será abordado o caso em que se têm $k > 2$ populações nas quais as observações podem ser classificadas. De forma análoga ao caso particular de $k = 2$, as populações possuem distribuição gaussiana multivariada e uma matriz comum de covariância. Após a coleta de um conjunto de dados para análise, cada população, Π_j, possuirá um número, $n_j (j = 1, 2, \cdots, k)$, de observações p dimensionais, $\boldsymbol{x}^T = [x_1, x_2, \cdots, x_p]$, onde cada elemento deste vetor representará uma variável discriminante e cada população terá um vetor médio, ou centroide, dado por $\boldsymbol{\mu}_j$ (FERREIRA, 2008). Assim, segundo Johnson e Wichern (2007) é possível obter um vetor médio dos centroides de todas as populações, dado por

$$\bar{\boldsymbol{\mu}} = \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\mu}_j \tag{4.26}$$

Para Härdle e Simar (2015), a análise discriminante linear de Fisher é descrita por uma combinação linear que independe de quantas populações existam. Portanto, igualmente ao tópico anterior, as funções discriminantes lineares de Fisher serão da forma $y = \boldsymbol{a}^T \boldsymbol{x}$. Johnson e Wichern (2007) afirmam que a esperança da variável y, ou seja o *score* médio para uma população Π_j será

$$E[y|\Pi_j] = E[\boldsymbol{a}^T \boldsymbol{x}|\Pi_j] = \boldsymbol{a}^T E[\boldsymbol{x}|\Pi_j] = \boldsymbol{a}^T \boldsymbol{\mu}_j = \bar{y}_j \tag{4.27}$$

e sua variância

$$Var[y|\Pi_j] = Var[\boldsymbol{a}^T \boldsymbol{x}|\Pi_j] = \boldsymbol{a}^T Var[\boldsymbol{x}|\Pi_j] \boldsymbol{a} = \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a} \tag{4.28}$$

onde $\boldsymbol{\Sigma}$ é a matriz comum de covariância. Consequentemente, como cada população possui um centroide diferente, a esperança total da variável y será

$$E[E[y|\Pi_j]] = E[\boldsymbol{a}^T\boldsymbol{\mu}_j] = \boldsymbol{a}^T E[\boldsymbol{\mu}_j] = \boldsymbol{a}^T \frac{1}{k}\sum_{j=1}^{k}\boldsymbol{\mu}_j = \boldsymbol{a}^T\boldsymbol{\mu}_j. \tag{4.29}$$

Ferreira (2008) afirma que Fisher teve como ponto de partida o objetivo de determinar o vetor $\boldsymbol{a}$ de forma que a razão entre a variância das populações e a variância comum dentro das populações fosse maximizada. Assim, seja a soma das diferenças entre a média da função discriminante para uma população, dada pela equação (4.27), e a média total da função, conforme a equação (4.29), dada por

$$\frac{\sum_{j=1}^{k}(E[y|\Pi_j] - E[E[y|\Pi_j]])^2}{Var[y|\Pi_j]} \tag{4.30}$$

e substituindo as equações (4.27) e (4.29) na equação (4.30), lembrando que a matriz de covariância é comum a todas as populações, obtém-se

$$\frac{\sum_{j=1}^{k}(\boldsymbol{a}^T\boldsymbol{\mu}_j - \boldsymbol{a}^T\bar{\boldsymbol{\mu}})^2}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} = \frac{\sum_{j=1}^{k}[\boldsymbol{a}^T(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})]^2}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} \tag{4.31}$$

que pode ser reescrita da forma

$$\frac{\sum_{j=1}^{k}[\boldsymbol{a}^T(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})]^2}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} = \frac{\sum_{j=1}^{k}\boldsymbol{a}^T(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})^T\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} \tag{4.32}$$

ou

$$\frac{\sum_{j=1}^{k}\boldsymbol{a}^T(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})^T\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} = \frac{\boldsymbol{a}^T\left[\sum_{j=1}^{k}(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})^T\right]\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}} \tag{4.33}$$

Härdle e Simar (2015) abordam que o termo $\sum_{j=1}^{k}(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})^T$ é uma matriz quadrada $(p \times p)$ denominada matriz de somas de quadrados e produtos entre os grupos. Assim, se $\boldsymbol{B} = \sum_{j=1}^{k}(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})^T$, então a razão que Fisher buscava maximizar é

$$\frac{\sum_{j=1}^{k}(E[y|\Pi_j] - E[E[y|\Pi_j]])^2}{Var[y|\Pi_j]} = \frac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}}, \tag{4.34}$$

em que o máximo desta razão é igual ao maior autovalor da matriz $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ e o vetor, $\boldsymbol{a}$, responsável por esse resultado é o autovetor correspondente a este autovalor.

Uma vez determinado como calcular a função discriminante linear de Fisher é importante estabelecer qual será o critério de classificação de novas observações. Antes disto, é importante observar que a solução do problema de maximização da equação (4.34) é dada pelo maior autovalor, e seu correspondente autovetor, da matriz $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$. Contudo o autovetor correspondente ao menor autovalor fornece o mínimo da equação (4.34). Portanto, se existir p autovalores $(\lambda_1 > \lambda_2 > \cdots > \lambda_p)$ não nulos de $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$, existirá p vetores $\boldsymbol{a}$ que formarão p funções discriminantes que são denominadas de primeira função discrminante para λ_1 até p-ésima função discriminante para λ_p (JOHNSON; WICHERN, 2007).

Tanto Johnson e Wichern (2007), Ferreira (2008) quanto Fávero et al. (2009) afirmam que o número de funções discriminantes necessário para realizar a análise deverá respeitar a condição de $m = \min(p, k-1)$, onde p é o número de variáveis discriminantes e k o número de populações. Portanto, diferentemente do caso particular de duas populações, existirá um vetor de funções discriminantes, ou *scores* discriminantes de uma nova observação, dado por

$$\boldsymbol{y} = \boldsymbol{a}^T\boldsymbol{x} = \begin{pmatrix} \boldsymbol{a}_1^T\boldsymbol{x} \\ \boldsymbol{a}_2^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{a}_m^T\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \tag{4.35}$$

onde, agora, $\boldsymbol{a}^T$ é uma matriz $(m \times p)$ em que cada linha é um vetor de coeficientes discriminantes p-dimensional, como pode ser observado

$$\boldsymbol{a}^T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{pmatrix}. \tag{4.36}$$

Haverá, também, um vetor dos *scores* médios calculados com os dados da amostra de teste, dado por

$$\bar{\boldsymbol{y}}_j = \boldsymbol{a}^T\boldsymbol{\mu}_j = \begin{pmatrix} \boldsymbol{a}_1^T\boldsymbol{\mu}_j \\ \boldsymbol{a}_2^T\boldsymbol{\mu}_j \\ \vdots \\ \boldsymbol{a}_m^T\boldsymbol{\mu}_j \end{pmatrix}. \tag{4.37}$$

Assim, conforme Johnson e Wichern (2007) a equação (4.35) é um vetor cujas componentes possuem variâncias unitárias e são não correlacionadas entre si, e uma maneira de expressar a distância entre o *score* discriminante de uma nova observação e o *score* médio de uma população é dada pela distância quadrática expressa por

$$(\boldsymbol{y} - \bar{\boldsymbol{y}})^T(\boldsymbol{y} - \bar{\boldsymbol{y}}) = \sum_{j=1}^{m}(\boldsymbol{a}^T\boldsymbol{x} - \boldsymbol{a}^T\boldsymbol{\mu}_j)^2 = \sum_{j=1}^{m}\left[\boldsymbol{a}^T(\boldsymbol{x} - \boldsymbol{\mu}_j)\right]^2. \tag{4.38}$$

Logo, Ferreira (2008) e Johnson e Wichern (2007) abordam que a regra de classificação de Fisher será a de alocar a observação $\boldsymbol{x}$ naquela população que minimizar a distância descrita pela equação (4.38). Em outras palavras, essa distância, assim como para o caso particular de $k = 2$, representa a distância de Mahalanobis entre o *score* discriminante da observação, $\boldsymbol{y}$, e o *score* médio da população Π_j, $\bar{\boldsymbol{y}}_j = \boldsymbol{a}^T\boldsymbol{\mu}_j$. Portanto, novamente a regra de discriminação é determinada pela distância mínima de Mahalanobis (FÁVERO et al., 2009; HÄRDLE; SIMAR, 2015).

Conforme Johnson e Wichern (2007), da mesma forma que no caso particular de $k = 2$, os centroides de cada população devem ser estimados através da equação (2.19) e

cada grupo terá uma média das observações dada por

$$\bar{\boldsymbol{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \boldsymbol{x}_i, \tag{4.39}$$

assim como a média global. Logo

$$\bar{\boldsymbol{x}}_G = \frac{\sum_{j=1}^{k} \bar{\boldsymbol{x}}_j}{k}. \tag{4.40}$$

Já a matriz de covariância comum é estimada através da equação (4.16). Portanto, a função discrminante de Fisher, $y = \boldsymbol{a}^T\boldsymbol{x}$, será determinada, como já observado, com a maximização de

$$\frac{\hat{\boldsymbol{a}}^T\hat{\boldsymbol{B}}\hat{\boldsymbol{a}}^T}{\hat{\boldsymbol{a}}^T\boldsymbol{S}_c\hat{\boldsymbol{a}}^T},$$

onde $\hat{\boldsymbol{B}}$ é o estimador da matriz $\boldsymbol{B}$ que pode ser calculado por

$$\hat{\boldsymbol{B}} = \sum_{j=1}^{k}(\bar{\boldsymbol{x}}_j - \bar{\boldsymbol{x}}_G)(\bar{\boldsymbol{x}}_j - \bar{\boldsymbol{x}}_G)^T, \tag{4.41}$$

e $\hat{\boldsymbol{a}}$ é o estimador de $\boldsymbol{a}$, que pode ser obtido solucionando o sistema de equações

$$(\hat{\boldsymbol{B}} - \hat{\lambda}_j\boldsymbol{S}_c)\hat{\boldsymbol{a}}_j = \boldsymbol{0}, \tag{4.42}$$

que manipulando a equação tem-se

$$(\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}} - \hat{\lambda}_j\boldsymbol{I})\hat{\boldsymbol{a}}_j = \boldsymbol{0}, \tag{4.43}$$

onde $\hat{\lambda}_j$ e $\hat{\boldsymbol{a}}_j$ são, respectivamente, os autovalores e autovetores da matriz $\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}}$.

Tanto Ferreira (2008) quanto Johnson e Wichern (2007) abordam que também é possível resolver o sistema de equações dado pela equação (4.42) realizando uma troca de variáveis. Assim, se $\hat{\boldsymbol{a}}_j = \boldsymbol{S}_c^{-\frac{1}{2}}\hat{\boldsymbol{e}}_j$ então a equação (4.42) se torna

$$(\hat{\boldsymbol{B}} - \hat{\lambda}_j\boldsymbol{S}_c)\boldsymbol{S}_c^{-\frac{1}{2}}\hat{\boldsymbol{e}}_j = \boldsymbol{0}, \tag{4.44}$$

e como $\boldsymbol{S}_c^{-\frac{1}{2}}\boldsymbol{S}_c\boldsymbol{S}_c^{-\frac{1}{2}} = \boldsymbol{I}$ então

$$(\boldsymbol{S}_c^{-\frac{1}{2}}\hat{\boldsymbol{B}}\boldsymbol{S}_c^{-\frac{1}{2}} - \hat{\lambda}_j\boldsymbol{I})\hat{\boldsymbol{e}}_j = \boldsymbol{0}, \tag{4.45}$$

onde λ_j permanece o mesmo, ou seja, os autovalores de $\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}}$ não variam com essa troca de variáveis realizada. Já $\hat{\boldsymbol{e}}_j$ são os autovetores da matriz $\boldsymbol{S}_c^{-1\frac{1}{2}}\hat{\boldsymbol{B}}\boldsymbol{S}_c^{-\frac{1}{2}}$, que uma vez calculados são usados para determinar os vetores da combinação linear de Fisher pela transformação apresentada, ou seja,

$$\hat{\boldsymbol{a}}_j = \boldsymbol{S}_c^{-\frac{1}{2}}\hat{\boldsymbol{e}}_j. \tag{4.46}$$

A eficácia da regra de classificação para $k > 2$ é desconhecida, contudo a avaliação do poder discriminante da função de Fisher pode ser feita através dos dois métodos abordados no tópico anterior, o Método de Amostra de Validação e o Método de Validação Cruzada, utilizando da mesma forma a matriz de classificação e calculando as probabilidades de classificação incorreta através das equações (4.21) e (4.22) (JOHNSON; WICHERN, 2007).

5 ANÁLISE DE CORRELAÇÃO CANÔNICA

A análise de Correlação Canônica ou do inglês *Canonical Correlation Analysis* (CCA), pode ser interpretada como uma generalização da análise de Regressão Linear Múltipla, pois para cada observação, além das p variáveis explicativas (independentes), $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^T$, pode haver, também, q variáveis a serem explicadas (dependentes), $\boldsymbol{y} = (y_1, y_2, \cdots, y_q)^T$, onde p e q não são necessariamente iguais (HAIR et al., 2009). Esta ferramenta busca determinar o comportamento linear entre dois grupos de variáveis. Com isso, o método consiste em encontrar novas variáveis (variáveis canônicas) que possuam máxima correlação entre si (correlação canônica), através de combinações lineares de cada um dos conjuntos de variáveis aleatórias (CHU et al., 2013).

Quando Hotelling (1936) demonstrou pela primeira vez a análise de correlação canônica, ele discutiu em seu trabalho que a correlação entre duas medidas não é aplicada apenas a variáveis unidimensionais, mas também a casos em que cada variável é um vetor multidimensional, como por exemplo, para a mesma pessoa, obter a pontuação de diversos testes de raciocínio juntamente com medidas de parâmetros físicos. Hotelling (1936) argumentou que a relação entre essas medidas (testes de raciocínio e parâmetros físicos) deveriam permanecer invariantes sob transformações lineares, e com isso ele demonstrou ser possível estabelecer pares de variáveis, onde cada variável seria uma função linear de um desses conjuntos de informações.

A CCA é uma técnica que, assim como as outras duas técnicas abordadas neste trabalho, também pode ser utilizada para reduzir o espaço dimensional de um determinado problema, interpretar o comportamento de diferentes fenômenos e fazer previsões de novos resultados (LEE, 2016). Jones et al. (2016) fez seu uso na previsão de possíveis cirurgias em pacientes com determinadas doenças, analisando dados médicos. Já Gao et al. (2017) utilizaram a CCA para desenvolver um algorítmo com melhor desempenho para solucionar problemas que envolvem *Empirical Risk Minimization* (ERM).

Com o avanço do poder computacional, novos métodos e aplicações com embasamento na análise de correlação canônica estão sendo desenvolvidos. Como exemplo, a técnica de análise de correlação canônica de vetores singulares, que segundo Raghu et al. (2017) é uma combinação da decomposição de valor singular com correlação canônica, capaz de realizar análises computacionalmente mais eficazes, e com aplicações em NRL (*Network Representation Learning*), uma ferramenta capaz de aprender com novas análises e amplamente utilizada para reconhecimento de imagens, busca de textos entre outras aplicações de inteligência artificial (YANG et al., 2015).

Como pode ser observado, a CCA tem um grande potencial para análises de dados multivariados, e neste capítulo será demonstrada toda a estrutura matemática necessária para realização da análise de correlação canônica, na ausência de informações sobre as

incertezas experimentais das medidas.

5.1 Análise de Correlações Canônicas por variáveis não padronizadas

Para realizar a análise de correlação canônica é necessário que cada indivíduo observado possua dois vetores de informações, $\boldsymbol{x} = [x_1, x_2, \cdots, x_p]$ e $\boldsymbol{y} = [y_1, y_2, \cdots, y_q]$, onde p e q não são necessariamente iguais, e um destes vetores pode ser denomiado de variável explicadora enquanto o outro é a variável a ser explicada. Assim, para um conjunto de n observações haverá um grupo com n vetores p-dimensionais e outro grupo com n vetores q-dimensionais que possuirão os vetores médios $\boldsymbol{\mu_x}$ e $\boldsymbol{\mu_y}$ com as respectivas dimensões. Os grupos terão uma matriz de covariância interna $\boldsymbol{\Sigma}_x$, de dimensão $(p \times p)$, e $\boldsymbol{\Sigma}_y$, de dimensão $(q \times q)$, bem como uma matriz de covariância entre eles, $\boldsymbol{\Sigma}_{xy}$, cuja dimensão é $(p \times q)$ (CHU et al., 2013).

A CCA, segundo Benton et al. (2017), consiste em determinar a projeção linear de dois vetores que possuam correlação máxima, ou seja, realizar uma transformação linear nos vetores $\boldsymbol{x}$ e $\boldsymbol{y}$ de modo que as novas variáveis, oriundas dessas transformações, possuam uma correlação linear maximizada. As novas variáveis podem ser expressas por

$$U = \boldsymbol{a}^T \boldsymbol{x} \tag{5.1}$$

e

$$V = \boldsymbol{b}^T \boldsymbol{y}, \tag{5.2}$$

onde $\boldsymbol{a}$ e $\boldsymbol{b}$ são vetores de transformação, p-dimensional e q-dimensional respectivamente, e U e V são escalares denominados como variáveis canônicas.

A correlação linear do par U e V, também denominada correlação canônica, é dada por

$$\rho_{U,V} = \frac{Cov[U,V]}{\sqrt{Var[U]Var[V]}}, \tag{5.3}$$

onde $Cov[U,V]$ é a covariância entre as variáveis canônicas e $Var[U]$ e $Var[V]$ são as variâncias das variáveis U e V respectivamente (SUO et al., 2017; MIN; LIU; ZHANG, 2017).

Ao substituir as equações (5.1, 5.2) em cada um dos termos de (5.3) tem-se que

$$Var[U] = Var[\boldsymbol{a}^T \boldsymbol{x}] = \boldsymbol{a}^T Var[\boldsymbol{x}]\boldsymbol{a} = \boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a}, \tag{5.4}$$

$$Var[V] = Var[\boldsymbol{b}^T \boldsymbol{y}] = \boldsymbol{b}^T Var[\boldsymbol{y}]\boldsymbol{b} = \boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b} \tag{5.5}$$

e

$$Cov[U,V] = Cov[\boldsymbol{a}^T \boldsymbol{x}, \boldsymbol{b}^T \boldsymbol{y}] = \boldsymbol{a}^T Cov[\boldsymbol{x}, \boldsymbol{y}]\boldsymbol{b} = \boldsymbol{a}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{b}, \tag{5.6}$$

e desta maneira, a equação (5.3) se torna

$$\rho_{U,V} = \frac{\boldsymbol{a}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{b}}{\sqrt{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})}}. \tag{5.7}$$

Portanto, o método consiste em determinar os vetores $\boldsymbol{a}$ e $\boldsymbol{b}$ tal que a correlação canônica, $\rho_{U,V}$, seja maximizada, conforme foi proposto Hotelling (1936).

Para encontrar os vetores $\boldsymbol{a}$ e $\boldsymbol{b}$ que maximizam a correlação canônica, Ferreira (2008) sugere derivar $\rho_{U,V}$ com relação a cada um dos vetores e igualar o resultado das duas equações a zero para então determinar as incógnitas. Assim, derivando com relação ao vetor $\boldsymbol{a}$, tem-se

$$\frac{\partial \rho_{U,V}}{\partial \boldsymbol{a}} = \frac{\boldsymbol{\Sigma}_{xy}\boldsymbol{b}(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}} - (\boldsymbol{a}^T \boldsymbol{\Sigma}_{xy}\boldsymbol{b})(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{-\frac{1}{2}}\boldsymbol{\Sigma}_x \boldsymbol{a}}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})} \tag{5.8}$$

que pode ser reescrita como

$$\frac{\partial \rho_{U,V}}{\partial \boldsymbol{a}} = \frac{\boldsymbol{\Sigma}_{xy}\boldsymbol{b}(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})} - \frac{(\boldsymbol{a}^T \boldsymbol{\Sigma}_{xy}\boldsymbol{b})(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{-\frac{1}{2}}\boldsymbol{\Sigma}_x \boldsymbol{a}}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})} \tag{5.9}$$

e

$$\frac{\partial \rho_{U,V}}{\partial \boldsymbol{a}} = \frac{\boldsymbol{\Sigma}_{xy}\boldsymbol{b}}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}} - \frac{(\boldsymbol{a}^T \boldsymbol{\Sigma}_{xy}\boldsymbol{b})}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})}\frac{(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}\boldsymbol{\Sigma}_x \boldsymbol{a}}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}} \tag{5.10}$$

como os termos $(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}$ e $(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}$ são escalares, podemos fatorá-los, e assim

$$\frac{\partial \rho_{U,V}}{\partial \boldsymbol{a}} = \frac{1}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}}\left[\boldsymbol{\Sigma}_{xy}\boldsymbol{b} - \frac{(\boldsymbol{a}^T \boldsymbol{\Sigma}_{xy}\boldsymbol{b})}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}}\frac{(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}\boldsymbol{\Sigma}_x \boldsymbol{a}}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}}\right]. \tag{5.11}$$

Igualando o resultado a zero e observando que o termo $\frac{(\boldsymbol{a}^T \boldsymbol{\Sigma}_{xy}\boldsymbol{b})}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}}$ é a própria correlação canônica, chega-se em

$$\boldsymbol{\Sigma}_{xy}\boldsymbol{b} - \rho_{U,V}\frac{(\boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b})^{\frac{1}{2}}}{(\boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a})^{\frac{1}{2}}}\boldsymbol{\Sigma}_x \boldsymbol{a} = \boldsymbol{0}. \tag{5.12}$$

Uma vez que o coeficiente de correlação linear é invariante com relação à mudança de escala pois $\rho_{kU,V} = \rho_{U,V}$ para qualquer que seja a constante k, então é possível redimensionar os vetores $\boldsymbol{a}$ e $\boldsymbol{b}$ tais que as variâncias sejam

$$Var[U] = \boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a} = 1$$

e

$$Var[V] = \boldsymbol{b}^T \boldsymbol{\Sigma}_y \boldsymbol{b} = 1$$

de modo que a correlação canônica dada pela equação (5.7) seja preservada (HÄRDLE; SIMAR, 2015). Assim, a equação (5.12) se torna

$$\boldsymbol{\Sigma}_{xy}\boldsymbol{b} - \rho_{U,V}\boldsymbol{\Sigma}_x \boldsymbol{a} = \boldsymbol{0}. \tag{5.13}$$

Ao realizar o mesmo procedimento de derivação com relação ao vetor $\boldsymbol{b}$ obtém-se

$$\boldsymbol{\Sigma}_{yx}\boldsymbol{a} - \rho_{U,V}\boldsymbol{\Sigma}_{y}\boldsymbol{b} = \mathbf{0}. \tag{5.14}$$

onde $\boldsymbol{\Sigma}_{yx} = \boldsymbol{\Sigma}_{xy}^{T}$.

Escolhendo, por exemplo, a equação (5.13) pode ser observado que

$$\rho_{U,V}\boldsymbol{\Sigma}_{x}\boldsymbol{a} = \boldsymbol{\Sigma}_{xy}\boldsymbol{b} \tag{5.15}$$

e portanto

$$\boldsymbol{a} = \frac{1}{\rho_{U,V}}\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{b}. \tag{5.16}$$

Substituindo a equação acima em (5.14), chega-se em

$$\frac{1}{\rho_{U,V}}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{b} - \rho_{U,V}\boldsymbol{\Sigma}_{y}\boldsymbol{b} = \mathbf{0} \tag{5.17}$$

que pode ser reescrito como

$$\frac{1}{\rho_{U,V}}(\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy} - \rho_{U,V}^{2}\boldsymbol{\Sigma}_{y})\boldsymbol{b} = \mathbf{0} \tag{5.18}$$

e multiplicando essa equação, pela esquerda, por $\boldsymbol{\Sigma}_{y}^{-1}$ tem-se

$$(\boldsymbol{\Sigma}_{y}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy} - \rho_{U,V}^{2}\boldsymbol{I})\boldsymbol{b} = \mathbf{0} \tag{5.19}$$

e portanto o vetor $\boldsymbol{b}$ que maximiza a correlação canônica é o autovetor correspondente ao maior autovalor $\rho_{U,V}^{2}$ da matriz $\boldsymbol{\Sigma}_{y}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy}$ (FERREIRA, 2008; MINGOTI, 2005). Como, para uma matriz com dimensão $n \times n$, por exemplo, o número de autovalores não nulos será igual a n, então existirão n autovetores que satisfazem a igualdade, considerando seus respectivos autovalores, e, portanto, n pares de variáveis canônicas. Desta forma, será abordado mais adiante neste capítulo, como escolher o número de pares para representar um determinado modelo.

Ao realizar o mesmo procedimento descrito acima para o vetor $\boldsymbol{a}$, a transformação linear que relaciona os vetores $\boldsymbol{a}$ e $\boldsymbol{b}$, bem como a equação homogênea resultante serão, respectivamente,

$$\boldsymbol{b} = \frac{1}{\rho_{U,V}}\boldsymbol{\Sigma}_{y}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{a}. \tag{5.20}$$

e

$$(\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{y}^{-1}\boldsymbol{\Sigma}_{yx} - \rho_{U,V}^{2}\boldsymbol{I})\boldsymbol{a} = \mathbf{0}. \tag{5.21}$$

Pode-se observar que $\rho_{U,V}^{2}$ é o autovalor tanto da matriz $\boldsymbol{\Sigma}_{y}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy}$ quanto da $\boldsymbol{\Sigma}_{x}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{y}^{-1}\boldsymbol{\Sigma}_{yx}$. Entretanto, apesar de ser possível determinar os dois vetores de maneira independente, Ferreira (2008) sugere determinar um deles por uma das duas equações (5.19, 5.21) e o outro pela transformação linear dada, respectivamente, pela equação (5.16) ou (5.20), principalmente se o objetivo for a redução de dimensionalidade do problema.

Uma vez determinado os vetores $\boldsymbol{a}$ e $\boldsymbol{b}$, deve-se substituí-los nas equações (5.1, 5.2) para obter as variáveis canônicas com máxima correlação entre si. Como $\boldsymbol{\mu_x}$, $\boldsymbol{\mu_y}$, $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$ e $\boldsymbol{\Sigma}_{xy}$ são apenas teóricos, na prática é preciso fazer uma estimação destes valores. Considerando que neste trabalho está sendo admitido que as variáveis aleatórias seguem uma distribuição normal de probabilidades, para um conjunto de n dados amostrais, os vetores médios podem ser estimados, na ausência de incertezas experimentais, através da equação (2.19) e as matrizes de covariâncias através da equação (2.31). Portanto, tem-se que

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i,$$

$$\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i,$$

são os vetores médios das variáveis $\boldsymbol{x}$ e $\boldsymbol{y}$ respectivamente, e

$$\boldsymbol{S}_x = \frac{1}{(n-1)}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T, \tag{5.22}$$

$$\boldsymbol{S}_y = \frac{1}{(n-1)}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T \tag{5.23}$$

e

$$\boldsymbol{S}_{xy} = \frac{1}{(n-1)}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T \tag{5.24}$$

são, na sequência, as matrizes de covariância das variáveis $\boldsymbol{x}$, das variáveis $\boldsymbol{y}$ e entre as variáveis $\boldsymbol{x}$ e $\boldsymbol{y}$ (MINGOTI, 2005). A partir destas equações, Johnson e Wichern (2007) discutem que as equações (5.16, 5.19) e (5.20, 5.21) tornam-se, na sequência,

$$\hat{\boldsymbol{a}} = \frac{1}{\hat{\rho}_{U,V}}\boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy}\boldsymbol{b} \tag{5.25}$$

$$(\boldsymbol{S}_y^{-1}\boldsymbol{S}_{yx}\boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy} - \hat{\rho}_{U,V}^2\boldsymbol{I})\hat{\boldsymbol{b}} = \boldsymbol{0} \tag{5.26}$$

e

$$\hat{\boldsymbol{b}} = \frac{1}{\hat{\rho}_{U,V}}\boldsymbol{S}_y^{-1}\boldsymbol{S}_{yx}\boldsymbol{a} \tag{5.27}$$

$$(\boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy}\boldsymbol{S}_y^{-1}\boldsymbol{S}_{yx} - \hat{\rho}_{U,V}^2\boldsymbol{I})\hat{\boldsymbol{a}} = \boldsymbol{0} \tag{5.28}$$

onde os vetores $\hat{\boldsymbol{a}}$ e $\hat{\boldsymbol{b}}$ são os vetores de transformação das variáveis originais. Por fim, as variáveis canônicas, geradas a partir dos dados originais, serão

$$\hat{U} = \hat{\boldsymbol{a}}^T\boldsymbol{x} \tag{5.29}$$

e

$$\hat{V} = \hat{\boldsymbol{b}}^T\boldsymbol{y} \tag{5.30}$$

5.2 Análise de Correlações Canônicas por variáveis padronizadas

É possível, também, realizar a análise de correlação canônica utilizando variáveis padronizadas, da mesma maneira como foi descrito para a análise de componentes principais. Assim, segundo Johnson e Wichern (2007), as i-ésimas observações, $\boldsymbol{x}_i = [x_1, x_2, \cdots, x_p]^T$ e $\boldsymbol{y}_i = [y_1, y_2, \cdots, y_p]^T$, com médias $\boldsymbol{\mu_x}$ e $\boldsymbol{\mu_y}$, bem como matrizes de covariâncias $\boldsymbol{\Sigma}_x$ e $\boldsymbol{\Sigma}_y$ respectivamente, podem ser transformadas em variáveis padronizadas por

$$\boldsymbol{z}_i = \boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{x}_i - \boldsymbol{\mu_x}) \tag{5.31}$$

e

$$\boldsymbol{w}_i = \boldsymbol{\Delta}^{-\frac{1}{2}}(\boldsymbol{y}_i - \boldsymbol{\mu_y}), \tag{5.32}$$

onde $\boldsymbol{z}_i$ e $\boldsymbol{w}_i$ são as variáveis padronizadas de $\boldsymbol{x}_i$ e $\boldsymbol{y}_i$, $\boldsymbol{\Lambda}^{-\frac{1}{2}}$ e $\boldsymbol{\Delta}^{-\frac{1}{2}}$ são matrizes diagonais com os elementos da diagonal principal de $\boldsymbol{\Sigma}_x$ e $\boldsymbol{\Sigma}_y$ respectivamente.

Uma consequência desta transformação, de acordo com Johnson e Wichern (2007), é que as matrizes de covariância dos dados padronizados (individualmente e entre si) são iguais às matrizes dos coeficientes de correlação linear dos dados originais, e podem ser descritas como

$$Var[\boldsymbol{z}] = \boldsymbol{\rho_x} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Sigma}_x\boldsymbol{\Lambda}^{-\frac{1}{2}}, \tag{5.33}$$

$$Var[\boldsymbol{w}] = \boldsymbol{\rho_y} = \boldsymbol{\Delta}^{-\frac{1}{2}}\boldsymbol{\Sigma}_y\boldsymbol{\Delta}^{-\frac{1}{2}} \tag{5.34}$$

e

$$Cov[\boldsymbol{z}, \boldsymbol{w}] = \boldsymbol{\rho_{xy}} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Sigma_{xy}}\boldsymbol{\Delta}^{-\frac{1}{2}}. \tag{5.35}$$

Portanto, as variáveis canônicas geradas a partir de variáveis padronizadas podem ser expressadas por

$$U^* = \boldsymbol{a}^{*T}\boldsymbol{z} \tag{5.36}$$

e

$$V^* = \boldsymbol{b}^{*T}\boldsymbol{w} \tag{5.37}$$

onde os vetores $\boldsymbol{a}^*$ e $\boldsymbol{b}^*$ são os vetores de transformação linear. Já suas matrizes de variância e covariância são, respectivamente

$$Var(U^*) = Var\left[\boldsymbol{a}^{*T}\boldsymbol{z}\right] = \boldsymbol{a}^{*T}Var[\boldsymbol{z}]\boldsymbol{a}^* = \boldsymbol{a}^{*T}\boldsymbol{\rho_x}\boldsymbol{a}^*, \tag{5.38}$$

$$Var(V^*) = Var\left[\boldsymbol{b}^{*T}\boldsymbol{w}\right] = \boldsymbol{b}^{*T}Var[\boldsymbol{w}]\boldsymbol{b}^* = \boldsymbol{b}^{*T}\boldsymbol{\rho_y}\boldsymbol{b}^* \tag{5.39}$$

e

$$Cov[U^*, V^*] = Cov\left[\boldsymbol{a}^{*T}\boldsymbol{z}, \boldsymbol{b}^{*T}\boldsymbol{w}\right] = \boldsymbol{a}^{*T}Cov[\boldsymbol{z}, \boldsymbol{w}]\boldsymbol{b}^* = \boldsymbol{a}^{*T}\boldsymbol{\rho_{xy}}\boldsymbol{b}^* \tag{5.40}$$

onde $\boldsymbol{\rho_x}$ é a matriz de correlação de $\boldsymbol{x}$, $\boldsymbol{\rho_y}$ de $\boldsymbol{y}$ e $\boldsymbol{\rho_{xy}}$ entre $\boldsymbol{x}$ e $\boldsymbol{y}$ (FERREIRA, 2008).

Da mesma maneira, o método consiste em determinar os vetores $\boldsymbol{a}^*$ e $\boldsymbol{b}^*$ que maximizam a correlação entre U^*e V^*, que, conforme a equação (5.3), pode ser dada por

$$\rho_{U^*,V^*} = \frac{\boldsymbol{a}^{*T}\boldsymbol{\rho_{xy}}\boldsymbol{b}^*}{\sqrt{\boldsymbol{a}^{*T}\boldsymbol{\rho_x}\boldsymbol{a}^*}\sqrt{\boldsymbol{b}^{*T}\boldsymbol{\rho_y}\boldsymbol{b}^*}} \tag{5.41}$$

De forma análoga ao desenvolvimento do tópico anterior, os vetores podem ser determinados resolvendo a equação

$$(\boldsymbol{\rho_y^{-1}}\boldsymbol{\rho_{yx}}\boldsymbol{\rho_x^{-1}}\boldsymbol{\rho_{xy}} - \rho^2_{U^*,V^*}\boldsymbol{I})\boldsymbol{b}^* = \boldsymbol{0}, \tag{5.42}$$

onde $\boldsymbol{\rho_{yx}} = \boldsymbol{\rho_{xy}^T}$, e substituindo $\boldsymbol{b}^*$ em

$$\boldsymbol{a}^* = \frac{1}{\rho_{U^*,V^*}}\boldsymbol{\rho_x^{-1}}\boldsymbol{\rho_{xy}}\boldsymbol{b}^*. \tag{5.43}$$

Ou então, de forma equivalente, pode-se resolver a equação

$$(\boldsymbol{\rho_x^{-1}}\boldsymbol{\rho_{xy}}\boldsymbol{\rho_y^{-1}}\boldsymbol{\rho_{yx}} - \rho^2_{U^*,V^*}\boldsymbol{I})\boldsymbol{a}^* = \boldsymbol{0}, \tag{5.44}$$

e substituir $\boldsymbol{a}^*$ em

$$\boldsymbol{b}^* = \frac{1}{\rho_{U^*,V^*}}\boldsymbol{\rho_y^{-1}}\boldsymbol{\rho_{yx}}\boldsymbol{a}^*. \tag{5.45}$$

Johnson e Wichern (2007) discutem que há uma relação entre os vetores de transformação obtidos pelas variáveis padronizadas e aqueles obtidos pelas variáveis originais. Portanto, é possível calcular as variáveis canônicas dos dados originais realizando a análise com os dados padronizados, e vice e versa. Assim, os vetores de transformação podem ser expressados como

$$\boldsymbol{a}^* = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{a}, \tag{5.46}$$

e

$$\boldsymbol{b}^* = \boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{b}. \tag{5.47}$$

Isso pode ser demonstrado substituindo as equações (5.46, 5.47) na equação (5.41). Desta forma, tem-se que

$$\rho_{U^*,V^*} = \frac{\boldsymbol{a}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\rho_{xy}}\boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\rho_x}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{a}}\sqrt{\boldsymbol{b^T}\boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{\rho_y}\boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{b}}} \tag{5.48}$$

e conforme as equações (5.38, 5.39, 5.40), pode ser observado que

$$\rho_{U^*,V^*} = \frac{\boldsymbol{a}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\rho_{xy}}\boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\rho_x}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{a}}\sqrt{\boldsymbol{b^T}\boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{\rho_y}\boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{b}}} = \frac{\boldsymbol{a}^T\boldsymbol{\Sigma}_{xy}\boldsymbol{b}}{\sqrt{(\boldsymbol{a}^T\boldsymbol{\Sigma}_x\boldsymbol{a})(\boldsymbol{b}^T\boldsymbol{\Sigma}_y\boldsymbol{b})}} = \rho_{U,V}. \tag{5.49}$$

Esta é uma propriedade da análise de correlação canônica e que não pode ser observada em outras técnicas como análise de componentes principais, por exemplo. Na prática as i-ésimas obervações padronizadas, que serão utilizadas, são dadas por

$$\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{5.50}$$

e

$$\hat{\boldsymbol{w}}_i = \hat{\boldsymbol{\Delta}}^{-\frac{1}{2}}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}) \tag{5.51}$$

onde os vetores $\bar{\boldsymbol{x}}$ e $\bar{\boldsymbol{y}}$ são os valores médios de cada conjunto de variáveis, que podem ser estimados através da equação (2.19), e $\hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}$ e $\hat{\boldsymbol{\Delta}}^{-\frac{1}{2}}$ são os estimadores de $\boldsymbol{\Lambda}^{-\frac{1}{2}}$ e $\boldsymbol{\Delta}^{-\frac{1}{2}}$, respectivamente. Já as matrizes de correlação dos dados amostrais, serão

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{x}} = \hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}\boldsymbol{S}_x\hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}, \tag{5.52}$$

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{y}} = \hat{\boldsymbol{\Delta}}^{-\frac{1}{2}}\boldsymbol{S}_y\hat{\boldsymbol{\Delta}}^{-\frac{1}{2}} \tag{5.53}$$

e

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{xy}} = \hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}\boldsymbol{S}_{\boldsymbol{xy}}\hat{\boldsymbol{\Delta}}^{-\frac{1}{2}} \tag{5.54}$$

onde $\boldsymbol{S}_x$, $\boldsymbol{S}_y$ e $\boldsymbol{S}_{xy}$ são estimados pela equação (2.31) (MINGOTI, 2005). Com isso, Ferreira (2008) aborda que as equações (5.42, 5.43) e (5.44, 5.45) se tornarão, respectivamente,

$$(\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}^{-1}\hat{\boldsymbol{\rho}}_{\boldsymbol{yx}}\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\rho}}_{\boldsymbol{xy}} - \hat{\rho}^2_{U^*,V^*}\boldsymbol{I})\hat{\boldsymbol{b}}^* = \boldsymbol{0} \tag{5.55}$$

$$\hat{\boldsymbol{a}}^* = \frac{1}{\hat{\rho}_{U^*,V^*}}\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\rho}}_{\boldsymbol{xy}}\boldsymbol{b}^* \tag{5.56}$$

e

$$(\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\rho}}_{\boldsymbol{yx}}\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}^{-1}\hat{\boldsymbol{\rho}}_{\boldsymbol{yx}} - \hat{\rho}^2_{U^*,V^*}\boldsymbol{I})\hat{\boldsymbol{a}}^* = \boldsymbol{0} \tag{5.57}$$

$$\hat{\boldsymbol{b}}^* = \frac{1}{\hat{\rho}_{U^*,V^*}}\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}^{-1}\hat{\boldsymbol{\rho}}_{\boldsymbol{yx}}\boldsymbol{a}^* \tag{5.58}$$

onde os vetores $\hat{\boldsymbol{a}}^*$ e $\hat{\boldsymbol{b}}^*$ são os vetores de transformação das variáveis padronizadas. Finalmente as variáveis canônicas que serão estimadas, a partir dos dados padronizados, são dadas por

$$\hat{U}^* = \hat{\boldsymbol{a}}^{*T}\hat{\boldsymbol{z}} \tag{5.59}$$

e

$$\hat{V}^* = \hat{\boldsymbol{b}}^{*T}\hat{\boldsymbol{w}}. \tag{5.60}$$

5.3 Número de pares de variáveis canônicas e a qualidade de um modelo reduzido

Como visto, o método CCA consiste em estabelecer as transformações lineares dadas pelas equações (5.1, 5.2) encontrando os vetores $\boldsymbol{a}$ e $\boldsymbol{b}$ que maximizam a correlação linear entre as variáveis canônicas U e V. Para isso, foi demonstrado que pode-se calcular $\boldsymbol{a}$ a partir de $\boldsymbol{b}$, este que é o autovetor correspondente ao autovalor, $\rho^2_{U,V}$, da matriz $\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}$, onde $\rho_{U,V}$ é a correlação canônica. Ou então, de forma equivalente,

calcular $\boldsymbol{b}$ a partir de $\boldsymbol{a}$, que é o autovetor correspondente ao autovalor, $\rho^2_{U,V}$, da matriz $\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}$.

Contudo, pode-se observar que a matriz $\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}$ tem dimensões $(q \times q)$ e a matriz $\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}$ tem dimensões $(p \times p)$, o que leva o leitor a pensar que existe q autovetores para a primeira matriz e p autovetores para a segunda. De fato, isso é verdade, entretanto, Ferreira (2008) discute que o número de pares de variáveis será $m = min(p, q)$ pois, se, por exemplo $p < q$, então a primeira matriz terá seus últimos $q - p$ autovalores iguais a zero. Além disto, os primeiros p autovetores da primeira metriz serão idênticos aos p autovetores da segunda.

Uma vez que a análise de correlação canônica fornecerá m pares de variáveis, é possível calcular a correlação entre cada variável canônica e cada variável original, tanto de $\boldsymbol{x}$ quanto de $\boldsymbol{y}$. Para isso, as variáveis canônicas podem ser escritas na forma vetorial por

$$\hat{\boldsymbol{U}} = \begin{pmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_m \end{pmatrix} \tag{5.61}$$

e

$$\hat{\boldsymbol{V}} = \begin{pmatrix} \hat{V}_1 \\ \hat{V}_2 \\ \vdots \\ \hat{V}_m \end{pmatrix}. \tag{5.62}$$

Já os vetores de transformação também podem ser colocados na forma matricial, onde

$$\hat{\boldsymbol{A}} = \begin{pmatrix} \hat{\boldsymbol{a}}_1^T \\ \hat{\boldsymbol{a}}_2^T \\ \vdots \\ \hat{\boldsymbol{a}}_m^T \end{pmatrix} \tag{5.63}$$

e

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{b}}_1^T \\ \hat{\boldsymbol{b}}_2^T \\ \vdots \\ \hat{\boldsymbol{b}}_m^T \end{pmatrix} \tag{5.64}$$

são as matrizes contendo os vetores de transformação das variáveis $\boldsymbol{x}$ e $\boldsymbol{y}$ respectivamente. Desta forma, generalizando as equações (5.1, 5.2) tem-se

$$\hat{\boldsymbol{U}} = \hat{\boldsymbol{A}}\boldsymbol{X} \tag{5.65}$$

$$\hat{\boldsymbol{V}} = \hat{\boldsymbol{B}}\boldsymbol{Y} \tag{5.66}$$

onde $\boldsymbol{X}$ e $\boldsymbol{Y}$ são matrizes contendo, em cada coluna, os vetores de observações $\boldsymbol{x}$ e $\boldsymbol{y}$ respectivamente (FERREIRA, 2008).

Portanto, segundo Johnson e Wichern (2007) a covariância entre o vetor de variáveis canônicas $\hat{\boldsymbol{U}}$ e as variáveis originais $\boldsymbol{X}$ será

$$Cov(\hat{\boldsymbol{U}}, \boldsymbol{X}) = Cov(\hat{\boldsymbol{A}}\boldsymbol{X}, \boldsymbol{X}) = \hat{\boldsymbol{A}}Cov(\boldsymbol{X}) = \hat{\boldsymbol{A}}\boldsymbol{S}_x. \tag{5.67}$$

Já a correlação entre $\hat{\boldsymbol{U}}$ e $\boldsymbol{X}$ pode ser feita de maneira análoga a equação (5.54) porém como $Var(\hat{\boldsymbol{U}}) = 1$, a matriz diagonal com suas variâncias será igual a identidade. Com isso, tem-se que

$$\hat{\boldsymbol{\rho}}_{\hat{U},X} = \boldsymbol{I}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\boldsymbol{S}_x\hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}} = \hat{\boldsymbol{A}}\boldsymbol{S}_x\hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}. \tag{5.68}$$

Da mesma maneira, a covariância entre $\hat{\boldsymbol{U}}$ e $\boldsymbol{Y}$ será

$$Cov(\hat{\boldsymbol{U}}, \boldsymbol{Y}) = Cov(\hat{\boldsymbol{A}}\boldsymbol{X}, \boldsymbol{Y}) = \hat{\boldsymbol{A}}Cov(\boldsymbol{X}, \boldsymbol{Y}) = \hat{\boldsymbol{A}}\boldsymbol{S}_{xy}, \tag{5.69}$$

e a correlação

$$\hat{\boldsymbol{\rho}}_{\hat{U},Y} = \boldsymbol{I}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\boldsymbol{S}_{xy}\hat{\boldsymbol{\Delta}}^{-\frac{1}{2}} = \hat{\boldsymbol{A}}\boldsymbol{S}_{xy}\hat{\boldsymbol{\Delta}}^{-\frac{1}{2}}. \tag{5.70}$$

Analogamente, para a variável canônica $\hat{\boldsymbol{V}}$, chega-se em

$$Cov(\hat{\boldsymbol{V}}, \boldsymbol{Y}) = \hat{\boldsymbol{B}}\boldsymbol{S}_y, \tag{5.71}$$

$$\hat{\boldsymbol{\rho}}_{\hat{V},Y} = \hat{\boldsymbol{B}}\boldsymbol{S}_y\hat{\boldsymbol{\Delta}}^{-\frac{1}{2}}, \tag{5.72}$$

$$Cov(\hat{\boldsymbol{V}}, \boldsymbol{X}) = \hat{\boldsymbol{B}}\boldsymbol{S}_{yx} \tag{5.73}$$

e

$$\hat{\boldsymbol{\rho}}_{\hat{V},X} = \hat{\boldsymbol{B}}\boldsymbol{S}_{yx}\hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}. \tag{5.74}$$

Seguindo o mesmo raciocínio para as correlações com as variáveis padronizadas pode ser demonstrado que

$$\hat{\boldsymbol{\rho}}_{\hat{U}^*,Z} = Cov(\hat{\boldsymbol{U}}^*, \boldsymbol{Z}) = \hat{\boldsymbol{A}}^*\hat{\boldsymbol{\rho}}_x, \tag{5.75}$$

$$\hat{\boldsymbol{\rho}}_{\hat{U}^*,W} = Cov(\hat{\boldsymbol{U}}^*, \boldsymbol{W}) = \hat{\boldsymbol{A}}^*\hat{\boldsymbol{\rho}}_{xy}, \tag{5.76}$$

$$\hat{\boldsymbol{\rho}}_{\hat{V}^*,W} = Cov(\hat{\boldsymbol{V}}^*, \boldsymbol{W}) = \hat{\boldsymbol{B}}^*\hat{\boldsymbol{\rho}}_y \tag{5.77}$$

e, finalmente,

$$\hat{\boldsymbol{\rho}}_{\hat{V}^*,Z} = Cov(\hat{\boldsymbol{V}}^*, \boldsymbol{Z}) = \hat{\boldsymbol{B}}^*\hat{\boldsymbol{\rho}}_{yx}. \tag{5.78}$$

Tanto Mingoti (2005) quanto Johnson e Wichern (2007) discutem que essas correlações entre as variáveis canônicas e as variáveis originais, também denominadas por

canonical loadings, podem auxiliar na interpretação e análise da qualidade do modelo ajustado. Uma vez que as variáveis canônicas apresentem alta correlação com as variáveis originais $\boldsymbol{x}$ e $\boldsymbol{y}$, é possível, através de regressões lineares simples, estimar os valores de $\boldsymbol{y}$ tendo $\boldsymbol{x}$, ou então de $\boldsymbol{x}$ conhecendo $\boldsymbol{y}$.

Também é possível reduzir a dimensionalidade do problema escolhendo um número de par de variáveis canônicas $k \leq m$. Contudo, é necessário avaliar a qualidade do modelo reduzido e para isso pode-se utilizar a proporção explicativa de cada conjunto de variáveis comparada a varição total do grupo. Portanto, essa porporção pode ser dada para o conjunto de variáveis padronizadas $\boldsymbol{z}$ e $\boldsymbol{w}$, respectivamente, por

$$R^2_{(z)k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{p}\hat{\rho}^2_{\hat{U}_j,x_i}}{p} \tag{5.79}$$

e

$$R^2_{(w)k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{q}\hat{\rho}^2_{\hat{V}_j,y_i}}{q}. \tag{5.80}$$

Desta forma, pode-se mensurar o quanto o modelo reduzido pode explicar a variabilidade total dos dados originais em uma escala de 0 á 100%, uma vez que $0 \leq R^2 \leq 1$. Há também outras medidas que podem auxiliar na avaliação da qualidade do modelo reduzido que não foram abordadas neste trabalho. Dentre elas, pode-se destacar a matriz de resíduos, onde se obtem as matrizes de covariâncias das variáveis originais, a partir dos k pares de variáveis canônicas, e calcula-se a diferença com relação as matrizes de covariâncias originais, de modo que quanto mais próximo de zero for o desvio dos resultados melhor é a qualidade do modelo reduzido (FERREIRA, 2008).

6 DESENVOLVIMENTO DAS NOVAS FORMULAÇÕES

Até o momento, foi abordado neste trabalho uma revisão de ferramentas matemáticas e como são estruturadas as análises discriminante linear de Fisher, de componentes principais e de correlação canônica. Agora, neste capítulo, será descrito uma nova metodologia que altera cada uma dessas análises de forma a incorporar os efeitos das incertezas experimentais dos dados originais. Como pôde ser observado, a abordagem clássica destas técnicas estatísticas não levam em consideração as incertezas experimentais dos dados e como consequência disto, cada observação representa um ponto no espaço multidimensional em que está inserido, onde cada dimensão é uma variável do modelo. Mesmo que a distância de Mahalanobis, no caso da análise discriminante, seja relativa com as variâncias e covariâncias comuns entre os grupos, estes desvios são apenas estatísticos e não levam em consideração os erros instrumentais no momento da coleta dos dados, e o mesmo ocorre para as outras duas ferramentas.

Como um dos objetivos desta pesquisa é inserir as incertezas experimentais de cada valor observado para as multiplas variáveis, em cada um dos três métodos, é importante destacar que a incerteza experimental de uma observação é a incerteza total representada pela soma quadrática das incertezas estatísticas e instrumentais no momento da coleta. Assim, em uma possível representação gráfica as observações multidimensionais não seriam mais pontos no espaço, mas sim nuvens onde o valor verdadeiro teria uma determinada probabilidade de estar em agum lugar dentro desta nuvem. Logo, as observações, independente do método de análise, terão um vetor com os valores das variáveis e um vetor com suas respectivas incertezas dados por

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \tag{6.1}$$

e

$$\boldsymbol{\sigma}_i = \begin{pmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \vdots \\ \sigma_{ip} \end{pmatrix}, \tag{6.2}$$

onde i representa uma observação e p o número de variáveis para cada observação. No caso da análise discriminante, $\boldsymbol{x}_i$ representará um índivíduo que pertence ao grupo j que contém n_j indivíduos. Já para a análise de componentes principais, esse vetor será uma das n observações, e o mesmo vale para a análise de correlação canônica. Porém, nesta última, haverá um grupo de observações com p e outro com q variáveis.

É possível observar que os três métodos têm em comum o fato de utilizarem a matriz de covariância dos dados para obter seus resultados. Como já visto, a matriz de covariância depende diretamente do vetor médio das observações e Vuolo (1996) discute que para um conjunto de n valores de uma mesma medida, cada qual com sua incerteza, o método da máxima verossimilhança garante que a melhor estimativa para esta medida é a média ponderada dos dados onde os pesos são o inverso das variâncias.

Portanto, o método escolhido para inserir as incertezas experimentais nos cálculos foi construir as matrizes de covariância, em cada uma das três análises, utilizando a melhor estimativa do vetor médio, onde cada componente é dada por

$$\bar{x}_a = \frac{\sum_{i=1}^{n} \frac{x_{ia}}{\sigma_{ia}^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_{ia}^2}} \tag{6.3}$$

e sua variância é

$$\sigma_{\bar{x}_a}^2 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_{ia}^2}}. \tag{6.4}$$

Portanto, o vetor médio e sua variância serão

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \tag{6.5}$$

e

$$\boldsymbol{\sigma}_{\bar{\boldsymbol{x}}}^2 = \begin{pmatrix} \sigma_{\bar{x}_1}^2 \\ \sigma_{\bar{x}_2}^2 \\ \vdots \\ \sigma_{\bar{x}_p}^2 \end{pmatrix}, \tag{6.6}$$

onde as incertezas das coordendas do vetor médio serão a raíz quadrada positiva dos elementos de $\boldsymbol{\sigma}_{x_j}^2$. Este mesmo procedimento é realizado, na análise discriminante, para construir a matriz $\hat{\boldsymbol{B}} = \sum_{j=1}^{k}(x_j - x_G)(x_j - x_G)^T$, onde x_G é a média dos centroides que neste caso será a média ponderada pelas incertezas de cada centroide.

Outro objetivo deste trabalho é avaliar o impacto no resultado das três análises ao inserir as incertezas como descrito acima. Desta forma torna-se necessário propagar os erros das componentes principais, dos *scores* discriminantes e das variáveis canônicas. Para fazer esta propagação, foi utilizado neste trabalho um procedimento numérico com o auxílio de algoritmos desenvolvidos pelo autor, utilizando o software de análise de dados ROOT versão 5.34.36, construído por pesquisadores da Organização Europeia para a Pesquisa Nuclear (CERN). Para mais detalhes, pode-se consultar a referência (ANTCHEVA et al., 2011). O método utilizado pelos algoritmos foi o mesmo para as três análises.

Como já observado nos capítulos anteriores, as três ferramentas têm em comum o fato de que seus resultados dependem diretamente de se extrair autovalores e autovetores

(vetores da transformação linear) de determinadas matrizes. Como estas matrizes têm uma ligação com a matriz de covariância dos dados, $\boldsymbol{S} = \frac{1}{(N-1)} \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$, a metodologia aplicada à cada algoritmo, consiste em realizar um procedimento iterativo (um milhão de iterações)[1] onde, em cada iteração, gera-se um vetor médio aleatório com distribuição gaussiana, em que a média e desvio padrão de cada componente são, respectivamente, as componentes da média ponderada e suas incertezas. Com isso, calcula-se uma matriz de covariância aleatória e consequentemente, uma matriz aleatória de cada análise.

A partir do vetor médio original, calculam-se as matrizes originais, os vetores de transformação linear e os resultados de cada análise. Isso também é feito em cada iteração do algoritmo de modo que o procedimento consiste em armazenar estes resultados (autovalores, autovetores e resultados) em histogramas. Os histogramas gerados pelo ROOT fornecem, automatiacamente, o valor médio das entradas e o rms (*root mean square*). O rms foi utilizado como estimativa das incertezas dos resultados de cada análise.

Outro aspecto comum entre as três análises é que ao se utilizar o ROOT para calcular os autovalores de uma matriz e alimentar um vetor, o procedimento é realizado colocando os autovalores de forma decrescente. Como neste trabalho a metodologia consiste em gerar matrizes aleatórias e extrair seus autovalores e autovetores, estes se tornam variáveis aleatórias. Portanto, se suas distribuições estiverem significativamente sobrepostas em qualquer ponto, pode ocorrer de uma amostra referente ao maior autovalor ser menor que uma referente ao segundo maior, por exemplo. Isso faria com que os valores que deveriam ser utilizados nos cálculos referentes ao maior autovalor pudessem, na verdade, estarem sendo utilizados nos cálculos correspondentes ao segundo maior.

Uma vez que é intrínseco ao ROOT ordenar os autovalores de forma descrescente, não é possível identinficar as amostras que foram comutadas. Deste modo, para que este método seja válido, é fundamental que as distribuições dos autovalores não se sobreponham de forma significativa, pois desta forma a estimativa das incertezas desejadas estará sendo representativa aos dados originais. Neste capítulo será demonstrada uma aplicação em dados reais para a análise de componentes principais e uma simulação de possíveis aplicações para as análises discriminante e de correlação canônica.

6.1 Análise de Componentes Principais

Neste tópico, o método desenvolvido será demonstrado através de uma aplicação real. Como abordado no capítulo 3, a análise de componentes principais consiste em calcular os autovalores e autovetores da matriz de covariância, $\boldsymbol{S}$, ou de correlação, $\hat{\boldsymbol{\rho}}$, dos dados originais, caso a análise seja aplicada a dados não padronizados ou padronizados,

[1] Apenas para a análise de correlação canônica foi utilizado um procedimento com 100 mil iterações devido ao tempo necessário para o recurso computacional utilizado.

respectivamente. Ensor et al. (2017) aplicaram a PCA a um conjunto de 30 observações, cada uma com 23 variáveis com o objetivo de reduzir dimensionalidade e determinar quantos parâmetros eram necessários para explicar a variação do espectro das conhecidas *diffuse interstellar bands* (DIBs), com base nas 23 variáveis. Seguem os dados nas Figuras (6, 7, 8).

Figura 6 – Tabela de dados com oito variáveis de comprimento de onda (todas em $10^{-13}m$), onde cada linha representa uma observação (estrela) e cada coluna uma variável.

Target	λ4428	λ4964	λ5494	λ5513	λ5545	λ5546	λ5769	λ5780
HD 15137	$1163 \pm {}^{106}_{115}$	7.9 ± 2.5	11.1 ± 2.2	2.1 ± 3.0	6.9 ± 1.9	0.0 ± 1.9	3.9 ± 1.7	230.1 ± 9.1
HD 22951	$471 \pm {}^{68}_{76}$	6.4 ± 1.1	2.0 ± 1.1	1.3 ± 1.5	6.2 ± 0.7	3.6 ± 1.0	0.7 ± 0.8	102.8 ± 3.6
HD 23180	$403 \pm {}^{45}_{47}$	12.3 ± 1.4	6.4 ± 0.2	10.7 ± 1.7	10.3 ± 1.5	5.4 ± 1.5	7.2 ± 1.3	88.1 ± 5.0
HD 23630	$325 \pm {}^{48}_{39}$	1.2 ± 1.0	2.4 ± 0.9	0.2 ± 1.5	0.8 ± 1.1	1.5 ± 1.0	2.3 ± 0.9	40.7 ± 4.8
HD 24398	$450 \pm {}^{61}_{70}$	8.8 ± 0.9	5.4 ± 1.0	5.8 ± 1.1	6.1 ± 0.6	3.3 ± 1.0	2.5 ± 0.7	100.4 ± 2.7
HD 24534	$402 \pm {}^{49}_{55}$	13.4 ± 1.6	7.6 ± 1.2	5.3 ± 1.9	9.4 ± 1.2	4.8 ± 1.6	7.1 ± 1.1	95.1 ± 5.0
HD 24760	$322 \pm {}^{41}_{30}$	1.5 ± 0.8	3.3 ± 0.8	1.1 ± 1.0	1.5 ± 0.9	0.2 ± 0.8	1.6 ± 0.6	77.0 ± 3.4
HD 24912	$949 \pm {}^{89}_{65}$	9.7 ± 1.3	7.0 ± 1.0	2.7 ± 1.2	8.9 ± 1.0	2.4 ± 1.2	2.4 ± 0.8	198.3 ± 3.1
HD 27778	$490 \pm {}^{74}_{58}$	8.3 ± 1.4	4.6 ± 1.6	3.6 ± 1.4	8.0 ± 1.1	4.5 ± 1.0	2.2 ± 1.0	86.6 ± 4.6
HD 35149	$254 \pm {}^{43}_{38}$	2.8 ± 1.3	2.8 ± 1.7	1.0 ± 1.9	2.6 ± 1.3	0.0 ± 1.4	1.7 ± 1.5	58.0 ± 5.5
HD 35715	$221 \pm {}^{47}_{23}$	1.3 ± 0.8	1.1 ± 0.8	1.2 ± 0.9	1.1 ± 0.8	0.7 ± 0.9	0.7 ± 0.7	34.6 ± 3.6
HD 36822	$483 \pm {}^{78}_{69}$	1.6 ± 2.4	1.4 ± 2.8	2.9 ± 3.0	2.0 ± 2.4	2.9 ± 2.4	1.0 ± 2.0	84.5 ± 9.6
HD 36861	$402 \pm {}^{49}_{86}$	4.6 ± 1.0	3.2 ± 1.0	4.4 ± 1.1	3.2 ± 0.9	3.2 ± 0.9	1.5 ± 0.7	49.0 ± 3.5
HD 40111	$739 \pm {}^{109}_{81}$	2.2 ± 4.7	2.7 ± 4.9	0.0 ± 7.2	3.6 ± 4.4	3.2 ± 4.9	3.3 ± 3.7	157.7 ± 19.5
HD 110432	$880 \pm {}^{64}_{45}$	8.3 ± 1.0	4.1 ± 1.0	3.8 ± 1.4	5.2 ± 1.0	1.8 ± 0.8	0.3 ± 0.8	137.3 ± 3.7
HD 143275	$383 \pm {}^{21}_{12}$	2.1 ± 1.0	5.1 ± 0.1	2.1 ± 1.5	5.2 ± 1.1	1.4 ± 1.2	1.9 ± 1.1	92.7 ± 4.2
HD 144217	$430 \pm {}^{54}_{38}$	3.5 ± 0.8	2.6 ± 1.0	1.1 ± 1.6	4.1 ± 1.1	1.0 ± 1.0	0.7 ± 1.1	156.0 ± 4.9
HD 145502	$583 \pm {}^{50}_{48}$	3.3 ± 1.2	6.3 ± 2.0	2.8 ± 2.5	4.4 ± 1.0	2.0 ± 1.2	3.0 ± 0.9	186.9 ± 5.2
HD 147165	$872 \pm {}^{50}_{53}$	6.1 ± 1.0	8.2 ± 1.5	5.1 ± 1.6	4.5 ± 1.0	1.9 ± 1.2	0.8 ± 1.1	240.0 ± 4.2
HD 147933	$1254 \pm {}^{121}_{77}$	20.0 ± 0.8	7.6 ± 0.5	13.8 ± 0.7	8.3 ± 0.5	6.9 ± 0.6	11.7 ± 2.8	209.8 ± 16.1
HD 149757	$576 \pm {}^{52}_{47}$	6.6 ± 0.9	5.3 ± 1.1	3.0 ± 1.3	5.7 ± 0.9	2.5 ± 0.8	2.8 ± 1.1	65.1 ± 3.8
HD 164284	$686 \pm {}^{73}_{53}$	2.5 ± 1.3	1.8 ± 1.4	2.3 ± 1.8	3.4 ± 1.1	1.5 ± 1.4	0.7 ± 1.0	94.4 ± 4.4
HD 170740	$834 \pm {}^{107}_{91}$	10.5 ± 1.0	10.6 ± 1.0	8.6 ± 1.5	11.3 ± 1.0	4.6 ± 1.0	2.4 ± 0.8	240.3 ± 4.0
HD 198478	$1592 \pm {}^{191}_{108}$	14.2 ± 2.0	10.5 ± 1.8	5.8 ± 2.1	11.8 ± 1.4	5.0 ± 1.5	1.5 ± 1.3	315.6 ± 5.8
HD 202904	$541 \pm {}^{92}_{67}$	2.5 ± 1.5	2.6 ± 1.5	1.8 ± 1.6	1.0 ± 1.0	1.1 ± 1.2	1.0 ± 1.1	44.5 ± 4.6
HD 207198	$1282 \pm {}^{67}_{89}$	24.6 ± 1.0	19.3 ± 0.9	16.6 ± 1.1	20.5 ± 0.9	9.9 ± 0.9	9.8 ± 0.7	249.0 ± 2.8
HD 209975	$1032 \pm {}^{182}_{74}$	8.8 ± 1.4	13.0 ± 1.5	1.9 ± 1.9	11.2 ± 0.7	4.6 ± 1.6	0.4 ± 1.3	234.2 ± 4.7
HD 214680	$361 \pm {}^{55}_{64}$	0.9 ± 1.0	4.4 ± 0.8	1.8 ± 1.4	1.7 ± 1.0	2.0 ± 0.9	0.6 ± 0.5	58.8 ± 2.8
HD 214993	$232 \pm {}^{63}_{47}$	4.0 ± 0.7	0.2 ± 1.2	1.6 ± 1.3	1.4 ± 1.0	1.2 ± 0.9	0.6 ± 0.8	78.6 ± 4.8
HD 218376	$766 \pm {}^{64}_{108}$	5.1 ± 1.0	5.9 ± 1.1	3.9 ± 1.2	6.2 ± 0.8	1.1 ± 1.1	1.1 ± 0.8	138.7 ± 4.4

Fonte: Adaptado de Ensor et al. (2017).

Figura 7 – Tabela de dados com oito variáveis de comprimento de onda (todas em $10^{-13}m$), onde cada linha representa uma observação (estrela) e cada coluna uma variável.

Target	λ5797	λ5850	λ6196	λ6270	λ6284	λ6376	λ6379	λ6614
HD 15137	68.1 ± 3.1	20.8 ± 2.8	19.9 ± 2.7	33.4 ± 4.6	298.6 ± 19.4	12.7 ± 3.4	36.2 ± 4.2	80.6 ± 4.1
HD 22951	35.9 ± 1.3	18.8 ± 1.2	10.5 ± 2.2	9.0 ± 2.9	130.8 ± 8.5	5.1 ± 1.3	23.8 ± 1.5	41.0 ± 2.1
HD 23180	57.7 ± 2.0	27.8 ± 1.3	12.8 ± 1.9	18.0 ± 3.3	95.4 ± 9.4	10.5 ± 2.1	41.3 ± 3.0	53.7 ± 3.4
HD 23630	6.7 ± 1.3	1.5 ± 1.0	1.9 ± 1.3	3.8 ± 3.3	21.0 ± 7.7	2.0 ± 2.0	3.0 ± 2.1	8.9 ± 2.8
HD 24398	55.5 ± 1.3	27.3 ± 1.1	15.2 ± 1.2	11.0 ± 2.5	94.1 ± 6.7	12.2 ± 1.8	46.3 ± 2.5	59.3 ± 1.9
HD 24534	58.9 ± 1.3	29.0 ± 1.4	15.2 ± 1.4	18.8 ± 3.5	78.2 ± 8.2	10.5 ± 3.7	40.3 ± 2.3	66.1 ± 2.4
HD 24760	13.5 ± 1.0	2.9 ± 0.8	6.0 ± 1.2	11.6 ± 2.0	105.9 ± 5.5	0.3 ± 1.3	8.2 ± 1.5	23.3 ± 2.1
HD 24912	51.4 ± 1.2	22.3 ± 1.7	21.7 ± 1.0	33.0 ± 1.7	272.4 ± 9.6	13.0 ± 1.9	30.1 ± 2.3	79.7 ± 1.8
HD 27778	37.4 ± 2.0	12.7 ± 1.3	10.8 ± 1.5	6.9 ± 3.2	117.8 ± 10.2	8.0 ± 1.8	17.4 ± 2.1	45.7 ± 2.7
HD 35149	11.8 ± 2.1	6.8 ± 1.3	7.1 ± 1.9	12.4 ± 3.7	78.0 ± 14.4	0.9 ± 2.4	6.0 ± 3.3	21.9 ± 4.6
HD 35715	3.3 ± 1.2	0.5 ± 0.7	2.4 ± 1.1	4.0 ± 2.0	55.4 ± 8.4	0.7 ± 2.0	2.8 ± 1.9	9.5 ± 1.9
HD 36822	16.4 ± 3.1	3.7 ± 2.2	8.1 ± 3.1	9.5 ± 8.8	106.6 ± 15.9	3.5 ± 3.3	10.1 ± 5.4	18.0 ± 6.2
HD 36861	23.3 ± 1.2	12.3 ± 0.8	4.9 ± 1.0	4.8 ± 2.1	51.6 ± 10.8	4.7 ± 1.8	6.2 ± 1.4	14.9 ± 1.8
HD 40111	32.3 ± 5.3	3.6 ± 3.1	13.0 ± 5.6	17.1 ± 1..0	211.1 ± 22.5	8.0 ± 7.6	12.9 ± 9.5	41.1 ± 9.6
HD 110432	35.0 ± 1.7	19.4 ± 1.0	18.0 ± 1.0	29.6 ± 2.0	185.1 ± 5.1	7.0 ± 1.8	32.4 ± 1.8	74.3 ± 2.1
HD 143275	17.4 ± 1.3	6.3 ± 1.1	7.6 ± 0.9	10.0 ± 3.8	118.9 ± 13.1	4.3 ± 1.8	10.1 ± 3.0	23.9 ± 1.6
HD 144217	17.3 ± 1.6	6.5 ± 1.1	13.5 ± 1.5	25.0 ± 2.3	159.3 ± 9.1	5.0 ± 2.4	14.0 ± 3.6	50.9 ± 1.7
HD 145502	33.7 ± 1.7	12.2 ± 1.2	14.1 ± 2.6	20.5 ± 2.5	199.6 ± 8.8	7.8 ± 2.0	30.0 ± 2.0	58.8 ± 2.5
HD 147165	31.3 ± 1.6	16.7 ± 1.1	17.5 ± 1.1	26.4 ± 2.7	214.2 ± 7.7	10.9 ± 2.0	21.1 ± 2.0	61.3 ± 2.3
HD 147933	57.2 ± 5.3	30.6 ± 2.6	17.0 ± 2.7	24.9 ± 5.0	173.8 ± 16.9	15.5 ± 2.8	28.0 ± 3.7	62.5 ± 3.6
HD 149757	32.6 ± 1.6	14.2 ± 1.1	10.3 ± 1.2	16.8 ± 2.9	72.0 ± 6.9	10.9 ± 2.0	16.7 ± 1.9	46.4 ± 2.0
HD 164284	13.8 ± 1.7	0.4 ± 1.3	6.8 ± 1.5	15.7 ± 3.0	111.3 ± 9.2	1.8 ± 2.0	11.3 ± 2.2	26.9 ± 2.7
HD 170740	63.3 ± 1.8	24.6 ± 1.1	26.3 ± 1.2	52.7 ± 2.6	249.6 ± 9.9	20.9 ± 1.6	60.7 ± 1.7	122.4 ± 2.2
HD 198478	75.0 ± 2.2	34.6 ± 1.6	33.1 ± 1.5	53.3 ± 4.2	379.5 ± 11.6	21.2 ± 3.5	46.7 ± 4.1	130.6 ± 3.4
HD 202904	5.7 ± 2.3	1.9 ± 1.7	3.6 ± 1.8	15.2 ± 3.1	82.2 ± 10.6	3.0 ± 2.6	11.7 ± 3.5	18.4 ± 2.7
HD 207198	132.6 ± 1.1	61.1 ± 0.7	32.3 ± 1.0	43.2 ± 1.7	227.2 ± 9.6	30.0 ± 1.8	71.8 ± 2.1	121.8 ± 1.9
HD 209975	71.5 ± 1.4	26.5 ± 1.6	26.9 ± 4.5	43.1 ± 3.1	240.2 ± 10.0	25.5 ± 2.7	45.5 ± 2.6	114.1 ± 3.1
HD 214680	20.1 ± 0.9	3.9 ± 0.9	5.4 ± 1.0	9.2 ± 1.6	68.7 ± 7.9	6.4 ± 1.5	4.5 ± 1.4	16.1 ± 2.0
HD 214993	13.6 ± 1.3	0.9 ± 0.7	7.6 ± 1.4	10.0 ± 2.2	107.1 ± 10.0	4.4 ± 1.7	13.9 ± 1.7	18.0 ± 2.3
HD 218376	38.7 ± 1.3	17.2 ± 1.0	14.2 ± 1.2	31.6 ± 2.3	175.7 ± 10.0	11.2 ± 2.0	37.0 ± 2.2	66.0 ± 2.2

Fonte: Adaptado de Ensor et al. (2017).

Figura 8 – Tabela de dados com seis variáveis, onde cada linha representa uma observação (estrela) e cada coluna uma variável.

Target	E(B-V)	N(H I) $[10^{21}\mathrm{cm}^{-2}]$	$N(H_2)$ $[10^{20}\mathrm{cm}^{-2}]$	$f(H_2)$	$F_\star$	$\frac{W(\lambda5797)}{W(\lambda5780)}$
HD 15137	0.24	$1.29^{+0.57}_{-0.40}$	$1.86^{+0.26}_{-0.12}$	$0.22^{+0.09}_{-0.06}$	0.37±0.09	0.30±0.02
HD 22951	0.19	$1.10^{+0.35}_{-0.32}$	$2.88^{+1.48}_{-0.98}$	$0.35^{+0.27}_{-0.18}$	0.73±0.05	0.35±0.02
HD 23180	0.22	$0.76^{+0.26}_{-0.23}$	$3.98^{+1.64}_{-1.16}$	$0.51^{+0.33}_{-0.24}$	0.84±0.06	0.65±0.04
HD 23630	0.05	$0.22^{+0.10}_{-0.07}$	$0.35^{+0.18}_{-0.12}$	$0.28^{+0.23}_{-0.15}$	0.89±0.10	0.16±0.04
HD 24398	0.27	$0.63^{+0.06}_{-0.07}$	$4.68^{+2.40}_{-1.59}$	$0.59^{+0.46}_{-0.31}$	0.88±0.05	0.55±0.02
HD 24534	0.31	$0.54^{+0.08}_{-0.07}$	$8.32^{+0.80}_{-0.73}$	$0.76^{+0.13}_{-0.11}$	0.90±0.06	0.62±0.04
HD 24760	0.07	$0.25^{+0.05}_{-0.05}$	$0.33^{+0.27}_{-0.15}$	$0.21^{+0.25}_{-0.14}$	0.68±0.04	0.18±0.02
HD 24912	0.26	$1.29^{+0.26}_{-0.24}$	$3.39^{+1.40}_{-0.99}$	$0.35^{+0.21}_{-0.15}$	0.83±0.02	0.26±0.01
HD 27778	0.34	$0.22^{+0.55}_{-0.22}$	$5.25^{+1.06}_{-0.88}$	$0.82^{+0.45}_{-0.27}$	1.19±0.07	0.43±0.03
HD 35149	0.08	$0.43^{+0.12}_{-0.13}$	$0.03^{+0.00}_{-0.03}$	$0.02^{+0.00}_{-0.02}$	0.54±0.11	0.20±0.04
HD 35715	0.03	$0.31^{+0.13}_{-0.13}$	$6\pm2\times10^{-6}$	$4\pm2\times10^{-6}$	0.66±0.11	0.10±0.04
HD 36822	0.07	$0.65^{+0.13}_{-0.12}$	$0.21^{+0.09}_{-0.06}$	$0.06^{+0.04}_{-0.03}$	0.74±0.08	0.19±0.04
HD 36861	0.10	$0.60^{+0.16}_{-0.16}$	$0.13^{+0.08}_{-0.05}$	$0.04^{+0.04}_{-0.02}$	0.57±0.04	0.48±0.04
HD 40111	0.10	$0.79^{+0.16}_{-0.15}$	$0.54^{+0.31}_{-0.20}$	$0.12^{+0.10}_{-0.07}$	0.49±0.04	0.20±0.04
HD 110432	0.39	$0.71^{+0.29}_{-0.21}$	$4.37^{+0.42}_{-0.38}$	$0.55^{+0.13}_{-0.11}$	1.17±0.11	0.25±0.01
HD 143275	0.00	$1.41^{+0.29}_{-0.29}$	$0.26^{+0.15}_{-0.09}$	$0.03^{+0.03}_{-0.02}$	0.90±0.03	0.19±0.02
HD 144217	0.18	$1.23^{+0.12}_{-0.11}$	$0.68^{+0.10}_{-0.09}$	$0.10^{+0.02}_{-0.02}$	0.81±0.02	0.11±0.01
HD 145502	0.20	$1.17^{+0.56}_{-0.59}$	$0.78^{+0.32}_{-0.23}$	$0.12^{+0.08}_{-0.07}$	0.80±0.11	0.18±0.01
HD 147165	0.31	$2.19^{+0.90}_{-0.87}$	$0.62^{+0.25}_{-0.18}$	$0.05^{+0.04}_{-0.03}$	0.76±0.06	0.13±0.01
HD 147933	0.37	$4.27^{+0.98}_{-0.80}$	$3.72^{+1.53}_{-1.09}$	$0.15^{+0.09}_{-0.07}$	1.09±0.08	0.27±0.03
HD 149757	0.29	$0.52^{+0.02}_{-0.04}$	$4.47^{+0.90}_{-0.75}$	$0.63^{+0.20}_{-0.17}$	1.05±0.02	0.50±0.04
HD 164284	0.11	$0.42^{+0.23}_{-0.39}$	$0.71^{+0.29}_{-0.21}$	$0.25^{+0.18}_{-0.20}$	0.89±0.18	0.15±0.02
HD 170740	0.38	$1.07^{+0.59}_{-0.47}$	$7.24^{+1.47}_{-1.22}$	$0.58^{+0.22}_{-0.18}$	1.02±0.11	0.26±0.01
HD 198478	0.43	$2.04^{+0.84}_{-0.63}$	$7.41^{+3.06}_{-2.17}$	$0.42^{+0.27}_{-0.20}$	0.81±0.05	0.24±0.01
HD 202904	0.09	$0.23^{+0.21}_{-0.23}$	$0.14^{+0.07}_{-0.05}$	$0.11^{+0.12}_{-0.10}$	0.39±0.11	0.13±0.05
HD 207198	0.47	$3.39^{+0.59}_{-0.50}$	$6.76^{+0.65}_{-0.59}$	$0.28^{+0.05}_{-0.05}$	0.90±0.03	0.53±0.01
HD 209975	0.27	$1.29^{+0.41}_{-0.38}$	$1.20^{+0.62}_{-0.41}$	$0.16^{+0.12}_{-0.09}$	0.57±0.26	0.31±0.01
HD 214680	0.08	$0.50^{+0.14}_{-0.15}$	$0.17^{+0.05}_{-0.04}$	$0.06^{+0.03}_{-0.03}$	0.50±0.06	0.34±0.02
HD 214993	0.06	$0.58^{+0.20}_{-0.18}$	$0.43^{+0.22}_{-0.14}$	$0.13^{+0.10}_{-0.07}$	0.68±0.10	0.17±0.02
HD 218376	0.16	$0.89^{+0.28}_{-0.26}$	$1.41^{+0.73}_{-0.48}$	$0.24^{+0.19}_{-0.13}$	0.60±0.06	0.28±0.01

Fonte: Adaptado de Ensor et al. (2017).

Um dos pontos levantados pelos autores é que eles testaram a normalidade de

algumas variáveis e observaram que elas não possuíam uma distribuição normal, como pode ser observado nas Figuras 6 e 8, nas variáveis cujas incertezas são assimétricas. Entretanto, afirmaram que assumir a normalidade dos dados não era uma decisão crítica e não afetaria significativamente a análise, portanto, neste trabalho, optou-se também por assumir que os dados são distribuidos normalmente. Contudo, para as variáveis com incertezas assimétricas, foram considerados os desvios de maior valor por questão de conservadorismo.

Outro ponto é que, nas Figuras 6, 7 e 8 há apenas 22 variáveis, porém, segundo Ensor et al. (2017), foi construída uma nova variável como combinação de outras duas, pois, além de discutirem uma aplicação da PCA em um ajuste paralelo, demonstraram que o autovalor correspondente a esta variável era nulo. Assim, a nova variável é determinada por

$$N(H) = N(HI) + 2N(H_2) \tag{6.7}$$

onde $N(H)$ representa o total de hidrogênio, $N(HI)$ a densidade do hidrogênio atômico neutro e $N(H_2)$ a desidade molecular do hidrogênio do meio interestelar na linha de visada da estrela. Através da equação 2.48 tem-se que a variância da variável $N(H)$ será

$$\sigma^2_{N(H)} = \sigma^2_{N(HI)} + 4\sigma^2_{N(H_2)}. \tag{6.8}$$

Para realizar a análise, os autores optaram por fazer a padronização dos dados conforme a equação (3.20). Neste trabalho, será utilizada a mesma equação, contudo, o vetor médio será dado pela média pondera dos dados e $\boldsymbol{\Gamma} = diag(S^{aa})$, onde, neste caso, a matriz $\boldsymbol{S}$ corresponde à matriz de covariância construída a partir da média ponderada. Assim, as variáveis padronizadas são determinadas por

$$\boldsymbol{z}_i = \boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}), \tag{6.9}$$

onde $\bar{\boldsymbol{x}}$ é a média ponderada das observações $\boldsymbol{x}_i$. Para comprovar essa padronização, cada componente do vetor médio das novas variáveis deve ser calculada por

$$\bar{z}^a = \frac{\sum_i^N p_i^a z_i^a}{\sum_{i=1}^N p_i^a}, \tag{6.10}$$

onde

$$p_i^a = \frac{1}{\sigma^2_{x_i^a}}. \tag{6.11}$$

Desta forma, ao substituir a componente a da variável padronizada na equação acima, pode-se mostrar que

$$\bar{z}^a = \frac{\sum_i^N p_i^a \left(\frac{x_i^a - \bar{x}^a}{\sqrt{S^{aa}}}\right)}{\sum_{i=1}^N p_i^a} \tag{6.12}$$

e

$$\bar{z}^a = \frac{1}{\sqrt{S^{aa}}}\left(\frac{\sum_{i=1}^N p_i^a x_i^a}{\sum_{i=1}^N p_i^a} - \frac{\sum_{i=1}^N p_i^a \bar{x}^a}{\sum_{i=1}^N p_i^a}\right) = \frac{1}{\sqrt{S^{aa}}}\left(\bar{x}^a - \bar{x}^a \frac{\sum_{i=1}^N p_i^a}{\sum_{i=1}^N p_i^a}\right) = 0. \tag{6.13}$$

Já a variância será

$$S_z^{aa} = \frac{1}{(N-1)} \sum_{i=1}^{N} (z_i^a - \bar{z}^a)(z_i^a - \bar{z}^a) = \frac{1}{(N-1)} \sum_{i=1}^{N} (z_i^a)^2. \tag{6.14}$$

Novamente, substituindo a componente a da equação (6.9) na equação da variância, acima, tem-se que

$$S_z^{aa} = \frac{1}{(N-1)} \sum_{i=1}^{N} \frac{(x_i^a - \bar{x}^a)^2}{S^{aa}} = \frac{1}{(N-1)S^{aa}} \sum_{i=1}^{N} (x_i^a - \bar{x}^a)^2. \tag{6.15}$$

Uma vez que a variância de x_i^a é

$$S^{aa} = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i^a - \bar{x}^a)^2, \tag{6.16}$$

então a equação (6.15) se torna

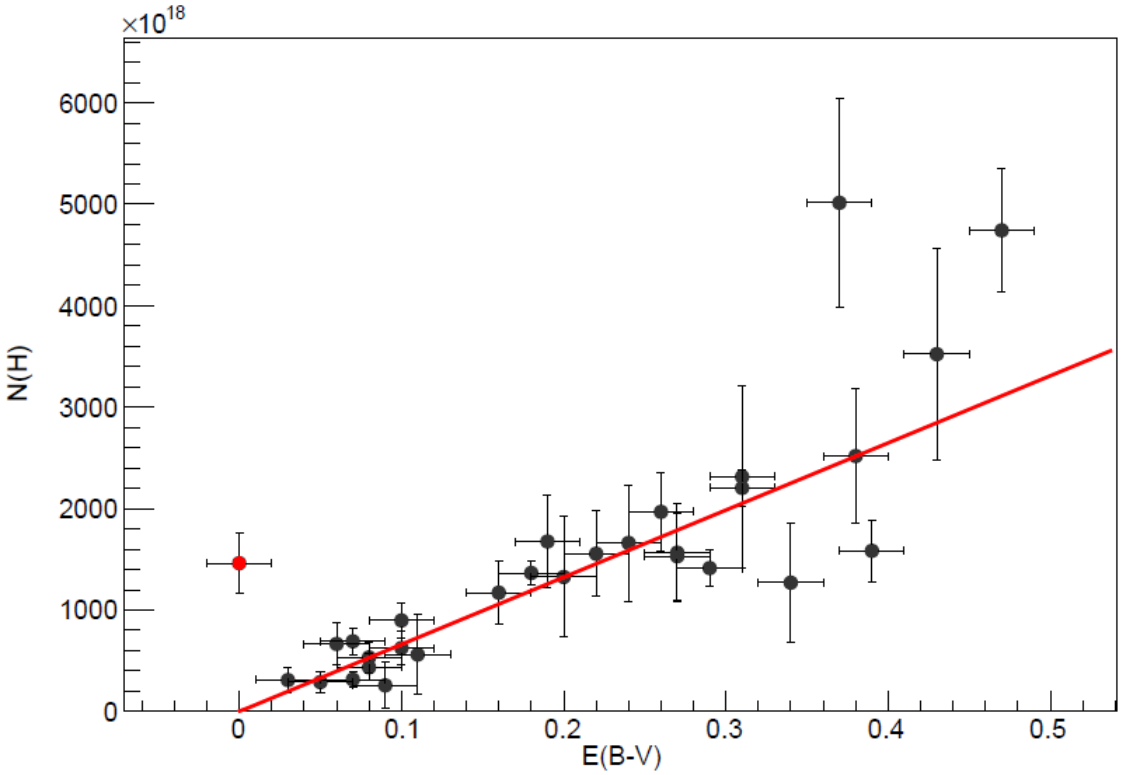
$$S_z^{aa} = \frac{S^{aa}}{S^{aa}} = 1. \tag{6.17}$$

Portanto está demonstrada a padronização utilizando a média ponderada dos dados.

Conforme discutido por Ensor et al. (2017), vários autores já mostraram a existência de correlação entre vários parâmetros, tanto entre si quanto com as DIBs. Como exemplo, eles citam a existência de correlação linear, mostrada por Bohlin, Savage e Drake (1978), entre o excesso de cor $E(B-V)$ e a densidade atômica do hidrogênio $N(H)$, e aplicam a PCA para buscar um ajuste linear entre essas variáveis e comparar com os resultados obtidos por Bohlin, Savage e Drake (1978).

Assim, devido à essa correlação, antes de realizar a análise completa com as 23 variáveis, será demonstrado um exemplo mais simples com as duas variáveis $E(B-V)$ e $N(H)$. Entretanto, inicialmente, foi realizada uma regressão linear ponderada, conforme a Figura 9, para avaliar o comportamento da reta ajustada para os pares $(E(B-V), N(H))$.

Figura 9 – Regressão linear para o conjunto de dados das variáveis $E(B-V)$, em mag, e $N(H)$, em cm^{-2}, onde $p0$ representa o coeficiente angular da reta, $p1$ o coeficiente linear e χ^2/ndf é o Chi-quadrado dividido pelo número de graus de liberdade.

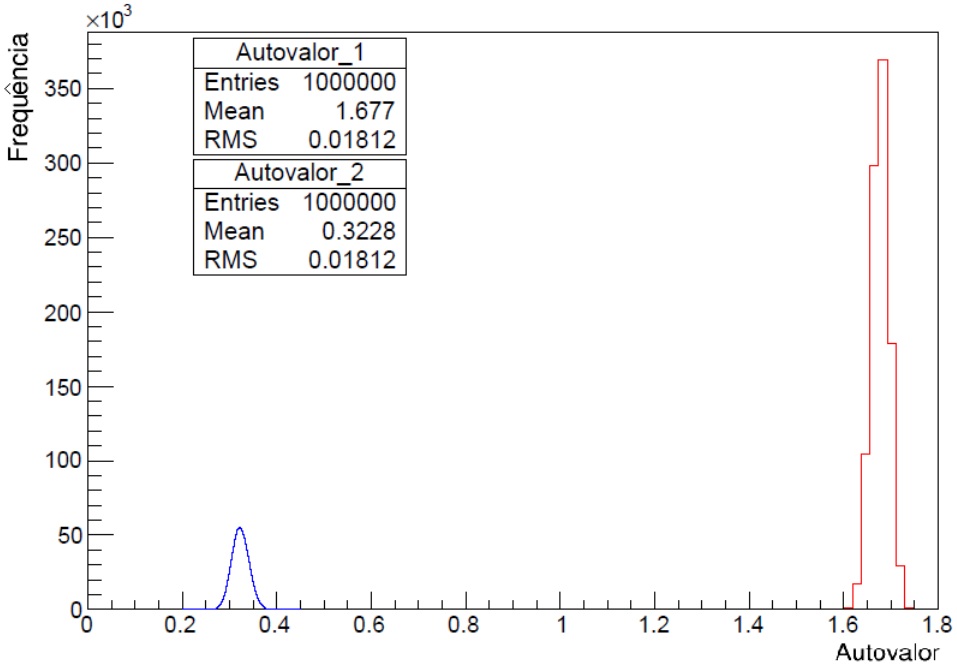


Fonte: Do autor.

Foi observado na Figura 9 que o ponto $(0,0 \pm 0,2 mag; 1,50 \pm 0,29 \cdot 10^{21} cm^{-2})$ (em vermelho), correspondente à estrela $HD143275$ é um possível *outlier*. Para averiguar isso, foi calculada a distância do ponto à reta com relação as suas barras de erro, em cada eixo coordenado. Foi constatado que a distância relativa do ponto à reta no eixo vertical é de $5,2\sigma$ e no eixo horizontal é de $11,5\sigma$. Portanto, dado a mínima probabilidade desta medida cair sobre a função ajustada, este ponto foi considerado um *outlier*, e a aplicação do modelo de PCA desenvolvida neste trabalho foi realizada com as 29 amostras restantes.

Para realizar esta aplicação, foi desenvolvido pelo autor um algoritmo que realiza os cálculos da média ponderada, matriz de covariância e correlação dos dados originais e variáveis padronizadas. A matriz de correlação foi construída conforme a equação (3.21), e em cada iteração era extraído os seus autovalores e autovetores para construir as componentes principais (PC). Ao realizar o procedimento metodológico deste trabalho, pôde-se observar que as distribuições dos autovalores estão completamente distintas (Figura 10), viabilizando, assim, o método proposto.

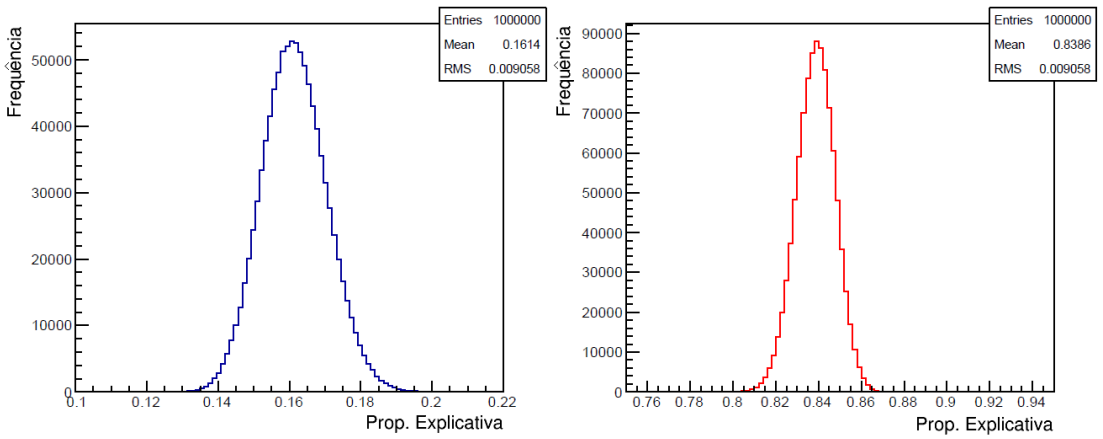
Figura 10 – Histogramas com as distribuições dos autovalores referentes as $PC1$(vermelho) e $PC2$(azul) da PCA aplicada ás variáveis $E(B-V)$ e $N(H)$, onde *Entries*, *Mean* e *RMS* representam o número de dados colocados nos histogramas, a média aritmética deles e a raíz do valor quadrático médio respectivamente.



Fonte: Do autor.

Para avaliar a variabilidade explicada por cada PC, em cada iteração, também foi aplicada a equação (3.22). Assim, a distribuição destas proporções segue demonstrada na Figura 11.

Figura 11 – Histogramas com as distribuições das proporções explicativas da $PC1$(vermelho) e $PC2$(azul).



Fonte: Do autor.

As proporções explicativas acumuladas também foram calculadas. Uma vez que sua equação é dada por

$$\gamma_{Cum_j} = \gamma_j + \gamma_{Cum_{j-1}}, \tag{6.18}$$

ao utilizar a equação (2.48) pode-se calcular sua incerteza por

$$\sigma^2_{\gamma_{Cum_j}} = \sigma^2_{\gamma_j} + \sigma^2_{\gamma_{Cum_{j-1}}}. \tag{6.19}$$

Como agora as proporções acumuladas possuem incertezas, é possível calcular uma incerteza relativa pela equação

$$\sigma_{Re} = \frac{\sigma_{\gamma_{Cum_j}}}{\gamma_{Cum_j}}. \tag{6.20}$$

Assim, os resultados obtidos por este novo método seguem, juntamente com suas incertezas[2], demonstrados na Tabela 2.

Tabela 2 – Resultados da PCA para as variáveis $E(B-V)$ e $N(H)$, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores, as proporções explicativas, as proporções acumuladas, as incertezas relativas do percentual acumulado e os autovetores, respectivamente, juntamente com suas incertezas.

PC	Autovalor (σ)	%Var (σ)	%Cum (σ)	σ_Re	Autovetor (σ)	
1	1,706(19)	85,3(9)	85,3(9)	0,011	0,707107(3)	0,707107(3)
2	0,294(19)	14,6(9)	100,0(1,3)	0,013	- 0,707107(3)	0,707107(3)

Fonte: Do autor.

A fim de comparação, segue, também, os resultados obtidos por Ensor et al. (2017).

Tabela 3 – Resultados da PCA para as variáveis $E(B-V)$ e $N(H)$, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores, as proporções explicativas, as proporções acumuladas e os autovetores.

PC	Autovalor	%Var	%Cum	Autovetor	
1	2	90,63	90,63	0,707	0,707
2	0,187	9,7	100	0,707	-0,707

Fonte: (ENSOR et al., 2017).

Como pôde ser observado, os resultados deste novo modelo apresentaram uma pequena diferença comparados ao trabalho de Ensor et al. (2017). Contudo, é razoável

[2] As incertezas entre parênteses correspondem aos últimos algarismos significativos da medida. Assim, por exemplo, o autovetor 1 é $1,706 \pm 0,019$ e o percentual cumulativo 2 é $100,1 \pm 1,3\%$.

pensar que dados com incertezas de ordens de grandezas maiores resultarão em valores menos precisos. Um fato que deve ser destacado, é que segundo Ensor et al. (2017) as duas componentes principais seriam capazes de replicar 100% da variabilidade do problema. Porém, o novo modelo mostra que isso não é necessariamente verdade. Uma vez que o vínculo dos pesos individuais de cada componente $\sum_j \gamma_j = 1$ deve ser respeitado, então isso significa que uma variação positiva em um dos pesos gera uma variação negativa em outros pesos. Ao final da análise, pode ser que mesmo que se escolha todas as componentes, devido as incertezas experimentais, não seja possível explicar 100% da variabilidade do problema.

Cada componente principal pode, então, ser descrita como uma combinação linear das variáveis, conforme a equação (3.1). Desta forma, as equações das componentes principais ficam

$$PC1 = 0,707107(3)z_{E(B-V)} + 0,707107(3)z_{N(H)} \tag{6.21}$$

$$PC2 = -0,707107(3)z_{E(B-V)} + 0,707107(3)z_{N(H)}, \tag{6.22}$$

onde $z_{E(B-V)}$ e $z_{N(H)}$ são as variáveis padronizadas de $E(B-V)$ e $N(H)$, respectivamente, que são calculadas a partir da equação (6.9). Uma vez que se tem a média ponderada e as variâncias de cada variável, é possível calcular as componentes principais em função das variáveis originais. Neste caso, tem-se que

$$PC1 = 5,590 \pm 0,003E(B-V) + 4,96 \times 10^{-22} \pm 0,07 \times 10^{-22} \pm N(H) - 1,54 \pm 0,03 \tag{6.23}$$

e

$$PC2 = -5,590 \pm 0,003E(B-V) + 4,96 \times 10^{-22} \pm 0,07 \times 10^{-22} N(H) + 0,82 \pm 0,03. \tag{6.24}$$

Ensor et al. (2017) discutem que como a primeira componente principal explica grande parte da variabilidade do problema, pode-se fazer uma aproximação para a segunda componente, de modo que $PC2 = 0$. Neste caso, pode-se utilizar a equação (6.24) para colocar $N(H)$ em função de $E(B-V)$, e a equação (2.48) para propagar a incerteza de cada parâmetro. Fazendo isso, chega-se em

$$N(H) = 11,3 \pm 0,2 \times 10^{21} E(B-V) - 1,65 \pm 0,05 \times 10^{21}. \tag{6.25}$$

Para fazer um comparativo, foi realizada uma regressão linear, construída através do método da máxima verossimilhança. Entretanto, uma vez que os valores da variável $E(B-V)$ possuíam a mesma incerteza, foi realizado um rebatimento das incertezas de $E(B-V)$ para as de $N(H)$, conforme a equação

$$\sigma^2_{N(H)} = \sigma^2_{N(H)_0} + \left(\frac{dN(H)}{dE(B-V)}\right)^2 \sigma^2_{E(B-V)}, \tag{6.26}$$

onde foi utilizado a incerteza inicial do valor de $N(H)$ para $\sigma_{N(H)_0}$ e a inclinação da reta ajustada na Figura 9 para $\left(\frac{dN(H)}{dE(B-V)}\right)$ (VUOLO, 1996). Assim, a função ajustada é dada por

$$N(H) = 6,2 \pm 0,5 \times 10^{21} E(B-V) + 5,27 \pm 7,54 \times 10^{19}, \tag{6.27}$$

em que o coeficiente angular foi menor e o coeficiente linear foi maior do que o ajuste pelas componentes principais. Ao fazer esses cálculos, Ensor et al. (2017) encontram a função linear

$$N(H) = 9,0 \times 10^{21} E(B-V) - 0,3 \times 10^{21}, \tag{6.28}$$

das componentes principais e

$$N(H) = 10,0 \times 10^{21} E(B-V) - 0,5 \times 10^{21} \tag{6.29}$$

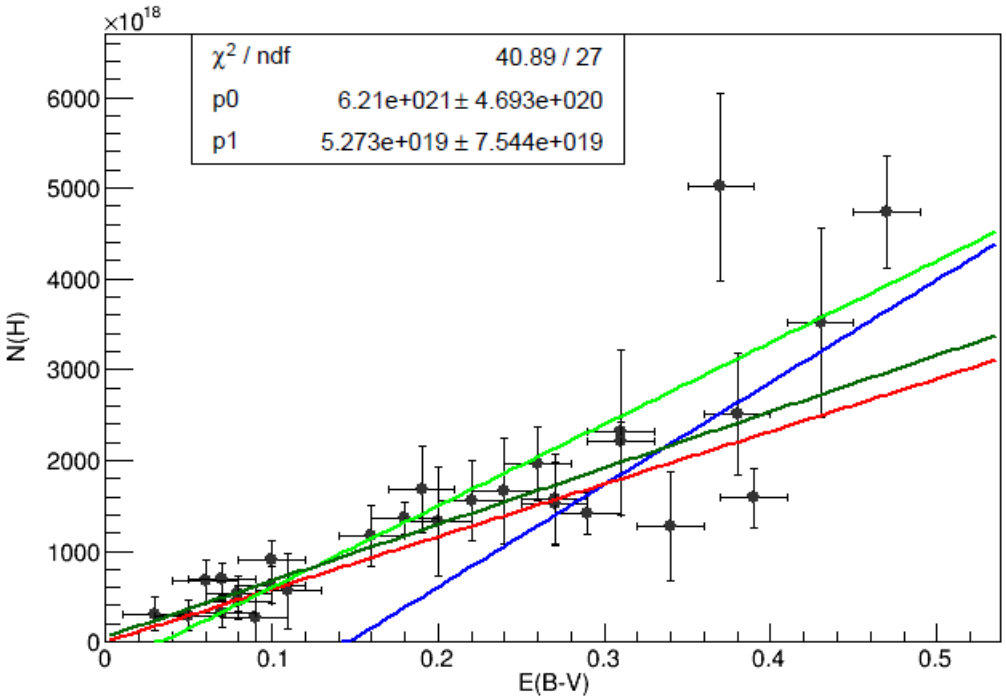
para o ajuste a partir do método dos mínimos quadrados[3]. Ensor et al. (2017) também discutem que (BOHLIN; SAVAGE; DRAKE, 1978) apresentou uma função linear que explicava o comportamento dessas duas variáveis, dada por

$$N(H) = 5,8 \times 10^{21} E(B-V). \tag{6.30}$$

Para uma melhor interpretação, as retas ajustadas pelas componentes principais, método da máxima verossimilhança e pelo Bohlin, Savage e Drake (1978) estão dispostas na Figura 12.

[3]Dadas as diferenças dos parâmetros da função, esse ajuste provavelmente não foi realizado com o método da máxima verossimilhança.

Figura 12 – Ajuste a partir da equação (6.25) com $PC2 = 0$ (reta azul), ajuste obtido por Ensor et al. (2017) pela equação (6.28) (reta verde claro), regressão linear (reta verde escuro), dada pela equação (6.27), em que $p0$ e $p1$ são os coeficientes angular e linear, respectivamente, e função proposta por Bohlin, Savage e Drake (1978) dada pela equação (6.30), (reta vermelha), onde $[N(H)] = [cm^{-2}]$ e $[E(B-V)] = [mag]$.

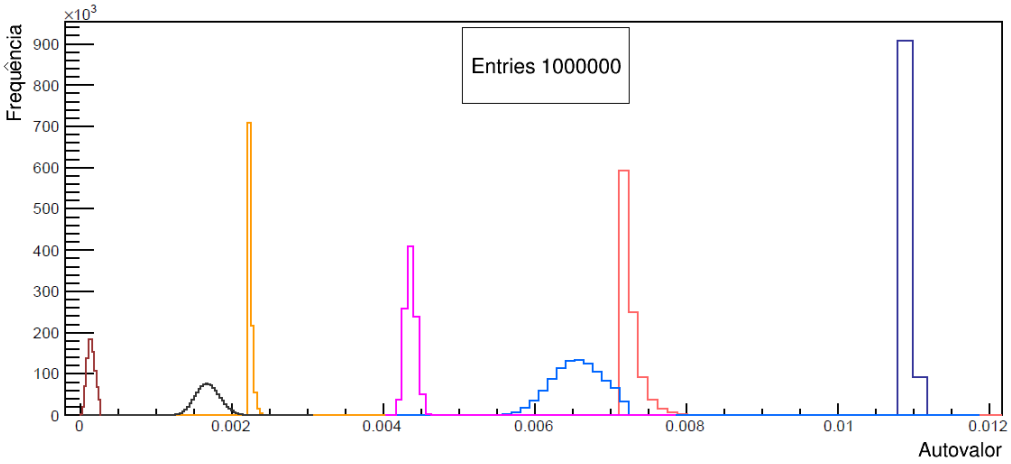


Fonte: Do autor.

Como pode ser observado na Figura 12, a regressão linear ponderada pelas incertezas dos pontos se assemelhou mais a função proposta por Bohlin, Savage e Drake (1978) do que a reta ajustada pelas componentes principais, pois mesmo sem as informações das incertezas experimentais, no caso de Ensor et al. (2017), quanto neste caso levando-as em conta, vale ressaltar que a PCA não tem como objetivo ajustar funções mas, sim, reduzir a dimensionalidade do problema. Assim, ao aplicar os autovetores obtidos nas variáveis padronizadas pode-se calcular os valores das componentes principais de cada observação que estão dispostos no Apêndice A.

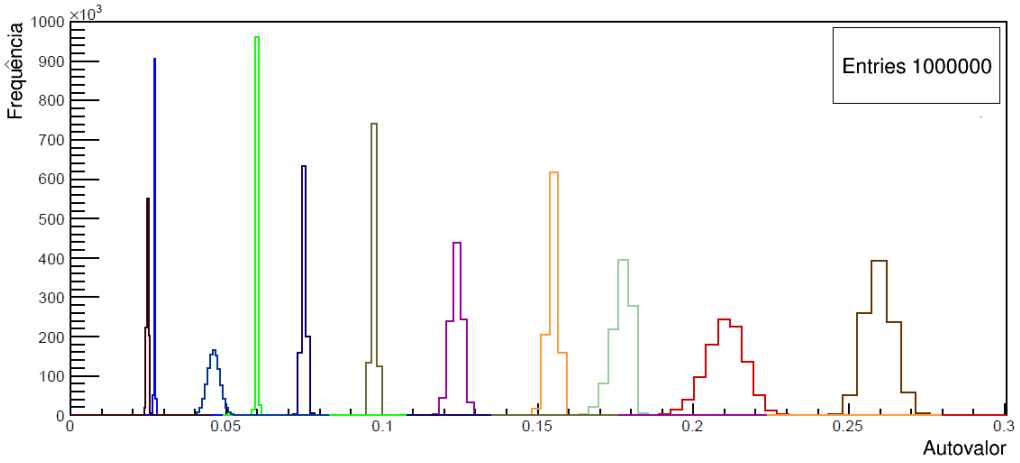
Já para a análise contemplando as 23 variáveis, o mesmo procedimento foi utilizado. Portanto, para fazer a mesma avaliação da distribuição de cada autovalor, seguem as distribuições nas Figuras 13-16 , em ordem crescente de valor.

Figura 13 – Da esquerda para a direita estão as distribuições de frequências do 23º ao 17º autovalor.



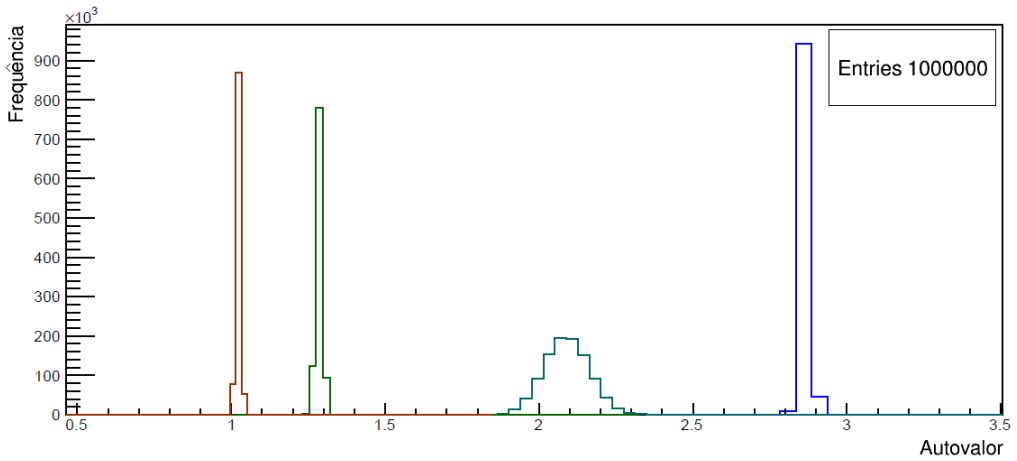
Fonte: Do autor.

Figura 14 – Da esquerda para a direita estão as distribuições de frequência do 16º ao 6º autovalor.



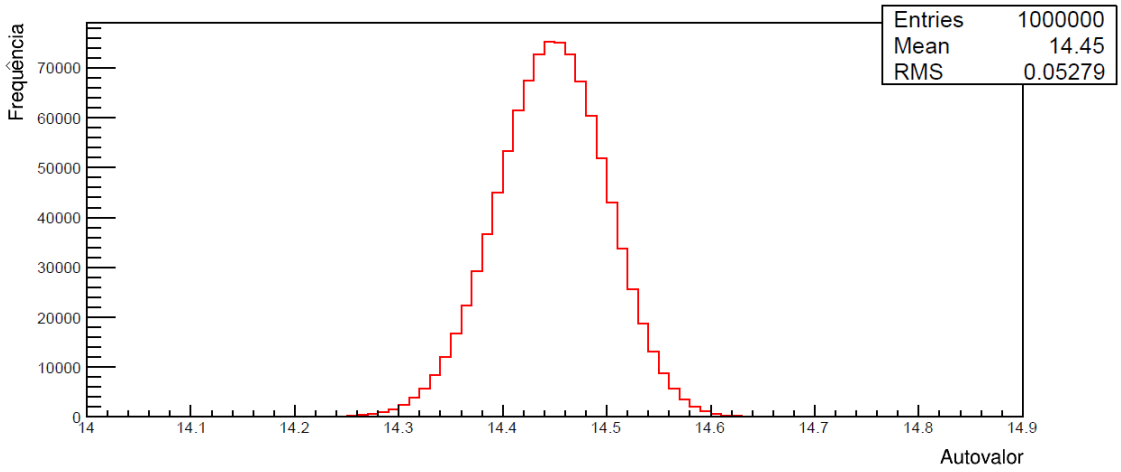
Fonte: Do autor.

Figura 15 – Da esquerda para a direita estão as distribuições de frequência do 5º ao 2º autovalor.



Fonte: Do autor.

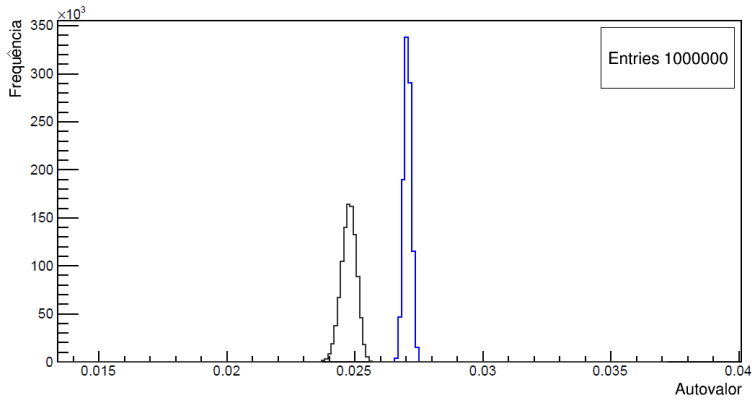
Figura 16 – Distribuição de frequência do autovalor correspondente à PC1.



Fonte: Do autor.

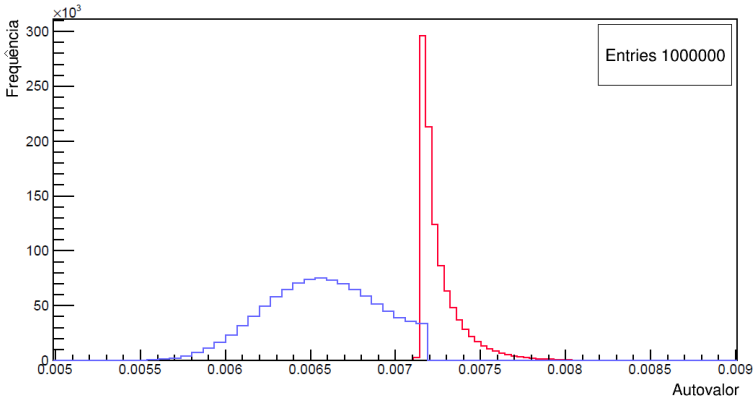
Como pode ser observado, a Figura 16 mostra a distribuição do maior autovalor (média 14,45) isolado, pois, ele se encontra muito distante do segundo maior (média 2,863). Como o método exige que para se utilizar os autovetores, seus correspondentes autovalores devem ter suas distribuições não superpostas, as Figuras 13-16 mostram que todas as distribuições estão distintas, com excessão dos autovalores 18 (vermelho claro) e 19 (azul claro) na Figura 13. Entretanto, devido a escala, pode ficar alguma dúvida se há superposição ou não das distribuições dos autovalores 21 (laranja) e 22 (preta) da Figura 13 e 15 (azul) e 16 (marrom) da Figura 14. Por isso, seguem suas distribuições em uma escala reajustada.

Figura 17 – Distribuição de frequência do 15º (azul) e 16º (marrom) autovalor.



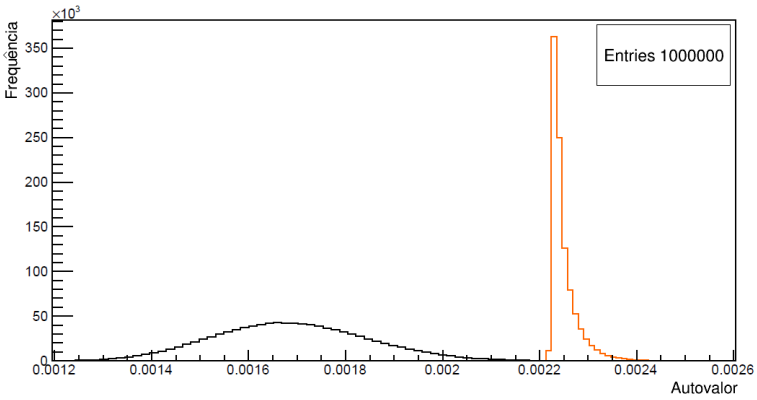
Fonte: Do autor.

Figura 18 – Distribuição de frequência do 18º (vermelho claro) e 19º (azul claro) autovalor.



Fonte: Do autor.

Figura 19 – Distribuição de frequência do 21º (laranja) e 22º (preta) autovalor.



Fonte: Do autor.

Com isso, pôde-se constatar que as distribuições do 15º e 16º autovalor, bem como do 21º e 22º autovalores não se sobrepõem significativamente, conforme as Figuras 17 e 19 respectivamente. Já a figura 18 mostra que as únicas distribuições superpostas, desta análise, correspondem ao 18º e 19º autovalores. É importante destacar que a validade do modelo depende das distribuições dos autovalores não serem superpostas. Neste caso, a superposição ocorreu nos autovalores correspondentes a 18ª e 19ª componentes principais e portanto não resultará em problema para a análise.

Entretanto, em casos que isso ocorrer nas primeiras componentes, uma análise de melhor qualidade pode ser comprometida, pois os autovetores correspondentes aos autovalores superpostos poderão ser comutados gerando valores distintos dentro do nível de confiança estipulado. Por exemplo, caso os autovalores se sobreponham dentro de um intervalo de 3σ de probabilidade, então estes falharão no teste de hipótese de que são distintos com um nível de confiança de 97%.

Assim, realizando os mesmos cálculos apresentados no caso de duas variáveis, seguem os resultados obtidos para as 23 variáveis do modelo.

Tabela 4 – Resultados da PCA para as 23 variáveis, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores e suas incertezas, as proporções explicativas e suas incertezas, as proporções acumuladas e suas incertezas e as incertezas relativas.

PC	Autovalor	σ_Autovalor	%Var	σ_%Var	%Cum	σ_%Cum	σ_Re
1	14,46	0,05	62,87	0,23	62,87	0,23	0,0036
2	2,863	0,013	12,45	0,06	75,32	0,24	0,0031
3	2,07	0,07	9,0	0,3	84,3	0,4	0,0047
4	1,289	0,010	5,60	0,04	89,9	0,4	0,0044
5	1,025	0,006	4,458	0,027	94,4	0,4	0,0042
6	0,260	0,004	1,130	0,019	95,5	0,4	0,0042
7	0,210	0,006	0,913	0,027	96,4	0,4	0,0042
8	0,178	0,003	0,772	0,014	97,2	0,4	0,0041
9	0,1554	0,0016	0,676	0,007	97,9	0,4	0,0041
10	0,1242	0,0019	0,540	0,008	98,4	0,4	0,0041
11	0,0976	0,0008	0,424	0,003	98,8	0,4	0,0041
12	0,0750	0,0007	0,326	0,003	99,2	0,4	0,0040
13	0,0599	0,0003	0,2602	0,0011	99,4	0,4	0,0040
14	0,0461	0,0020	0,200	0,009	99,6	0,4	0,0040
15	0,02709	0,00015	0,1178	0,0006	99,7	0,4	0,0040
16	0,02481	0,00029	0,1079	0,0013	99,9	0,4	0,0040
17	0,01090	0,00005	0,04739	0,00023	99,9	0,4	0,0040
18	0,00718	0,00013	0,0312	0,0006	99,9	0,4	0,0040
19	0,0066	0,0003	0,0287	0,0014	100,0	0,4	0,0040
20	0,00437	0,00007	0,0190	0,0003	100,0	0,4	0,0040
21	0,002235	0,000028	0,00972	0,00012	100,0	0,4	0,0040
22	0,00170	0,00016	0,0074	0,0007	100,0	0,4	0,0040
23	0,00015	0,00005	0,00065	0,00025	100,0	0,4	0,0040

Fonte: Do autor.

Ao aplicar os autovetores nas 29 observações, já excluído o *outlier* $HD143275$, foram obtidos os vetores de componentes principais para cada observação, juntamente com suas incertezas estimadas. Já os resultados obtidos por Ensor et al. (2017), para os autovalores, seguem na Tabela 5.

Tabela 5 – Resultados da PCA para as 23 variáveis, onde cada linha representa uma componente principal e cada coluna, da esquerda para a direita, representa os autovalores, as proporções explicativas e as proporções acumuladas.

PC	Autovalor	%Var	%Cum
1	15,248	66,30	66,30
2	3,158	13,73	80,03
3	1,801	7,83	87,86
4	1,139	4,95	92,81
5	0,355	1,54	94,35
6	0,262	1,14	95,49
7	0,192	0,84	96,33
8	0,186	0,81	97,14
9	0,157	0,68	97,82
10	0,117	0,51	98,33
11	0,096	0,42	98,75
12	0,074	0,32	99,07
13	0,066	0,29	99,36
14	0,055	0,24	99,60
15	0,032	0,14	99,74
16	0,025	0,11	99,85
17	0,012	0,05	99,90
18	0,008	0,03	99,93
19	0,006	0,03	99,96
20	0,005	0,02	99,98
21	0,003	0,01	99,99
22	0,002	0,01	100,00
23	0,000	0,00	100,00

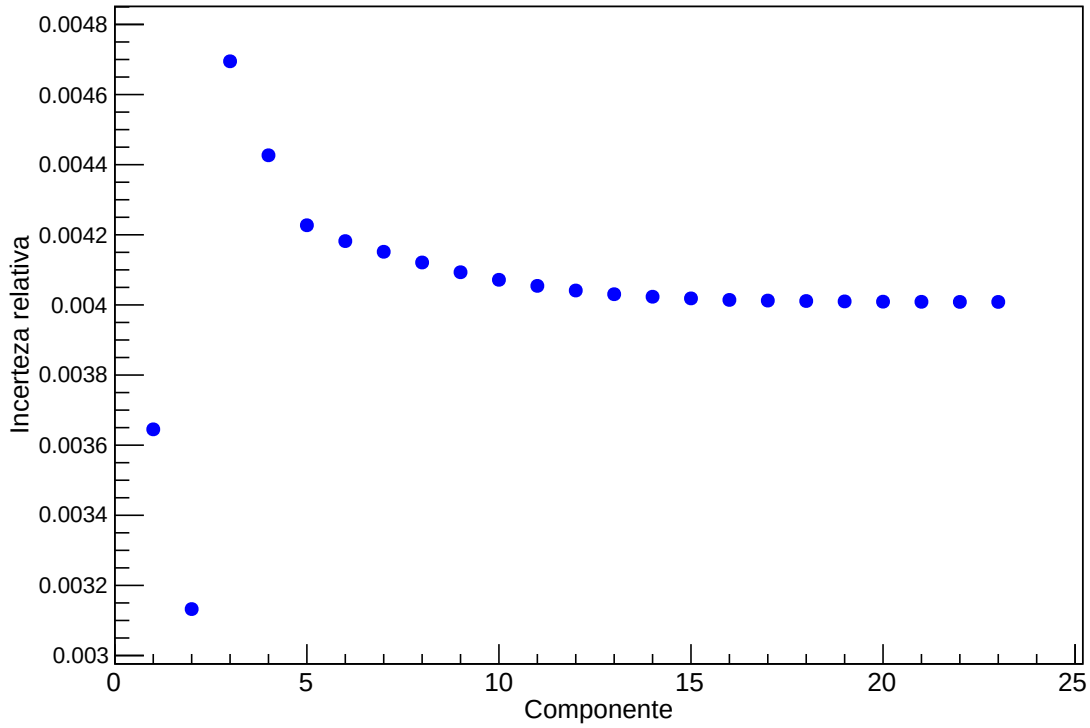
Fonte: (ENSOR et al., 2017).

De acordo com as Tabelas 4 e 5, os autovalores calculados pelo novo modelo, bem como a variabilidade explicada por cada componente, diferiram um pouco dos resultados obtidos por Ensor et al. (2017). Entretanto, isso é o que se espera ao inserir as incertezas experimentais na análise, pois a matriz de covariância, da qual foi extraído os autova-

lores e autovetores, foi modificada, uma vez que se usou a média ponderada dos dados em sua construção. Também é importante destacar que este novo modelo gera, como consequência, incertezas para os autovalores e demais resultados.

Assim como no caso de duas variáveis, a Tabela 4 mostra que mesmo que fossem utilizadas todas as componentes principais a variabilidade explicada poderia não ser 100% devido as incertezas do percentual cumulativo. Isso, é um dos pontos que diferencia esta metodologia do modelo padrão, pois a incerteza do percentual cumulativo é afetada pelas incertezas experimentais dos dados. Outro fato a ressaltar é que as incertezas relativas apresentaram um decaimento suave (Figura 20), com excessão da segunda para terceira que apresentou um salto de $0,00156$ unidades. Esse salto pode ser interpretado como uma fase de transição até que as incertezas relativas caiam assintoticamente. Contudo, esse valor inferido de incerteza relativa sugere uma nova forma de selecionar o número de componentes principais que irá representar o modelo trabalhado, pois pode-se estipular um limite de corte e utilizar um número de componentes cuja incerteza relativa da proporção explicativa acumulada já tenha estabilizado e seja igual ou inferior a este limite.

Figura 20 – Gráfico das incertezas relativas as proporções acumuladas de cada componente principal.

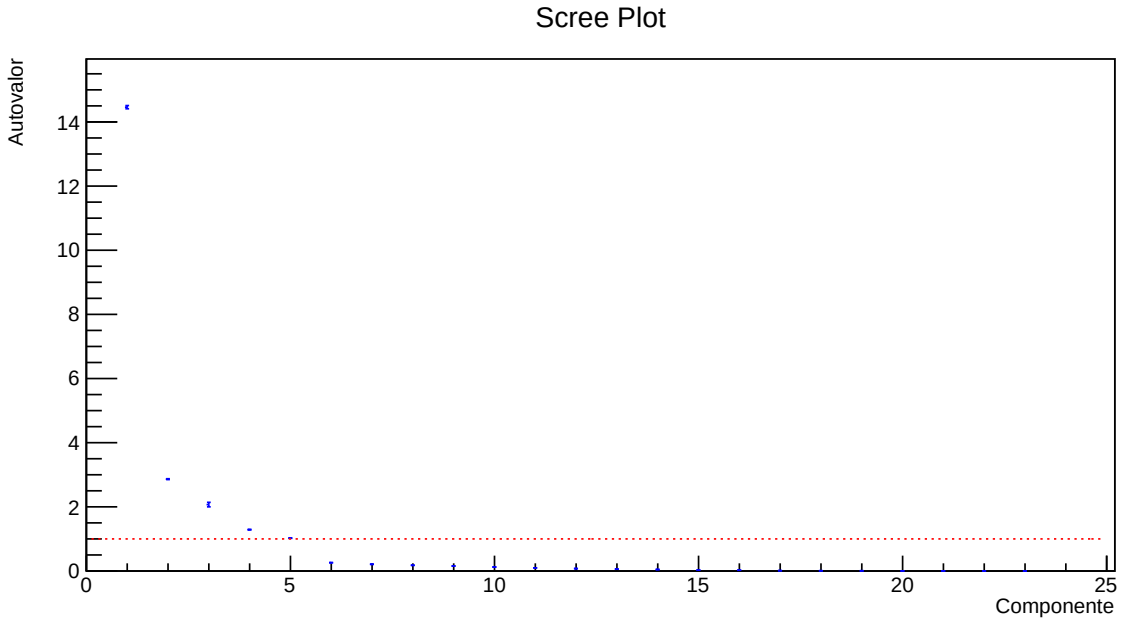


Fonte: Do autor.

Para selecionar o número de componentes principais a ser utilizado, Ensor et al.

(2017) utilizaram dois critérios: selecionar o número de componentes que juntas somassem pelo menos 90% da variabilidade total; e as componentes cujos autovalores fossem maiores que 1. E para auxiliar nesta interpretação utilizaram um *screeplot*, conforme foi mostrado na Figura 4. Pela Tabela 5 é possível perceber que as quatro primeiras componentes satisfazem essas duas condições. Para manter o mesmo raciocínio, foram utilizados os mesmos critérios e construído um *screeplot* (Figura 21).

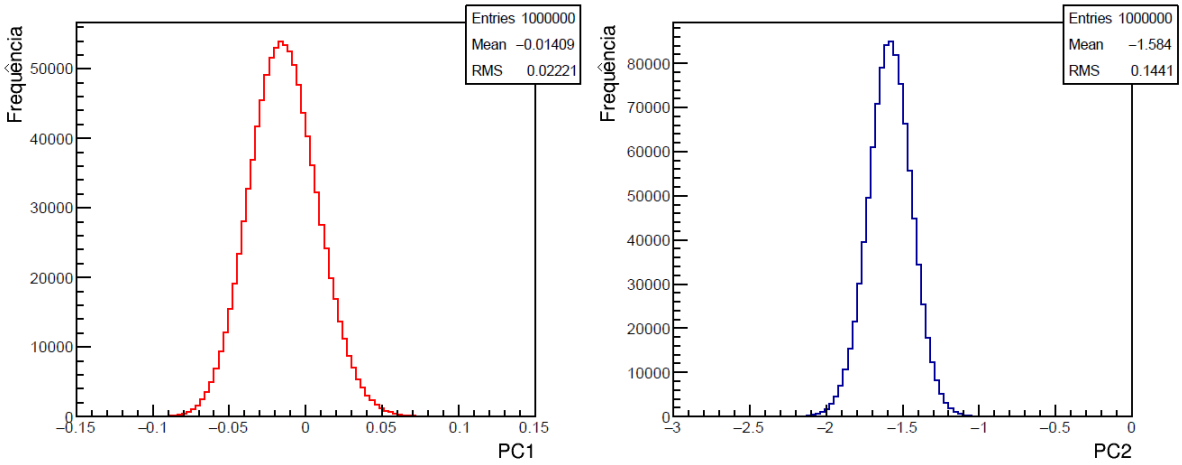
Figura 21 – Gráfico *Scree Plot* de autovalores por número da componente principal.



Fonte: Do autor.

Observando a Figura 21 e, também, com o auxílio da Tabela 4, percebe-se que as quatro primeiras componentes satisfazem as duas condições pois possuem um percentual explicativo acumulado igual à $89,9 \pm 0,4$ e os autovalores maiores que 1. Rigorosamente, a componente 5 também poderia ser escolhida dado que seu autovalor é $1,025 \pm 0,006$, entretanto, como o objetivo da análise é captar o menor número de variáveis possível e dado que o quinto autovalor, comparado ao quarto, é pouco maior que 1, optou-se por manter a mesma escolha de quatro componentes principais, uma vez que as quatro primeiras componentes cumprem satisfatoriamente as duas condições. Para ilustrar as componentes principais, a Figura 22 mostra a distribuição de frequência da primeira e da segunda componente principal referente à estrela $HD15137$.

Figura 22 – Componentes principais 1(vermelho) e 2(azul) referentes à estrela $HD15137$.



Fonte: Do autor.

Como pode ser observado, as médias das distribuições da PC1 e PC2, da estrela $HD15137$, são, respectivamente, $-0,01409$ e $-1,584$ e seus rms são, na sequência $0,02221$ e $0,1441$. Portanto, as PCs originais são dadas por $PC1 = -0,015 \pm 0,022$ e $PC2 = -1,59 \pm 0,14$, onde a incerteza de cada componente é o rms dos seus respectivos histogramas. Todas as componentes principais de cada uma das 29 estrelas estão disponíveis no Apêndice B. Já as interpretações e aplicabilidades destas componentes principais no estudo das DIBs permaneceram as mesmas após a aplicação deste novo método. Para uma melhor interpretação destas aplicações no estudo das DIBs, sugere-se a leitura de Ensor et al. (2017).

6.2 Análise Discriminante

Neste tópico, será demonstrado um exemplo genérico, desenvolvido pelo autor, de aplicação da análise discriminante, bem como possíveis aplicações reais deste método. Portanto, como já visto no capítulo 4, o método tradicional consiste em determinar os m vetores de coficientes discriminantes, $\hat{\boldsymbol{a}}$, que maximizam a razão dada pela equação (4.34), vetores estes, que são os m autovetores correspondentes aos m maiores autovalores da matriz $\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}}$, onde $m = \min(p, k-1)$ e então realizar as transformações lineares no centroide de cada grupo e em novas observações, a fim de se obter um vetor de *scores* discriminantes, $\hat{\boldsymbol{y}} = \hat{\boldsymbol{a}}^T\boldsymbol{x}$, conforme a equação (4.35). De acordo com os objetivos deste trabalho, seja um problema com k diferentes grupos, cada um possuindo uma matriz de

observações

$$\boldsymbol{X}_j = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n_j 1} \\ x_{12} & x_{22} & \cdots & x_{n_j 2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{n_j p} \end{pmatrix} \tag{6.31}$$

e de incertezas

$$\boldsymbol{\sigma}_{\boldsymbol{X}_j} = \begin{pmatrix} \sigma_{11} & \sigma_{21} & \cdots & \sigma_{n_j 1} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{n_j 2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{n_j p} \end{pmatrix}, \tag{6.32}$$

onde n_j representa a quantidade de observações do grupo j e p é o número de variáveis discriminantes. O novo método consiste em utilizar a média ponderada e sua incerteza, segundo as equações (6.5, 6.6), para estimar a posição do centroide de cada grupo, pois como os dados possuem incertezas, fica evidente que valores com maiores desvios devem influenciar menos na construção do método de classificação.

Com os novos centroides, pode-se determinar a matriz de covariância interna de cada população, $\boldsymbol{S}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$, e então construir a matriz $\boldsymbol{S}_c$ conforme a equação (4.16) e calcular sua inversa. Já a matriz $\hat{\boldsymbol{B}}$ é calculada segundo a equação (4.41), onde $\bar{\boldsymbol{x}}_G$ é, agora, a média ponderada dos centroides. Para realizar estes cálculos, dois algoritmos foram desenvolvidos pelo autor. Um para realizar a análise com dados de treinamento e o outro para avaliar novas observações. O primeiro programa, que deve ser alimentado com os dados de cada grupo juntamente com suas incertezas, calcula a matriz $\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}}$ original, como descrito no Capítulo 4, e extrai seus autovalores e autovetores que são utilizados nas funções discriminantes. Então, para cada centroide, é gerada uma posição aleatória com distribuição normal e com isso uma matriz $\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}}$ aleatória.

Esse processo é repetido um milhão de vezes e em cada uma delas, os autovalores, bem como as componentes dos autovetores alimentam um histograma individual. Como resposta o primeiro algoritmo fornece quatro matrizes, a dos centroides,

$$\bar{\boldsymbol{X}} = \begin{pmatrix} \bar{X}_{11} & \bar{X}_{12} & \cdots & \bar{X}_{1k} \\ \bar{X}_{21} & \bar{X}_{22} & \cdots & \bar{X}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{X}_{p1} & \bar{X}_{p2} & \cdots & \bar{X}_{pk} \end{pmatrix}, \tag{6.33}$$

onde cada coluna é o centroide de uma população, sua incerteza,

$$\boldsymbol{\sigma}_{\bar{X}} = \begin{pmatrix} \boldsymbol{\sigma}_{\bar{X}_{11}} & \boldsymbol{\sigma}_{\bar{X}_{12}} & \cdots & \boldsymbol{\sigma}_{\bar{X}_{1k}} \\ \boldsymbol{\sigma}_{\bar{X}_{21}} & \boldsymbol{\sigma}_{\bar{X}_{22}} & \cdots & \boldsymbol{\sigma}_{\bar{X}_{2k}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\sigma}_{\bar{X}_{p1}} & \boldsymbol{\sigma}_{\bar{X}_{p2}} & \cdots & \boldsymbol{\sigma}_{\bar{X}_{pk}} \end{pmatrix}, \tag{6.34}$$

a de vetores discriminantes,

$$\hat{\boldsymbol{a}}^T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{pmatrix}, \tag{6.35}$$

e sua incerteza

$$\boldsymbol{\sigma}_{\hat{\boldsymbol{a}}^T} = \begin{pmatrix} \sigma_{a_{11}} & \sigma_{a_{12}} & \cdots & \sigma_{a_{1p}} \\ \sigma_{a_{21}} & \sigma_{a_{22}} & \cdots & \sigma_{a_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_{m1}} & \sigma_{a_{m2}} & \cdots & \sigma_{a_{mp}} \end{pmatrix}. \tag{6.36}$$

Já o segundo algoritmo será alimentado com estes valores e também com novas observações. Desta forma, pode-se calcular

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{a}}^T \hat{\boldsymbol{X}} = \begin{pmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1N} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_{m1} & \hat{y}_{m2} & \cdots & \hat{y}_{mN} \end{pmatrix} \tag{6.37}$$

e

$$\bar{\boldsymbol{Y}} = \hat{\boldsymbol{a}}^T \bar{\boldsymbol{X}} = \begin{pmatrix} \bar{y}_{11} & \bar{y}_{12} & \cdots & \bar{y}_{1N} \\ \bar{y}_{21} & \bar{y}_{22} & \cdots & \bar{y}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{y}_{m1} & \bar{y}_{m2} & \cdots & \bar{y}_{mN} \end{pmatrix}. \tag{6.38}$$

onde $\hat{\boldsymbol{X}}$, $(p \times N)$, é a matriz de N novas observações que será usada para calcular $\hat{\boldsymbol{Y}}$, $(m \times N)$, que é a matriz dos novos *scores* discriminantes, e $\bar{\boldsymbol{Y}}$, $(m \times k)$, é a matriz de *scores* médios em que cada coluna representa o vetor de *scores* de cada centroide.

Considerando que os centroides, os vetores discriminantes e as novas observações possuem incertezas, torna-se possível gerar matrizes aleatórias para os *scores* discriminantes e médios. Portanto, novamente, o algoritmo gera valores aleatórios normalmente distribuidos, com as entradas das matrizes $\hat{\boldsymbol{a}}^T$, $\hat{\boldsymbol{X}}$ e $\bar{\boldsymbol{X}}$ como média e suas incertezas como desvio padrão. Este procedimento é repetido um milhão de vezes e em cada iteração, os valores gerados são armazenados em histogramas, de modo que suas incertezas estimadas sejam o rms de cada um.

É importante observar que devido os elementos da matriz de *scores* discriminantes bem como os da matriz de *scores* médios possuírem incertezas, não podem mais serem interpretados como pontos em um determinado espaço, mas sim como nuvens de probabilidade. Com isso, a classificação de uma nova observação em um determinado grupo, como é realizada pelo método clássico, torna-se sem fundamento. Isso ocorre devido ao fato de que tanto os *scores* discriminantes da observação quanto os *scores* médios dos

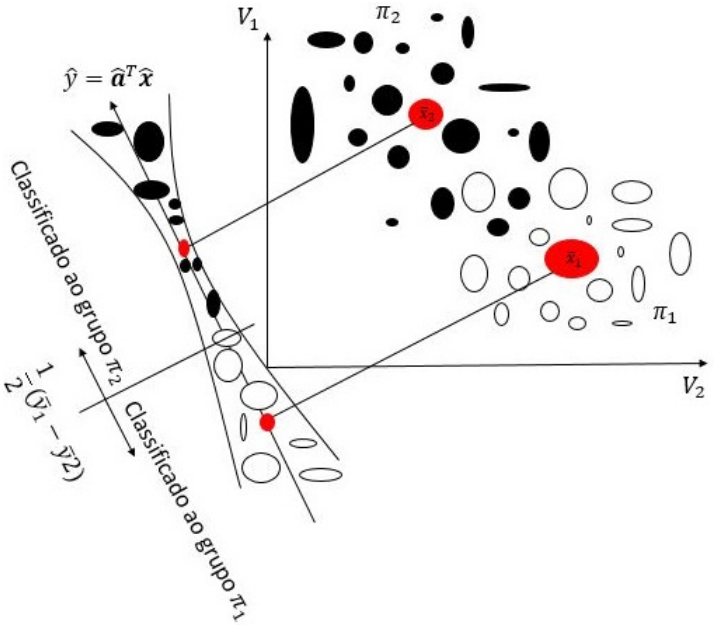
grupos possuem uma determinada probabilidade de estar em qualquer lugar dentro de suas respectivas nuvens de incerteza. Portanto, passa a ser inviável afirmar com certeza que uma nova observação possui um *score* discriminante que esteja mais próximo de um determinado *score* médio, pois existe a possibilidade da verdadeira posição desse valor estar mais próxima de outro.

Assim, a avaliação realizada por este último algoritmo, consiste em determinar qual é probabilidade de uma nova observação pertencer a cada um dos k grupos. Dado que, em cada iteração é gerado um *score* discriminante para a nova observação e para cada centroide. O algoritmo também calcula a distância de Mahalanobis com relação a cada população e classifica a nova observação àquele grupo no qual essa distância é menor. Como se conhece o número de iterações e o número de vezes que esta nova observação foi classificada em cada grupo é possível calcular a probabilidade desta medida pertencer a cada uma das populações. Assim, seja I o número de iterações e C_j o número de vezes que a observação foi classificada na população j, sua probabilidade de pertencer a este grupo será

$$P_j = \frac{C_j}{I}. \tag{6.39}$$

Desta forma, para uma melhor interpretação, pode-se fazer uma analogia à Figura (5), e construir um diagrama como é mostrado na Figura (23), onde tanto as observações quanto seus *scores* discriminantes são representados por nuvens de acordo com suas incertezas, e isso se aplica também aos centroides.

Figura 23 – Diagrama com a representação da nova função discriminante aplicada ao caso de dois grupos e duas variáveis discriminantes.



Fonte: Do autor.

Uma possível aplicação desta ferramenta está no campo da astronomia, em que se pode determinar a probabilidade de novos exoplanetas pertencerem ao grupo dos planetas telúricos (mais densos e menores) ou gigantes gasosos (menos densos e maiores), utilizando como variáveis discriminantes a massa do planeta, seu diâmetro, dentre outras.

A título de exemplo, uma aplicação da LDA, seria considerar uma situação que se tem dois grupos e duas variáveis discriminantes, como mostram as matrizes abaixo.

$$\boldsymbol{X}^{(1)} = \begin{pmatrix} 220 & 250 & 180 \\ 56 & 78 & 45 \end{pmatrix}, \quad \boldsymbol{X}^{(2)} = \begin{pmatrix} 45 & 67 & 88 \\ 115 & 131 & 99 \end{pmatrix} \tag{6.40}$$

Dado que o limite de aplicabilidade do método depende das distribuições de frequências dos autovalores da matriz $\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}}$ não estarem significativamente sobrepostos e isto, por sua vez, depende da dimensão da incerteza de cada observação, neste exemplo foi elaborada uma simulação com incertezas relativas às observações nas escalas de 1%, 5%, 15% e 25%.

Desta forma, a matriz de incertezas de 1%, por exemplo, será

$$\boldsymbol{\sigma}_{1\%}^{(1)} = \begin{pmatrix} 2{,}2 & 2{,}5 & 1{,}8 \\ 0{,}56 & 0{,}78 & 0{,}45 \end{pmatrix}. \tag{6.41}$$

Com os valores da equação (6.41), pode-se construir a matriz de centroides, utilizando a equação (6.5), e obter

$$\bar{\boldsymbol{X}} = \begin{pmatrix} 208{,}83 & 57{,}33 \\ 54{,}14 & 112{,}03 \end{pmatrix}, \quad \boldsymbol{\sigma}_{\bar{\boldsymbol{X}}_{1\%}} = \begin{pmatrix} 1{,}22 & 0{,}34 \\ 0{,}32 & 0{,}65 \end{pmatrix}. \tag{6.42}$$

Uma vez que em cada simulação os valores de todas as incertezas serão alterados na mesma proporção, a matriz de centroides permanece a mesma, mudando apenas os valores de suas incertezas. Já as matrizes de covariância individual de cada grupo, podem ser construídas através da equação (2.31). Portanto, tem-se que

$$\boldsymbol{S}_1 = \begin{pmatrix} 1326 & 633 \\ 633 & 328 \end{pmatrix}, \quad \boldsymbol{S}_2 = \begin{pmatrix} 592 & -127 \\ -127 & 269 \end{pmatrix}. \tag{6.43}$$

Com isso, através da equação (4.16) chega-se na matriz de covariância comum, dada por

$$\boldsymbol{S}_c = \begin{pmatrix} 959 & 253 \\ 253 & 299 \end{pmatrix}, \tag{6.44}$$

cuja inversa é

$$\boldsymbol{S}_c^{-1} = \begin{pmatrix} 0{,}0013 & -0{,}0011 \\ -0{,}0011 & 0{,}0043 \end{pmatrix}. \tag{6.45}$$

Calculando-se a média ponderada dos centroides, através da equação (6.5), tem-se que

$$\bar{\boldsymbol{X}}_G = \begin{pmatrix} 68{,}56 \\ 65{,}40 \end{pmatrix} \tag{6.46}$$

e utilizando a equação (4.41), pode-se calcular a matriz $\hat{\boldsymbol{B}}$, dada por

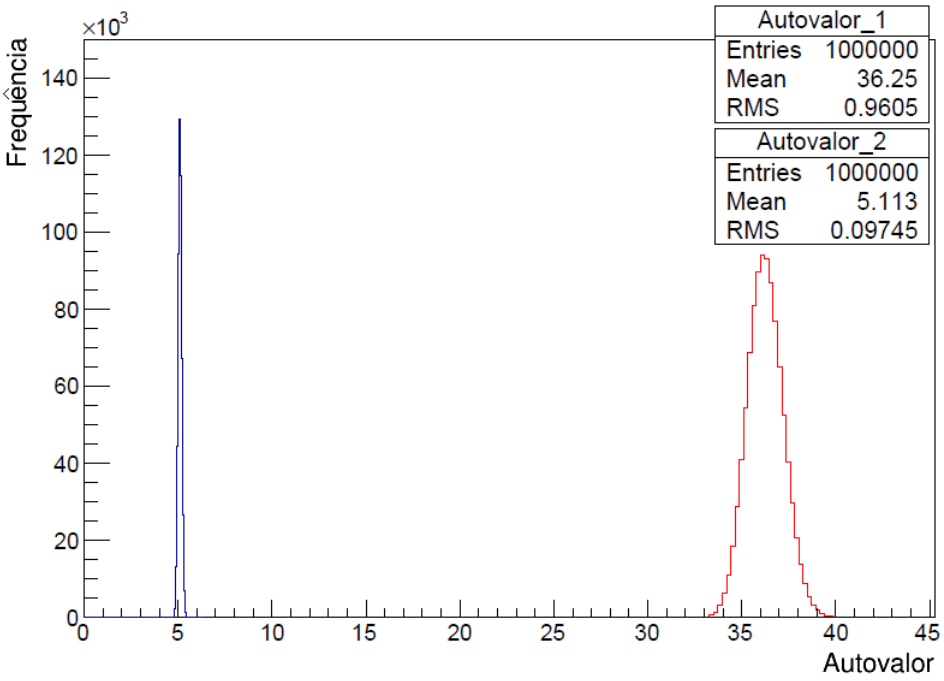
$$\hat{\boldsymbol{B}} = \begin{pmatrix} 19{,}800 & -2{,}102 \\ -2{,}102 & 2{,}301 \end{pmatrix}. \tag{6.47}$$

Finalmente, a matriz cujos autovalores e autovetores devem ser extraídos é

$$\boldsymbol{S}_c^{-1}\hat{\boldsymbol{B}} = \begin{pmatrix} 29 & -5 \\ -32 & 12 \end{pmatrix}. \tag{6.48}$$

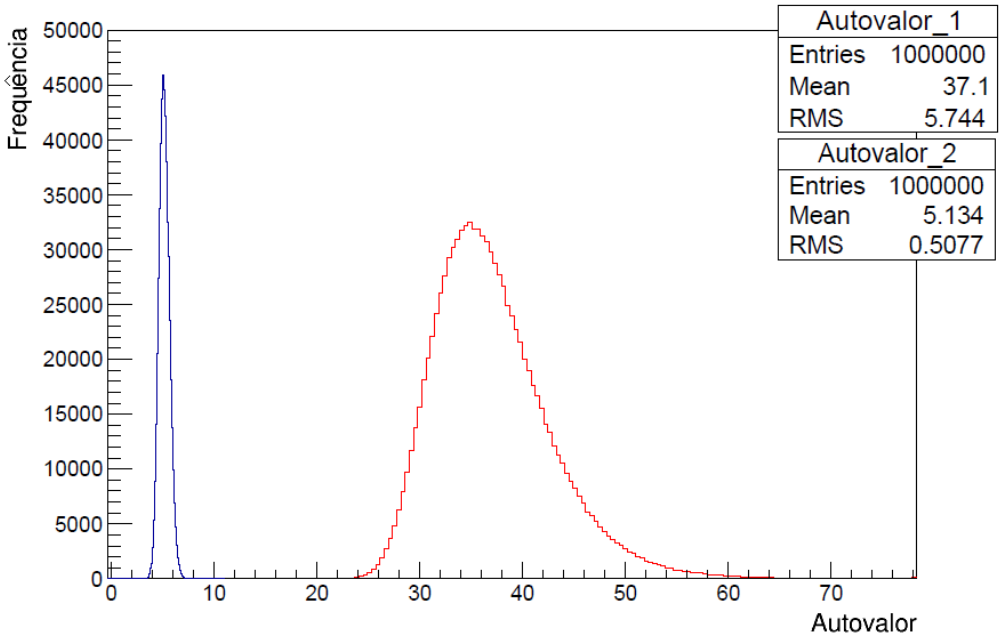
O mesmo procedimento foi realizado para os casos em que os dados possuiam incertezas de 5%, 15% e 25%. Para avaliar as distribuições de frequências dos autovalores, seguem as Figuras 24 - 27.

Figura 24 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 1%.



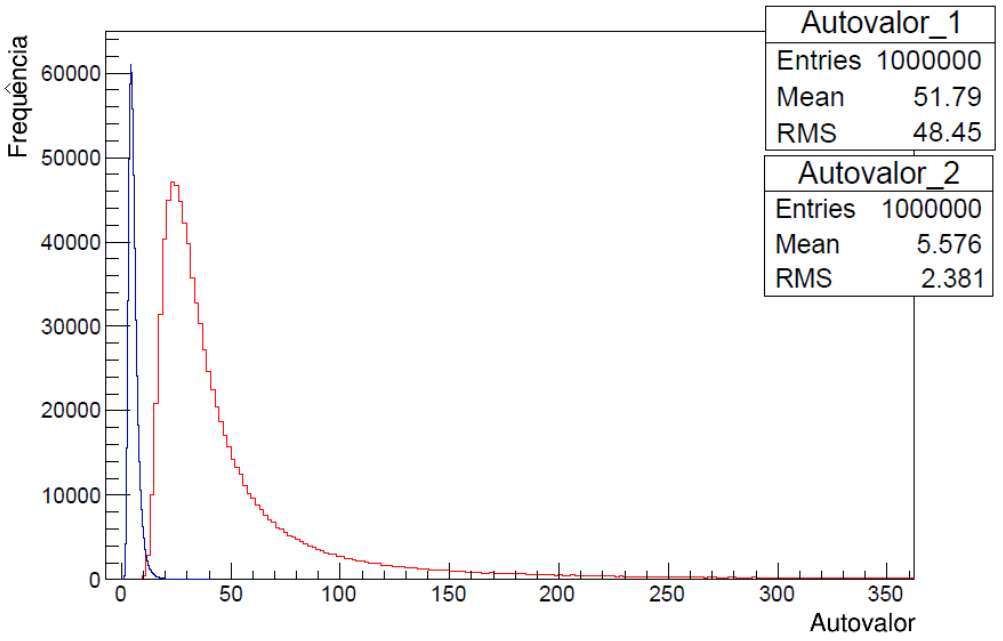
Fonte: Do autor.

Figura 25 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 5%.



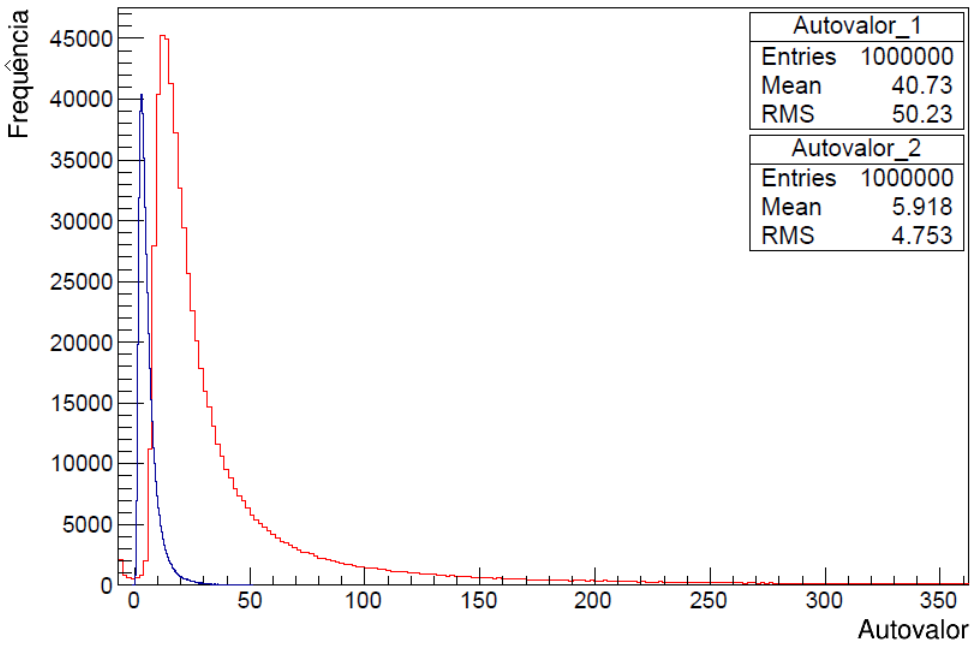
Fonte: Do autor.

Figura 26 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 15%.



Fonte: Do autor.

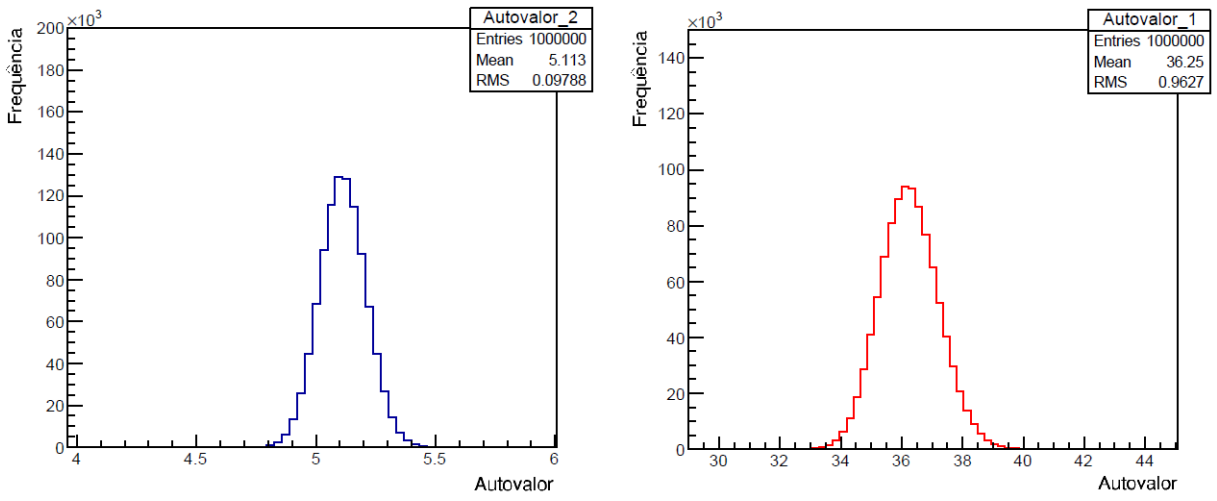
Figura 27 – Distribuições de frequência dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 25%.



Fonte: Do autor.

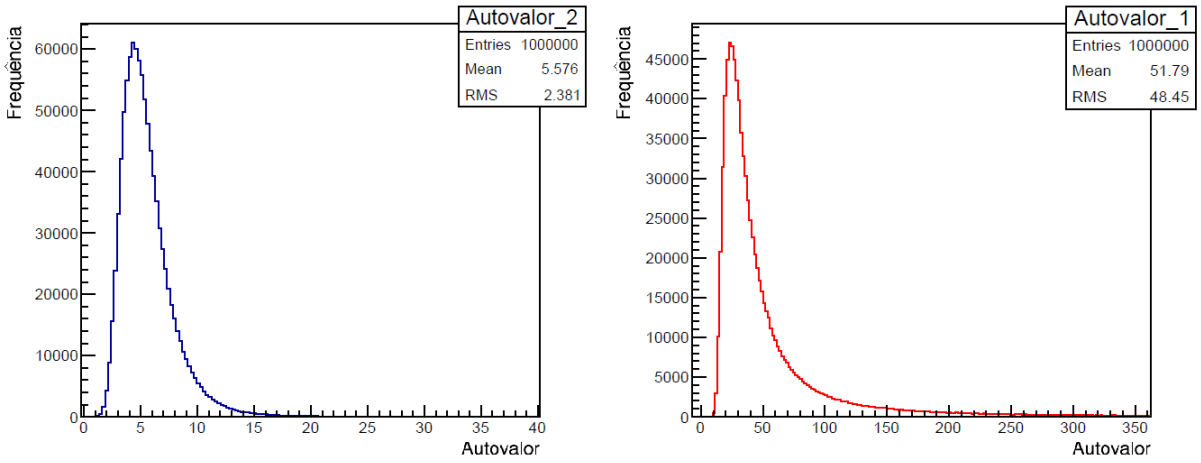
Para uma melhor visualização das distribuições individuais dos autovalores, referente aos dados com incertezas de 1% e 15%, seus histogramas foram plotados, respectivamente, nas Figuras 28 e 29.

Figura 28 – Distribuições de frequência, individuais, dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 1%.



Fonte: Do autor.

Figura 29 – Distribuições de frequência, individuais, dos autovalores correspondentes à primeira (vermelho) e segunda (azul) funções discriminantes, da análise dos dados cujas incertezas eram de 15%.



Fonte: Do autor.

De acordo com as Figuras 24 - 26, é possível observar que conforme as medidas passam a ter incertezas experimentais maiores, as distribuições de frequências dos autovalores sofrem deslocamentos e passam a ter uma variância maior. Neste exemplo, o limite da aplicabilidade do método ocorre quando as observações medidas possuem incertezas iguais ou superiores a 15%. Outro ponto importante é que pode-se observar que as distribuições que tinham um formato bem definido de uma gaussiana (Figura 28) passaram a ficar assimétricas (Figura 29), à medida que as incertezas experimentais aumentaram. Isso é um indício de que mesmo que os elementos do vetor médio tenham distirbuição normal, a verdadeira natureza das distribuições dos autovalores pode não ser gaussiana.

Para o caso particular dos dados possuírem incertezas de 5%, como há apenas dois grupos e duas variáveis discriminantes, o número de funções discriminantes será $m = \min(p, k-1) = (2, 1) = 1$. Ao aplicar o vetor de *scores* discriminantes na matriz de centroides, encontram-se os valores dos *scores* médios de cada população. Assim, os resultados obtidos se encontram na Tabela 6.

Tabela 6 – Resultados da análise com amostra de treinamento. Da esquerda para a direita estão os coeficientes discriminantes , suas incertezas, o centroide da população 1, sua incertezas, o centroide da população 2 e sua incerteza.

$\hat{a}$	$\sigma_{\hat{a}}$	C_1	σ_{C1}	C_2	σ_{C1}
0,60	0,05	208,8	1,2	57,4	0,3
-0,79	0,04	54,1	0,3	112,0	0,7

Fonte: Do autor.

Desta maneira, a função discriminante, que servirá para avaliar novas observações é dada por

$$Y = \hat{\boldsymbol{a}}^T \hat{\boldsymbol{X}} = 0,60\hat{X}_1 - 0,79\hat{X}_2. \tag{6.49}$$

Portanto, para avaliar novas medidas, dadas por

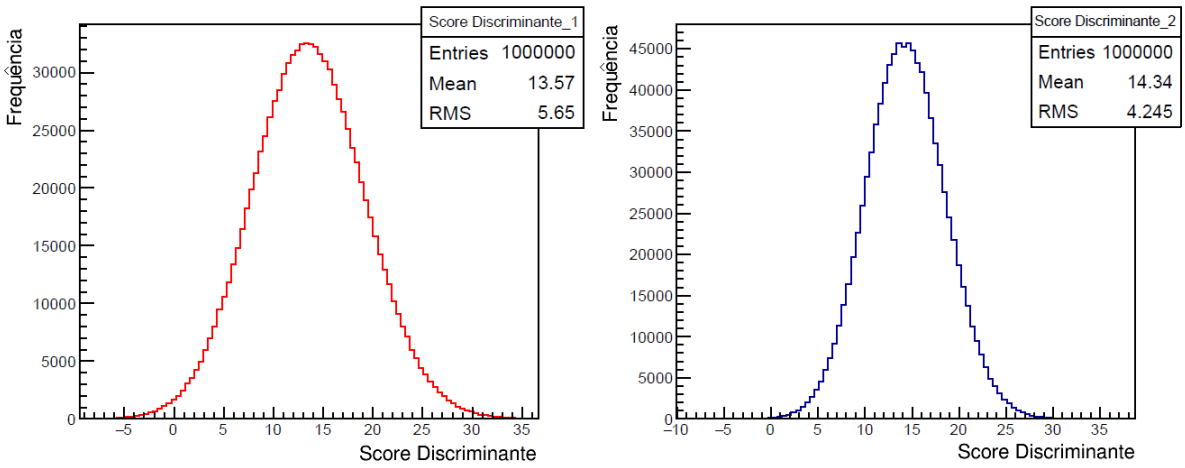
$$\hat{\boldsymbol{X}} = \begin{pmatrix} 86,2 & 66,7 \\ 48,1 & 32,4 \end{pmatrix}, \tag{6.50}$$

e suas incertezas, da ordem, por exemplo, de 5%,

$$\boldsymbol{\sigma}_{\hat{\boldsymbol{X}}} = \begin{pmatrix} 4,3 & 3,3 \\ 2,4 & 1,6 \end{pmatrix}, \tag{6.51}$$

o algoritmo gerou valores aleatórios (um milhão), normalmente distribuídos, para os coeficientes discriminantes, centroides (Tabela 6) e novas observações, dadas pela equação (6.50), e os aplicou na equação (6.49). Como resultado, foi obtida uma distribuição de frequência para os *scores* discriminantes, como mostra a Figura 30.

Figura 30 – Distribuições de frequência dos *scores* discriminantes das observações 1 (vermelho) e 2 (azul).



Fonte: Do autor.

Desta forma, utilizando o rms dos histogramas da Figura 30 como estimativa das incertezas dos *scores* discriminantes, tem-se que

Tabela 7 – *Scores* discriminantes das observações 1 (coluna 1) e 2 (colunas 2).

SD_1	SD_1
13,7 ± 5,7	14,3 ± 4,2

Fonte: Do autor.

Conforme o método clássico de LDA, ao calcular a distância de Mahalanobis dos *scores* discriminantes e médios seria determinado que as observações 1 e 2 pertenceriam aos grupos 2 e 1 respectivamente. Entretanto, para esse novo método, ao utilizar equação (6.39) foi possível estimar quais eram as probabilidades das novas observações pertencerem a cada um dos grupos, conforme estão dispostas na Tabela 8.

Tabela 8 – Resultados das probabilidades de pertencer a cada um dos grupos, onde cada linha representa uma nova observação e as colunas dois e três representam as probabilidades.

Observação	%Grupo 1	%Grupo 2
1	48	52
2	53	47

Fonte: Do autor.

Neste caso, conforme a Tabela 8, a observação 1 possui 48% de chance de pertencer ao grupo 1 e 52% de pertencer ao grupo 2. Já a observação 2 possui, para os mesmo grupos, respectivamente, 53% e 47% de chance. Esses resultados mostram que a classificação pelo método tradicional, de que a observação 1 pertence ao grupo 2, por exemplo, não faz sentido dado que as chances dela pertencer aos dois grupos são próximas. Assim, fica a critério do pesquisador que aplicar o método, utilizar uma técnica para estipular um valor de corte para então determinar se uma nova observação será considerada como pertencente ao um grupo ou a outro, de modo que a observação seja considerada de uma determinada população se a probabilidade de pertencer a esta população for superior ao valor estipulado.

Portanto, é importante destacar que a resposta em probabilidade do modelo desenvolvido neste trabalho não deve ser confundida com a probabilidade conhecida, *a priori*, de uma observação pertencer a uma população, utilizada no método clássico. A análise discriminante tradicional exige que se conheça os k diferentes grupos, as observações que pertencem a eles e como se comportam, ou seja, deve-se conhecer a função densidade de probabilidade que descreve o comportamento dos dados.

A partir disto, um critério de classificação é construído e novas observações são discriminadas de maneira determinística, criando assim um risco de estar errado. O que este modelo traz de novo é que usa-se, sim, o conhecimento, *a priori*, das populações mas ao analizar uma nova medida o método não determina a qual população ela pertence, mas estima e deixa explícito a probabilidade dela pertencer a cada grupo. Com isso o aplicador conhece as chances de estar errado e tem a oportunidade de avaliar mais cuidadosamente uma determinada observação antes de tomar qualquer decisão.

Entretanto, é importante destacar que este método difere da Análise Discriminante Linear Probabilística ou PLDA (*Probabilistic Linear Discriminant Analysis*), que foi desenvolvida com objetivo, a princípio, de realizar reconhecimento facial, mas também pode ser utilizada para outros fins, como por exemplo reconhecimento de voz. Para uma melhor descrição da PLDA pode-se consultar as referências (FERRER, 2017; LU; RENALS, 2014; SIZOV; LEE; KINNUNEN, 2014). Assim, essa nova formulação da análise discriminante será denominada como análise probabilística de classificação, dados os resultados que ela produz.

6.3 Análise de Correlação Canônica

Neste tópico, assim como no tópico de análise discriminante, será demonstrada uma aplicação hipotética da análise de correlações canônicas e discutidos os efeitos das incertezas experimentais. Desta forma, seja um problema com 20 observações, em que cada uma possui os vetores $\boldsymbol{X} = [X_1, X_2, X_3]^T$ e $\boldsymbol{Y} = [Y_1, Y_2, Y_3, Y_4]^T$, e cujos valores estão dispostos na Tabela 9.

Tabela 9 – Dados dos vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$, onde cada linhar representa uma observação.

Obs	X1	X2	X3	Y1	Y2	Y3	Y4
1	159,8	622,5	1924,1	98,5	314,9	181,2	200,0
2	117,9	844,7	985,1	27,5	112,7	151,1	1091,6
3	17,8	553,6	1297,3	15,6	175,3	156,4	29,7
4	38,1	608,8	607,0	49,5	242,1	46,3	445,6
5	23,4	76,5	719,3	72,2	107,2	218,4	1277,2
6	148,6	574,7	161,1	38,7	370,7	310,0	62,7
7	146,0	644,0	1936,3	12,9	393,7	227,1	1322,9
8	93,5	463,3	1189,9	40,6	94,7	170,5	165,0
9	243,6	239,8	1062,2	84,0	262,5	198,7	340,4
10	220,8	513,1	1131,6	44,8	197,8	200,0	675,0
11	69,3	632,3	1346,0	86,6	296,9	243,1	368,8
12	93,3	981,8	718,6	55,1	309,6	309,7	685,4
13	96,9	947,8	1310,0	3,5	189,7	170,8	1127,6
14	88,9	561,5	1366,4	53,0	149,4	78,0	1429,7
15	230,1	637,9	721,4	53,2	116,0	275,8	130,4
16	62,7	737,2	338,1	22,7	77,5	280,1	1300,7
17	146,1	937,4	1558,6	47,6	394,8	52,8	76,1
18	58,8	121,2	648,5	61,0	6,4	32,1	1160,4
19	62,4	266,9	1945,7	10,2	67,5	196,2	824,0
20	207,3	713,9	1086,8	52,0	84,7	240,1	355,6

Fonte: Do autor.

Assim, para construir as matrizes dadas pelas equações (5.26, 5.28), foi utilizado, como exemplo, os dados com incertezas relativas de 10%. Os resultados para as médias ponderadas se encontram na Tabela 10.

Tabela 10 – Resultados obtidos para as médias ponderadas (linha 2) e suas incertezas (linha 3) para cada uma das variáveis dos vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$ (linha 1).

Variável	X1	X2	X3	Y1	Y2	Y3	Y4
MP	38,1	173,6	399,4	7,9	11,1	68,7	59,3
σ_MP	1,2	5,7	12,3	0,3	0,6	2,1	2,4

Fonte: Do autor.

A partir dos resultados da Tabela 10, pode-se construir as matrizes de covariâncias

$$\boldsymbol{S}_x = \begin{pmatrix} 11214 & 36352 & 60869 \\ 36352 & 242884 & 310647 \\ 60869 & 310647 & 777742 \end{pmatrix} \tag{6.52}$$

e

$$\boldsymbol{S}_y = \begin{pmatrix} 2254 & 8132 & 4757 & 20422 \\ 8132 & 50833 & 24749 & 96824 \\ 4757 & 24749 & 21675 & 69108 \\ 20422 & 96824 & 69108 & 617024 \end{pmatrix} \tag{6.53}$$

cujas inversas são

$$\boldsymbol{S}_x^{-1} = \begin{pmatrix} 0,0001912 & -0,00001938 & -0,000007224 \\ -0,00001938 & 0,00001038 & -0,000002630 \\ -0,000007224 & -0,000002630 & 0,000002902 \end{pmatrix} \tag{6.54}$$

e

$$\boldsymbol{S}_y^{-1} = \begin{pmatrix} 0,001161 & -0,0001313 & -0,00007466 & -0,000009444 \\ -0,0001313 & 0,00006080 & -0,00003740 & -0,000001006 \\ -0,00007466 & -0,00003740 & 0,0001223 & -0,000005360 \\ -0,000009444 & -0,000001006 & -0,000005360 & 0,000002691 \end{pmatrix} \tag{6.55}$$

e, também, a matriz de covriância entre $\boldsymbol{X}$ e $\boldsymbol{Y}$,

$$\boldsymbol{S}_{xy} = \begin{pmatrix} 3615 & 17447 & 11377 & 37304 \\ 14728 & 93437 & 54533 & 240786 \\ 27665 & 152295 & 77683 & 439032 \end{pmatrix}. \tag{6.56}$$

Finalmente, a matrizes das quais devem ser extraídos os autovalores e autovetores são

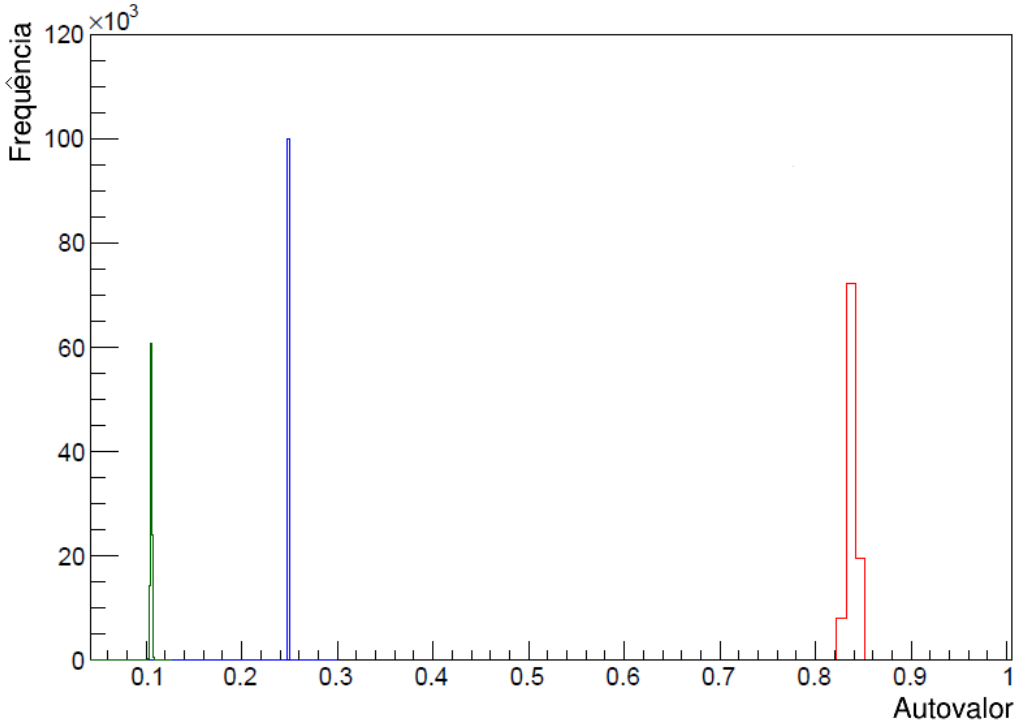
$$\boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy}\boldsymbol{S}_y^{-1}\boldsymbol{S}_{yx} = \begin{pmatrix} 0,36 & 0,66 & 0,83 \\ 0,066 & 0,51 & 0,68 \\ 0,015 & 0,12 & 0,32 \end{pmatrix} \tag{6.57}$$

e

$$\boldsymbol{S}_y^{-1}\boldsymbol{S}_{yx}\boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy} = \begin{pmatrix} 0,16 & 0,10 & 0,18 & -0,64 \\ 0,075 & 0,55 & 0,28 & 1,67 \\ 0,054 & 0,25 & 0,26 & 0,07 \\ 0,0041 & 0,048 & 0,010 & 0,22 \end{pmatrix}. \tag{6.58}$$

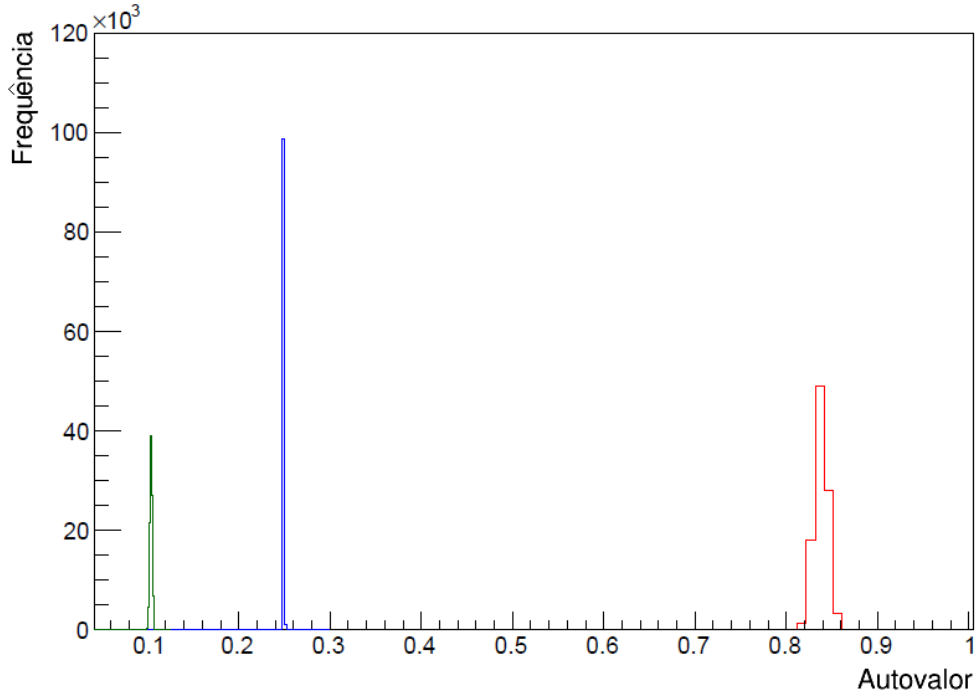
A fim de avaliar o comportamento das distribuições dos autovalores para diferentes dimensões de incertezas, esses cálculos foram realizados para simulações com incertezas relativas às medidas de 3%, 5%, 10% e 15%. Desta forma, seguem os histogramas nas Figuras 31 - 34, onde a distribuição do primeiro autovalor está representada em vermelho, a do segundo em azul e a do terceiro em verde.

Figura 31 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 3%.



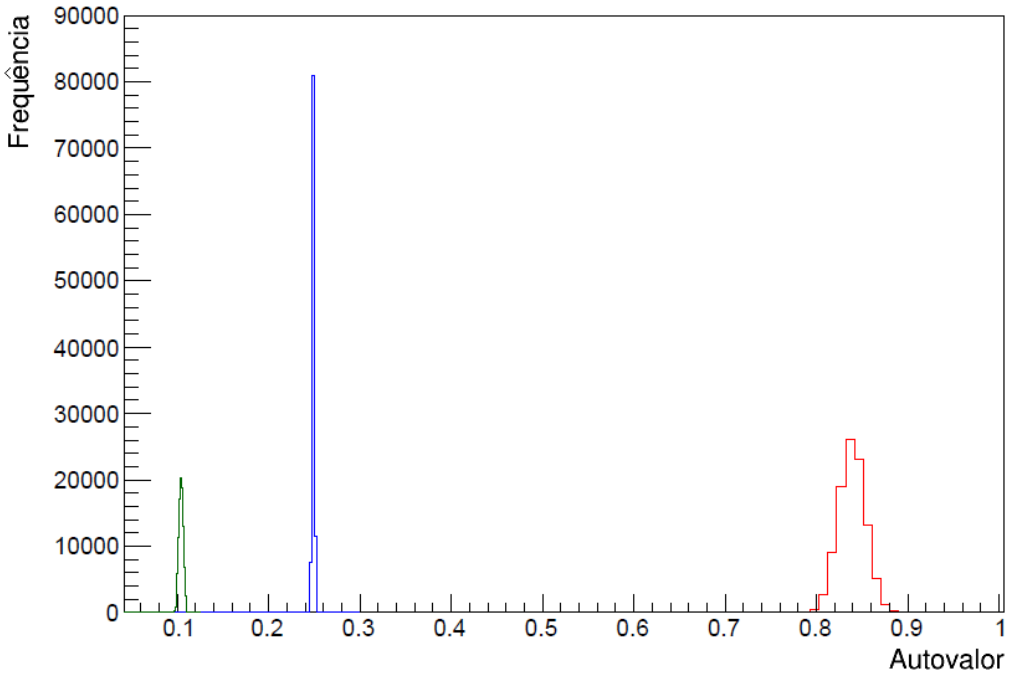
Fonte: Do autor.

Figura 32 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 5%.



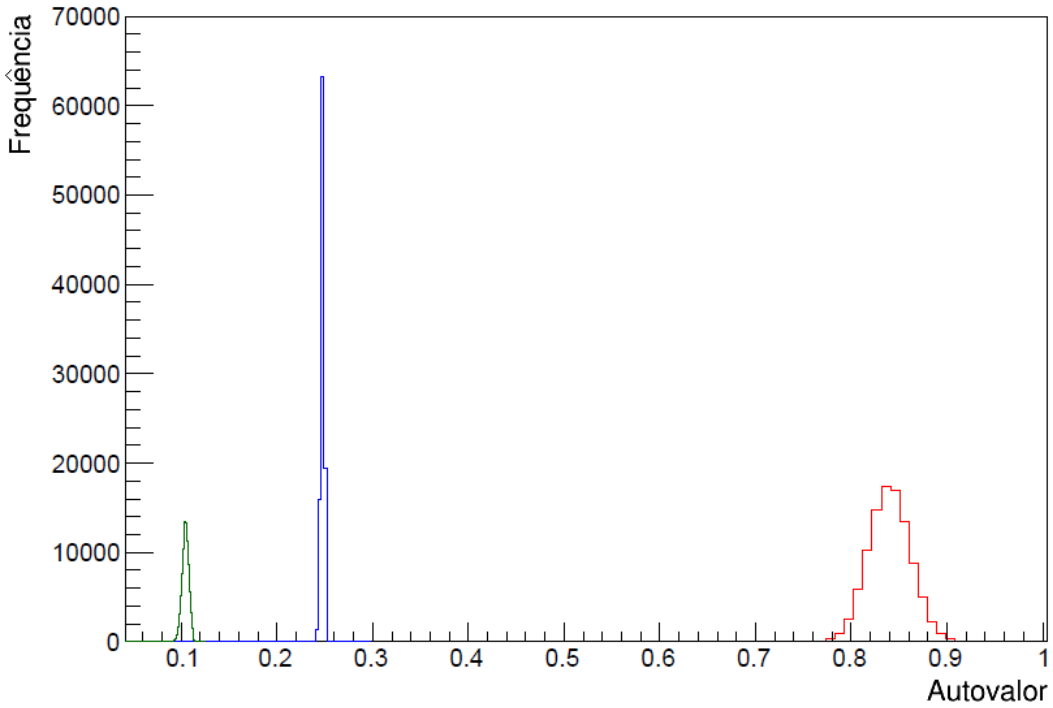
Fonte: Do autor.

Figura 33 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 10%.



Fonte: Do autor.

Figura 34 – Distribuições de frequência dos autovalores para os dados com incertezas relativas de 15%.



Fonte: Do autor.

Como o autovalor é o quadrado da correlação canônica de um determinado par de variáveis canônicas, então seguem, na Tabela 11, os resultados obtidos.

Tabela 11 – Resultados dos autovalores, suas incertezas (colunas 2 e 3), correlações canônicas e suas incertezas (colunas 4 e 5).

Simulação	Autovalor	σ_Autovalor	CC (%)	σ_CC (%)
3%	0,838	0,004	91,53	0,23
	0,2493	0,0003	49,93	0,03
	0,1050	0,0007	32,41	0,11
5%	0,838	0,007	91,5	0,4
	0,2489	0,0005	49,89	0,05
	0,1038	0,0012	32,21	0,18
10%	0,838	0,014	91,5	0,8
	0,2490	0,0011	49,90	0,11
	0,1040	0,0023	32,3	0,4
15%	0,837	0,022	91,5	1,2
	0,2481	0,0016	49,81	0,16
	0,104	0,003	32,3	0,5

Fonte: Do autor.

Com isso, é possível perceber que em nenhuma simulação as distribuições de frequências dos autovalores se sobrepuseram, ficando completamente distintas entre si. Entretanto, a medida que as incertezas relativas cresceram, as larguras das distribuições também aumentaram (FIGURAS 31 - 34). Embora nas simulações deste exemplo não se tenha encontrado em qual ordem de grandeza as incertezas dos dados devem estar para atingir o limite de aplicabilidade do método, pode-se entender que se as incertezas continuarem subindo, em algum ponto as distribuições irão se sobrepor.

Conforme a Tabela 11, pode-se perceber que o primeiro par de variáveis canônicas apresentou a maior correlação canônica, $91,5\%$, e com o aumento das incertezas relativas, sua incerteza estimada também aumentou. Para dar continuidade ao caso particular em que os dados possuem incertezas relativas de 10%, os autovetores já redimensionados para que as variâncias das variáveis canônicas sejam unitárias seguem dispostos na Tabela 12.

Tabela 12 – Resultados dos autovetores de transformação linear e suas incertezas.

a1	σ_a1	a2	σ_a2	a3	σ_a3
-0,0022	0,0007	0,0134	0,0029	0,0027	0,0005
-0,0012	0,0004	-0,00106	0,00024	-0,0028	0,0005
-0,00034	0,00011	-0,00089	0,00019	0,00141	0,00023
b1	σ_b1	b2	σ_b2	b3	σ_b3
0,0005	0,0003	-0,01839	0,00018	0,02880	0,00011
0,00302	0,00008	0,00385	0,00006	-0,00168	0,00006
0,00141	0,00019	-0,00665	0,00004	-0,00720	0,00005
0,000259	0,000014	0,001205	0,000007	0,000375	0,000013

Fonte: Do autor.

Ao aplicar os resultados da Tabela 12 conforme as equações (5.29, 5.30), encontram-se os valores dos pares de variáveis canônicas apresentados na Tabela 13.

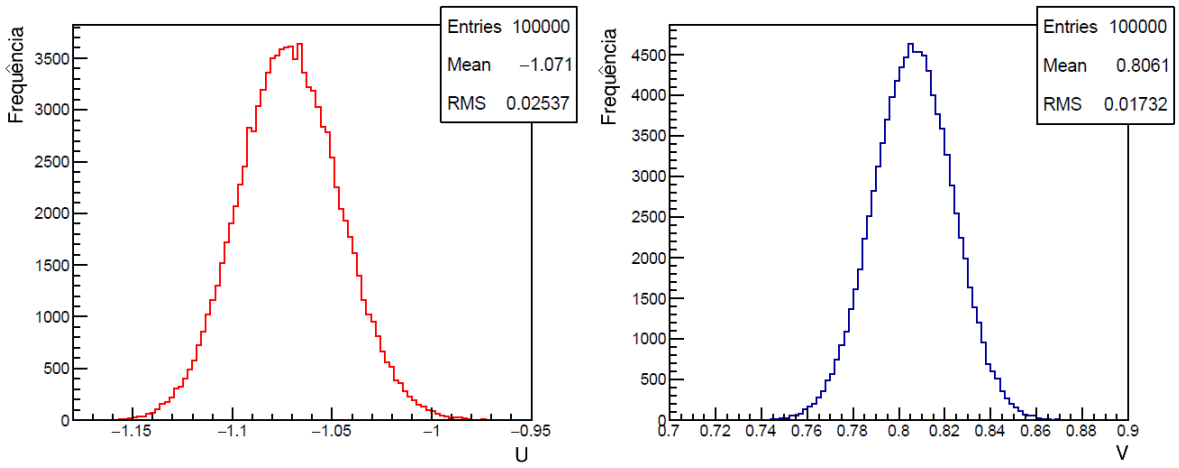
Tabela 13 – Pares de variáveis canônicas das 20 observações da análise realizada com dados com incertezas relativas de 10%.

Obs	U1	σ_U1	V1	σ_V1	U2	σ_U2	V2	σ_V2	U3	σ_U3	V3	σ_V3
1	-1,734	0,025	1,305	0,017	-0,225	0,021	-1,564	0,017	1,41	0,03	1,078	0,018
2	-1,579	0,012	0,849	0,007	-0,190	0,023	0,238	0,012	-0,655	0,027	-0,076	0,011
3	-1,131	0,028	0,765	0,017	-1,497	0,019	-0,617	0,014	0,329	0,017	-0,960	0,010
4	-1,001	0,017	0,936	0,017	-0,673	0,021	0,250	0,011	-0,746	0,016	0,853	0,013
5	-0,388	0,017	0,996	0,019	-0,405	0,004	-0,829	0,018	0,867	0,006	0,806	0,020
6	-1,047	0,029	1,59	0,03	1,236	0,023	-1,272	0,028	-0,978	0,021	-1,717	0,020
7	-1,734	0,025	1,860	0,018	-0,443	0,025	1,361	0,029	1,33	0,03	-1,428	0,027
8	-1,152	0,013	0,588	0,023	-0,293	0,017	-1,318	0,009	0,639	0,020	-0,156	0,009
9	-1,17	0,04	1,201	0,017	2,065	0,016	-1,447	0,016	1,498	0,026	0,675	0,016
10	-1,466	0,028	1,076	0,014	1,409	0,024	-0,580	0,015	0,766	0,029	-0,229	0,013
11	-1,351	0,018	1,376	0,021	-0,935	0,023	-1,623	0,019	0,319	0,020	0,383	0,017
12	-1,594	0,026	1,576	0,026	-0,43	0,03	-1,057	0,024	-1,484	0,027	-0,906	0,020
13	-1,766	0,018	1,108	0,013	-0,87	0,03	0,888	0,017	-0,542	0,028	-1,024	0,019
14	-1,318	0,016	0,957	0,018	-0,615	0,022	0,804	0,016	0,600	0,022	1,250	0,018
15	-1,49	0,03	0,80	0,04	1,764	0,028	-2,210	0,012	-0,139	0,022	-0,600	0,016
16	-1,111	0,025	0,98	0,03	-0,242	0,021	-0,415	0,016	-1,420	0,019	-1,006	0,023
17	-1,946	0,008	1,310	0,021	-0,42	0,03	0,384	0,018	-0,026	0,007	0,357	0,020
18	-0,492	0,013	0,394	0,019	0,084	0,006	0,088	0,012	0,738	0,009	1,950	0,014
19	-1,12	0,04	0,699	0,023	-1,170	0,013	-0,240	0,011	2,174	0,018	-0,923	0,015
20	-1,655	0,019	0,71	0,03	1,05	0,03	-1,799	0,011	0,102	0,020	-0,241	0,014

Fonte: Do autor.

Como exemplo, seguem as distribuições de frequência (Figura 35) do primeiro par de variáveis canônicas para a primeira observação.

Figura 35 – Distribuições de frequência do primeiro par de variáveis canônicas (U_1 e V_1) das variáveis X_1 (vermelho) e Y_1 (azul).



Fonte: Do autor.

Já as correlações entre cada variável canônica e cada grupo de variáveis originais foram calculadas conforme as equações (5.68, 5.70, 5.74, 5.72), e estão dispostas na Tabela abaixo.

Tabela 14 – Correlações entre as variáveis canônicas e cada uma das variáveis originais dos dois grupos de dados $\boldsymbol{X}$ e $\boldsymbol{Y}$, bem como suas incertezas estimadas.

Variável	U1	σ_U1	U2	σ_U2	U3	σ_U3	V1	σ_V1	V2	σ_V2	V3	σ_V3
X1	-82,8	27,1	54,3	11,7	13,8	2,4	75,8	0,5	-27,1	0,3	4,47	0,14
X2	-95,2	31,3	-9,4	2,4	-29,1	4,8	87,1	0,3	4,7	0,4	-9,38	0,22
X3	-86,4	28,3	-23,2	5,2	44,6	7,4	79,1	0,3	11,6	0,3	14,39	0,15
Y1	-72,2	23,7	17,6	3,8	16,2	2,7	78,9	0,7	-35,3	0,4	50,2	0,3
Y2	-88,4	29,0	-0,3	0,5	0,3	0,3	96,62	0,21	0,6	0,5	0,8	0,4
Y3	-78,2	25,7	17,3	3,7	-8,3	1,4	85,4	1,2	-34,7	0,6	-25,8	0,5
Y4	-65,3	21,4	-18,5	4,1	5,9	1,0	71,3	0,4	37,1	0,4	18,2	0,6

Fonte: Do autor.
Nota: Todos os valores estão em porcentagem.

Como pode ser obsevado na Tabela 14 o primeiro par de variáveis canônicas apresentou as maiores correlações com as variáveis $\boldsymbol{X}$ e $\boldsymbol{Y}$. Entretanto, também apresentaram as maiores incertezas. Isso mostra que apesar dos dados apresentarem uma alta correlação, pode não ser seguro utilizar esses pares para estimar valores de conjunto de variáveis. Com os dados da Tabela 14 também é possível calcular a proporção explicativa da variância

total nos casos de se escolher um, dois ou os três pares, como pode ser visto na Tabela 15.

Tabela 15 – Proporções explicativas em termos de variância total, para cada variável canônica individual.

Variável	R2X	σ_R2X	R2Y	σ_R2Y
1	78,0	0,6	69,9	0,4
2	11,93	0,11	9,57	0,07
3	10,10	0,07	8,79	0,04

6 Fonte: Do autor.
Nota: Todos os valores, com excessão da coluna 1, estão em porcentagem.

O que pode ser observado na Tabela 15 é que a primeira variável canônica referente ao grupo $\boldsymbol{X}$ explica, sozinha, $78,0\%$ de toda a variabilidade do grupo. Já a primeira variável canônica do grupo $\boldsymbol{Y}$ explica $69,9\%$. Ao utilizar mais variáveis esse percentual explicativo aumentaria, mas mesmo que essa proporção chegasse a 100%, devido as incertezas experimentais das observações, esse conjunto pode de fato explicar um percentual menor.

Assim, o que este método traz de novo é a possibilidade de correlações que, pelo método clássico, podiam ter uma alta relevância, e agora podem apresentar um valor menos expressivo, e vice e versa, dependendo das suas incertezas que são propagadas das incertezas experimentais das variáveis originais. Isso permite ao pesquisador uma avaliação mais realista dos dados coletados em seus experimentos, diminuindo as chances de superestimar ou subestimar suas conclusões.

Outro ponto é que ao buscar estimar variáveis de um grupo a partir do outro, utilizando, por exemplo, uma regressão linear, este método permite realizar uma regressão ponderada, uma vez que tanto as variáveis canônicas, coeficientes de transformação quanto os dados originais possuem incertezas.

7 CONSIDERAÇÕES FINAIS

O que pôde ser observado com este estudo é que a inserção das incertezas experimentais nas ferramentas de estatística multivariada abordadas gerou a possibilidade de interpretações mais condizentes com a realidade, uma vez que a qualidade da análise está diretamente ligada à qualidade das medidas realizadas em um experimento. Diferentemente da teoria clássica, estas novas formulações consideram diferentes graus de importância para os dados de acordo com suas incertezas.

No estudo das Componentes Principais o método tradicional busca determinar um número reduzido de componentes que explicam quase toda a variabilidade do problema. Entretanto, dado que não se leva em conta as incertezas de cada medida, a avaliação do poder explicativo das componentes selecionadas pode ser superestimada ou subestimada, prejudicando assim as interpretações e futuras análises com as componentes principais.

Outro fato a se destacar é que este método trouxe uma nova maneira de selecionar o número de componentes a ser utilizado. Já que todas as informações do método passaram a ter incertezas, inclusive as proporções explicativas acumuladas, foi possível calcular uma incerteza relativa à proporção cumulativa e demonstrar que ela decai até estabilizar em um determinado valor. Desta forma, uma escolha de componentes baseada apenas nas proporções explicativas, como no método tradicional, pode resultar em um poder explicativo com uma incerteza relativa grande o suficiente para prejudicar a confiabilidade da análise. Com isso o pesquisador pode estipular um valor mínimo para essa incerteza relativa e selecionar o número de componentes que satisfazer a condição.

A análise discriminante linear de Fisher foi a única ferramenta abordada neste estudo, que teve uma mudança de interpretação com a aplicação desta nova formulação. Como visto, o método clássico consiste em, a partir de informações de uma amostra de treinamento, construir uma regra de classificação que seja capaz de avaliar novas observações e determinar a qual grupo ou população elas pertencem. Contudo esse procedimento não leva em consideração as incertezas intrínsecas à coleta dos dados e pode em alguns casos classificar uma observação em um grupo erroneamente, causando prejuízos aos resultados da análise.

Neste trabalho foi considerado uma nova abordagem, pois uma vez que se leva em consideração que as informações da amostra de treinamento possuem uma determinada probabilidade de serem aquele valor observado, ou seja, possuem um limite de confiança, perde-se o sentido de uma classificação determinística, pois, embora o resultado aponte que ela pertence à uma população, devido à sua incerteza, ela pode de fato pertencer à outra população. Assim, esta nova análise desenvolvida neste trabalho, não determina a qual grupo a nova observação pertence, mas qual é a probabilidade dela pertencer a cada um dos grupos existentes.

Portanto, fica a critério do pesquisador avaliar a qual grupo aquela observação pertence. Porém, esta nova formulação diminui a possibilidade de se cometer um erro em uma classificação, pois, uma vez que o pesquisador conhece as chances de erro, ele pode estipular um valor de corte classificando as observações em um determinado grupo cujas probabilidades delas pertencerem a este grupo seja igual ou superior ao valor de corte. Outro ponto diferencial deste método, é que o aplicador tem a possibilidade de avaliar mais cuidadosamente as observações que estiverem próximas ao limite de pertencer a uma população ou a outra, evitando, assim, a possibilidade de uma classificação precipitada.

Já para análise de correlação canônica, foi possível concluir que o impacto desta metodologia, diferentemente da análise clássica, proporcionou a possibilidade de avaliar a qualidade das correlações e com isso poder ter uma melhor interpretação das possíveis predições de um grupo de variáveis a partir do outro. Isso ocorre pois, mesmo que uma análise gere correlações altas, o que seria muito satisfatório para o método tradicional, agora, com as incertezas dessas correlações, pode ser que esses valores não sejam tão confiáveis. Assim, o pesquisador tem a possibilidade de reavaliar a qualidade das suas observações medidas. Outro ponto, é que este método permite realizar análises de regressões lineares ponderadas com as incertezas, com o intuito de estimar os valores de um grupo de variáveis a partir do outro. Com isso, o método torna-se mais sensível à qualidade dos dados.

Por fim, pode-se concluir que a utilização das incertezas experimentais em análises estatísticas proporciona uma interpretação de qualidade e mais condizente com a realidade do experimento. Como continuidade deste estudo, é possível buscar a inserção de incertezas experimentais em outras ferramentas e também buscar aplicações das ferramentas abordadas neste estudo, não só na Física, mas nas mais diversas áreas da ciência.

REFERÊNCIAS

AERTS, S.; WILMS, I. Cellwise robust regularized discriminant analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 10, n. 6, p. 436–447, 2017.

AKTEKIN, T.; POLSON, N.; SOYER, R. Sequential bayesian analysis of multivariate count data. *Bayesian Analysis*, International Society for Bayesian Analysis, p. 1–25, 2017.

AL-SAYED, A. Principal component analysis within nuclear structure. *Nuclear Physics A*, Elsevier, v. 933, p. 154–164, 2015.

AMBIKASARAN, S. et al. Fast direct methods for gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 38, n. 2, p. 252–265, 2016.

ANDRIOTTI, J. L. S. *Fundamentos de estatística e geoestatística*. São Leopoldo: Universidade do Vale do Rio dos Sinos, 2003.

ANTCHEVA, I. et al. Root—a c++ framework for petabyte data storage, statistical analysis and visualization. *Computer Physics Communications*, Elsevier, v. 182, n. 6, p. 1384–1385, 2011.

BALAKRISHNAN, N.; LAI, C. *Continuous bivariate distributions*. 2. ed. New York: Springer, 2009.

BENTON, A. et al. Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519*, 2017.

BODNAR, T. et al. Discriminant analysis in small and large dimensions. *arXiv preprint arXiv:1705.02826*, 2017.

BOHLIN, R.; SAVAGE, B.; DRAKE, J. A survey of interstellar hi from l-alpha absorption measurements. ii. *The Astrophysical Journal*, v. 224, p. 132–142, 1978.

BOYER, C. B.; MERZBACH, U. C. *História da matemática*. 3. ed. São Paulo: Editora Blucher, 2012.

CASTRO, L. S. V. d. *Pontos de estatística*. 15. ed. Rio de Janeiro: Científica, 1970.

CHESPERITO. *Chaves HD - Bombinhas são perigosas, ainda mais em mãos erradas (1976)*. 1976. Disponível em: <https://www.youtube.com/watch?v=tOEuULLRBkE>. Acesso em: 20 fev. 2018.

CHU, D. et al. Sparse kernel canonical correlation analysis. In: *PROCEEDINGS OF INTERNATIONAL MULTICONFERENCE OF ENGINEERS AND COMPUTER SCIENTISTS*. Hong Kong: IMECS, 2013. v. 1.

COWAN, G. *Statistical data analysis*. New York: Oxford University Press, 1998.

ENSOR, T. et al. A principal component analysis of the diffuse interstellar bands. *The Astrophysical Journal*, IOP Publishing, v. 836, n. 2, p. 162–186, 2017.

ESFAHANI, E. N.; LIU, X.; LI, J. Imaging ferroelectric domains via charge gradient microscopy enhanced by principal component analysis. *arXiv preprint arXiv:1706.02345*, 2017.

FÁVERO, L. P. et al. *Análise de dados: modelagem multivariada para tomada de decisões*. Rio de Janeiro: Elsevier, 2009.

FERREIRA, D. F. *Estatística multivariada*. Lavras: Editora UFLA, 2008.

FERRER, L. Joint probabilistic linear discriminant analysis. *arXiv preprint arXiv:1704.02346*, 2017.

GAO, C. et al. Stochastic canonical correlation analysis. *arXiv preprint arXiv:1702.06533*, 2017.

GRATIER, P. et al. Dissecting the molecular structure of the orion b cloud: insight from principal component analysis. *Astronomy & Astrophysics*, EDP Sciences, v. 599, p. A100, 2017.

HAIR, J. F. et al. *Análise multivariada de dados*. 6. ed. Porto Alegre: Editora Bookman, 2009.

HAMILL, J. et al. Fast data sorting with modified principal component analysis to distinguish unique single molecular break junction trajectories. *Physical review letters*, APS, v. 120, n. 1, p. 016601(5), 2018.

HAN, X.; PAN, G.; YANG, Q. A unified matrix model including both cca and f matrices in multivariate analysis: the largest eigenvalue and its applications. *arXiv preprint arXiv:1606.04417*, 2016.

HÄRDLE, W.; SIMAR, L. *Applied multivariate statistical analysis*. 4. ed. London: Springer Science & Business Media, 2015.

HAUBOLD, H.; MATHAI, A.; THOMAS, S. A pathway to multivariate gaussian density. *Mathematica Aeterna Journal*, v. 2, p. 51 – 61, 2007.

HENZE, N.; KOCH, S. On a test of normality based on the empirical moment generating function. *Statistical Papers*, p. 1–13, 2016.

HOTELLING, H. Relations between two sets of variates. *Biometrika*, v. 28, n. 3/4, p. 321–377, 1936.

HUBERT, M.; DRIESSEN, K. V. Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, v. 45, n. 2, p. 301 – 320, 2004.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 6. ed. New Jersey: Pearson Prentice-Hall, 2007.

JONES, C. L. et al. Canonical correlation analysis for analyzing sequences of medical billing codes. *arXiv preprint arXiv:1612.00516*, 2016.

LEE, S. High-dimension, low sample size asymptotics of canonical correlation analysis. *arXiv preprint arXiv:1609.02992*, 2016.

LU, L.; RENALS, S. Probabilistic linear discriminant analysis for acoustic modeling. *IEEE Signal Processing Letters*, v. 21, n. 6, p. 702–706, 2014.

MAGALHÃES, M. N.; LIMA, A. C. P. de. *Noções de probabilidade e estatística*. 6. ed. São Paulo: Editora da Universidade de São Paulo, 2008.

MAHMOUDI, E.; MAHMOODIAN, H. A new bivariate distribution obtained by compounding the bivariate normal and geometric distributions. *Journal of Statistical Theory and Applications*, v. 16, n. 2, p. 198–208, 2017.

MAJUMDAR, R.; MAJUMDAR, S. On the regular conditional distribution of a multivariate normal given a linear transformation. *arXiv preprint arXiv:1612.01210*, 2016.

MIN, W.; LIU, J.; ZHANG, S. Sparse weighted canonical correlation analysis. *arXiv preprint arXiv:1710.04792*, 2017.

MINGOTI, S. A. *Análise de dados através de métodos de estatística Multivariada: Uma abordagem aplicada*. Belo Horizonte: Editora UFMG, 2005.

MURSULA, K.; HOLAPPA, L. Principal component analysis of geomagnetic activity: New information on solar wind. *arXiv preprint arXiv:1709.06885*, 2017.

RAGHU, M. et al. Svcca: Singular vector canonical correlation analysis for deep understanding and improvement. In: *Conference on Advances in Neural Information Processing Systems*. Long Beach: [s.n.], 2017. p. 6072–6081.

SIRUNYAN, A. M. et al. Principal-component analysis of two-particle azimuthal correlations in pbpb and p pb collisions at cms. *Physical Review C*, APS, v. 96, n. 6, p. 064902, 2017.

SIZOV, A.; LEE, K. A.; KINNUNEN, T. Unifying probabilistic linear discriminant analysis variants in biometric authentication. In: SPRINGER. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Berlin, 2014. p. 464–475.

SUO, X. et al. Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865*, 2017.

TANG, Y.; LI, W. lfda: An r package for local fisher discriminant analysis and visualization. *arXiv preprint arXiv:1612.09219*, 2016.

TAYLOR, J. R. *Introdução à análise de erros: o estudo de incertezas em medições físicas*. 2. ed. Porto Alegre: Editora Bookman, 2012.

THULIN, M. Tests for multivariate normality based on canonical correlations. *Statistical Methods & Applications*, v. 23, n. 2, p. 189 – 208, 2014.

VIALI, L. Algumas considerações sobre a origem da teoria da probabilidade. *Revista Brasileira de História da Matemática*, v. 8, n. 16, p. 143–153, 2008.

VUOLO, J. H. *Fundamentos da teoria dos erros*. 2. ed. São Paulo: Editora Blucher, 1996.

YANG, C. et al. Network representation learning with rich text information. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. Buenos Aires, 2015. p. 2111–2117.

ZHENG, D. et al. Main and interaction effects selection for quadratic discriminant analysis via penalized linear regression. *arXiv preprint arXiv:1702.04570*, 2017.

APÊNDICES

APÊNDICE A – PCs das 29 estrelas para a análise com duas variáveis

Obs	Estrela	PC1(σ)	PC2(σ)
1	HD 15137	0,6138670(2,5)	0,2952670(1,5)
2	HD 22951	0,3457140(9)	0,577078(4)
3	HD 23180	0,4521700(1,8)	0,3535560(3)
4	HD 23630	-1,1002900(2,1)	0,670972(4)
5	HD 24398	0,732030(3)	0,0834509(3)
6	HD 24534	1,263220(8)	0,1746660(1,0)
7	HD 24760	-0,977619(6)	0,573661(4)
8	HD 24912	0,8731200(2,4)	0,3345330(2,4)
9	HD 27778	0,972624(5)	-0,4459060(2,4)
10	HD 35149	-0,864090(6)	0,577198(3)
11	HD 35715	-1,2005300(2,8)	0,7907210(2,6)
12	HD 36822	-0,794216(4)	0,757065(3)
13	HD 36861	-0,6614190(2,1)	0,559882(5)
14	HD 40111	-0,5287440(2,0)	0,692557(4)
15	HD 110432	1,400770(10)	-0,5677270(1,0)
16	HD 144217	0,1395070(5)	0,4808640(3)
17	HD 145502	0,2299890(1,4)	0,3513600(9)
18	HD 147165	1,316870(7)	0,2283210(1,8)
19	HD 147933	2,963850(1,7)	1,21534(5)
20	HD 149757	0,7678810(2,2)	-0,10068400(2,5)
21	HD 164284	-0,6376410(8)	0,4736680(2,2)
22	HD 170740	1,801350(7)	-0,0571481(4)
23	HD 198478	2,566060(8)	0,1575960(9)
24	HD 202904	-0,895917(6)	0,435377(3)
25	HD 207198	3,38114(3)	0,532697(4)
26	HD 209975	0,7144700(1,7)	0,0658909(3)
27	HD 214680	-0,816288(4)	0,625000(5)
28	HD 214993	-0,861894(4)	0,799379(3)
29	HD 218376	-0,0651149(3)	0,4962290(2,2)

APÊNDICE B – PCs das 29 estrelas para a análise com 23 variáveis

Componente	HD 15137	σ	HD 22951	σ	HD 23180	σ	HD 23630	σ	HD 24398	σ
PC1	-0,015	0,022	0,017	0,009	2,655	0,014	-2,909	0,012	2,725	0,012
PC2	-1,59	0,14	1,27	0,12	7,38	0,17	-1,44	0,09	5,84	0,17
PC3	2,1	0,3	2,03	0,16	3,0	0,7	1,12	0,21	3,09	0,54
PC4	1,02	0,10	-0,93	0,07	-4,50	0,13	1,77	0,06	-4,72	0,16
PC5	-7,01	0,11	-2,98	0,07	-5,29	0,13	-0,17	0,06	-2,31	0,13
PC6	-0,33	0,18	0,44	0,18	1,5	0,8	-1,00	0,42	1,20	0,72
PC7	0,61	0,20	0,50	0,12	5,8	0,4	-5,03	0,50	5,45	0,40
PC8	1,06	0,19	-0,19	0,09	0,9	0,6	-3,24	0,56	1,64	0,56
PC9	1,14	0,22	0,63	0,04	1,81	0,20	-1,05	0,16	0,99	0,15
PC10	1,50	0,12	0,46	0,04	1,3	0,3	-1,22	0,19	0,95	0,24
PC11	0,39	0,14	-1,34	0,03	-6,82	0,08	1,25	0,10	-5,13	0,05
PC12	2,29	0,14	0,23	0,06	2,41	0,27	-3,96	0,16	2,04	0,17
PC13	2,85	0,07	-0,57	0,03	-3,10	0,27	-2,82	0,08	-1,30	0,16
PC14	1,46	0,23	-0,98	0,07	-2,72	0,28	-3,19	0,35	-1,15	0,18
PC15	-1,77	0,24	-0,25	0,05	0,68	0,05	0,93	0,56	-0,33	0,16
PC16	2,38	0,18	0,53	0,06	1,12	0,11	-3,22	0,34	2,04	0,15
PC17	0,11	0,03	-1,155	0,021	-1,25	0,05	-3,49	0,07	-0,256	0,015
PC18	0,10	0,14	0,11	0,23	1,0	0,5	-0,91	0,23	0,79	0,48
PC19	1,4	0,3	0,92	0,22	0,22	0,4	0,87	0,33	0,84	0,34
PC20	-1,15	0,17	0,34	0,12	1,59	0,23	0,74	0,18	1,07	0,22
PC21	0,72	0,12	0,54	0,08	-0,44	0,12	1,51	0,16	-0,34	0,03
PC22	0,31	0,14	-0,002	0,012	-1,20	0,14	0,16	0,31	-0,40	0,10
PC23	0,24	0,05	0,36	0,07	0,54	0,11	0,59	0,12	0,33	0,07

Componente	HD 24534	σ	HD 24760	σ	HD 24912	σ	HD 27778	σ	HD 35149	σ
PC1	4,020	0,015	-2,946	0,013	1,027	0,005	3,808	0,021	-3,064	0,016
PC2	7,65	0,20	-2,35	0,11	0,44	0,07	6,01	0,15	-2,97	0,10
PC3	3,83	0,68	1,34	0,25	1,41	0,08	2,99	0,60	0,91	0,34
PC4	-5,98	0,20	2,43	0,07	-1,24	0,10	-6,19	0,31	3,17	0,13
PC5	-2,38	0,16	-3,31	0,08	1,00	0,05	5,45	0,14	-5,59	0,09
PC6	1,42	0,95	-0,98	0,35	-0,16	0,10	0,83	0,59	-0,80	0,30
PC7	7,37	0,49	-4,51	0,38	0,85	0,07	5,43	0,32	-4,08	0,31
PC8	2,03	0,76	-2,44	0,54	0,54	0,09	1,70	0,62	-1,93	0,52
PC9	0,97	0,21	-0,17	0,14	-0,54	0,04	-1,09	0,19	0,63	0,16
PC10	1,26	0,31	-0,42	0,18	0,11	0,04	-0,11	0,23	0,10	0,18
PC11	-6,30	0,07	1,44	0,07	0,12	0,03	-3,67	0,08	1,53	0,07
PC12	2,99	0,20	-2,55	0,11	0,25	0,07	0,64	0,12	-1,39	0,09
PC13	-1,06	0,19	-1,48	0,07	1,24	0,05	-0,12	0,06	-0,86	0,06
PC14	-0,85	0,23	-2,32	0,25	0,77	0,10	0,05	0,16	-1,67	0,19
PC15	-0,63	0,30	0,36	0,43	-1,19	0,18	-1,14	0,33	0,29	0,36
PC16	3,25	0,23	-2,26	0,26	1,60	0,13	3,00	0,22	-1,89	0,21
PC17	0,126	0,022	-3,01	0,06	0,36	0,04	0,93	0,06	-2,52	0,06
PC18	1,00	0,61	-0,70	0,14	-0,11	0,29	0,37	0,74	-0,47	0,11
PC19	1,18	0,43	0,98	0,28	1,26	0,20	1,58	0,32	0,54	0,20
PC20	1,25	0,29	0,12	0,20	-0,24	0,13	1,07	0,29	-0,29	0,18
PC21	-0,58	0,03	1,57	0,15	0,22	0,14	-0,48	0,16	1,37	0,08
PC22	-0,40	0,13	0,11	0,30	0,63	0,07	0,65	0,10	-0,11	0,24
PC23	0,34	0,07	0,58	0,12	0,041	0,016	0,029	0,021	0,50	0,10

Componente	HD 35715	σ	HD 36822	σ	HD 36861	σ	HD 40111	σ	HD 110432	σ
PC1	-4,131	0,014	-2,777	0,011	-1,222	0,017	-2,564	0,019	3,363	0,021
PC2	-4,09	0,07	-1,95	0,06	1,96	0,12	-2,89	0,12	2,51	0,07
PC3	0,32	0,41	0,49	0,23	1,52	0,36	1,30	0,37	1,46	0,43
PC4	4,10	0,11	2,39	0,07	0,22	0,17	3,10	0,12	-4,52	0,31
PC5	-3,11	0,10	-2,53	0,06	-8,92	0,08	-6,01	0,10	8,67	0,10
PC6	-1,20	0,59	-0,51	0,37	0,94	0,31	-0,48	0,28	-0,11	0,29
PC7	-6,60	0,52	-4,40	0,37	0,53	0,25	-3,97	0,31	3,93	0,32
PC8	-3,25	0,76	-2,53	0,51	-0,73	0,23	-2,00	0,53	2,35	0,51
PC9	-0,26	0,20	-0,02	0,09	2,33	0,09	0,59	0,18	-1,83	0,14
PC10	-0,74	0,26	-0,50	0,16	1,06	0,12	0,35	0,18	-0,43	0,15
PC11	2,86	0,09	1,14	0,07	-3,35	0,07	1,54	0,09	-0,38	0,07
PC12	-3,42	0,14	-2,50	0,12	0,94	0,20	-0,95	0,07	0,44	0,16
PC13	-1,64	0,10	-1,92	0,05	-3,06	0,18	0,004	0,025	2,49	0,10
PC14	-2,42	0,31	-2,35	0,25	-3,08	0,25	-1,01	0,16	2,56	0,25
PC15	0,73	0,55	0,87	0,45	1,37	0,34	-0,11	0,25	-2,05	0,51
PC16	-3,38	0,33	-2,65	0,27	-1,45	0,20	-1,03	0,16	3,53	0,32
PC17	-3,34	0,07	-2,76	0,06	-2,59	0,11	-2,27	0,04	2,65	0,12
PC18	-0,95	0,13	-0,59	0,07	0,56	0,32	-0,51	0,12	0,02	0,12
PC19	0,54	0,30	0,40	0,20	-0,47	0,23	1,08	0,27	1,54	0,21
PC20	-0,07	0,22	0,33	0,14	0,90	0,03	-0,41	0,22	-0,09	0,18
PC21	1,74	0,13	1,23	0,08	0,44	0,14	1,46	0,13	-0,57	0,22
PC22	0,11	0,30	-0,11	0,23	-1,33	0,13	0,11	0,26	1,44	0,10
PC23	0,53	0,11	0,48	0,10	0,63	0,13	0,51	0,10	-0,42	0,09

Componente	HD 144217	σ	HD 145502	σ	HD 147165	σ	HD 147933	σ	HD 149757	σ
PC1	-1,410	0,003	-0,6133	0,0021	0,672	0,008	3,513	0,016	2,956	0,015
PC2	-2,73	0,03	-1,44	0,03	-2,363	0,019	2,70	0,07	5,93	0,14
PC3	0,25	0,24	0,43	0,13	-0,08	0,22	-0,43	0,32	2,84	0,55
PC4	1,52	0,05	0,79	0,03	0,39	0,07	-2,54	0,17	-5,33	0,21
PC5	1,59	0,04	0,687	0,024	2,73	0,03	6,70	0,07	1,37	0,13
PC6	-0,95	0,34	-0,64	0,17	-0,58	0,18	-0,03	0,17	0,90	0,67
PC7	-3,05	0,25	-1,57	0,15	-0,02	0,11	3,73	0,23	5,44	0,34
PC8	-1,01	0,34	-0,55	0,18	1,32	0,08	1,53	0,49	1,57	0,59
PC9	-0,95	0,10	-0,66	0,06	-0,84	0,14	-1,26	0,15	-0,04	0,16
PC10	-0,62	0,15	-0,31	0,08	0,14	0,13	-0,01	0,08	0,30	0,24
PC11	2,61	0,05	1,40	0,03	2,54	0,07	-0,59	0,07	-4,61	0,06
PC12	-1,75	0,10	-0,90	0,06	0,65	0,21	1,33	0,14	1,26	0,15
PC13	0,78	0,13	0,74	0,08	3,90	0,17	2,26	0,07	-1,36	0,13
PC14	0,27	0,13	0,32	0,08	3,15	0,30	2,87	0,25	-0,92	0,18
PC15	-0,58	0,08	-0,481	0,016	-1,87	0,38	-1,21	0,50	-0,38	0,18
PC16	-0,50	0,07	0,02	0,03	2,30	0,23	2,83	0,29	1,99	0,14
PC17	-0,525	0,018	-0,260	0,018	2,04	0,09	2,82	0,11	0,199	0,017
PC18	-0,65	0,19	-0,43	0,15	-0,30	0,26	0,09	0,24	0,62	0,56
PC19	0,87	0,22	0,75	0,15	0,99	0,18	0,41	0,11	0,83	0,29
PC20	-0,51	0,15	-0,32	0,11	-1,22	0,10	-0,18	0,12	1,18	0,25
PC21	0,81	0,16	0,52	0,12	0,07	0,16	-1,02	0,06	-0,51	0,05
PC22	0,77	0,15	0,51	0,10	1,16	0,07	0,80	0,19	-0,05	0,10
PC23	0,049	0,022	0,059	0,017	-0,37	0,08	-0,57	0,11	0,18	0,04

Componente	HD 164284	σ	HD 170740	σ	HD 198478	σ	HD 202904	σ	HD 207198	σ
PC1	-2,047	0,008	3,575	0,015	3,768	0,014	-3,432	0,023	6,603	0,017
PC2	-1,56	0,07	2,13	0,07	0,58	0,05	-4,68	0,13	6,36	0,06
PC3	0,86	0,18	1,59	0,32	1,30	0,23	1,36	0,51	0,73	0,58
PC4	1,23	0,06	-3,77	0,24	-2,85	0,18	4,27	0,15	-5,58	0,14
PC5	1,05	0,04	5,97	0,09	3,92	0,08	-6,62	0,12	1,50	0,13
PC6	-0,89	0,36	0,07	0,24	0,29	0,32	-1,00	0,43	1,75	0,90
PC7	-3,94	0,37	3,36	0,28	3,94	0,41	-5,34	0,36	9,32	0,72
PC8	-2,20	0,43	2,10	0,42	3,12	0,43	-1,98	0,70	4,26	1,02
PC9	-1,02	0,11	-1,55	0,11	-0,84	0,18	0,58	0,26	0,51	0,25
PC10	-0,95	0,15	0,12	0,14	0,86	0,18	0,47	0,26	1,77	0,36
PC11	1,55	0,07	-0,32	0,06	0,72	0,11	2,93	0,12	-4,30	0,13
PC12	-3,08	0,10	0,98	0,18	2,72	0,29	-1,29	0,12	5,44	0,24
PC13	-1,42	0,08	3,21	0,10	5,50	0,15	1,02	0,14	3,36	0,14
PC14	-1,90	0,23	2,60	0,29	4,62	0,47	-0,61	0,21	3,52	0,48
PC15	0,27	0,35	-2,34	0,54	-3,19	0,79	-0,63	0,26	-1,85	0,87
PC16	-1,98	0,21	3,94	0,33	5,43	0,48	-1,02	0,16	5,95	0,52
PC17	-2,23	0,03	2,34	0,12	3,56	0,15	-2,42	0,04	4,11	0,12
PC18	-0,75	0,24	-0,05	0,25	0,15	0,45	-0,74	0,18	1,18	0,26
PC19	0,92	0,28	1,95	0,28	2,01	0,29	1,40	0,35	0,42	0,36
PC20	0,31	0,17	-0,33	0,20	-1,20	0,16	-0,86	0,29	-0,07	0,21
PC21	1,19	0,16	-0,34	0,23	-0,48	0,22	1,92	0,20	-1,98	0,09
PC22	0,43	0,24	1,34	0,08	1,53	0,12	0,45	0,33	0,03	0,18
PC23	0,37	0,08	-0,28	0,06	-0,48	0,10	0,52	0,11	-0,51	0,10

Componente	HD 209975	σ	HD 214680	σ	HD 214993	σ	HD 218376	σ
PC1	1,078	0,013	-2,385	0,019	-3,164	0,013	-0,980	0,013
PC2	-0,36	0,07	-0,83	0,13	-2,60	0,09	-0,89	0,12
PC3	1,15	0,19	1,47	0,32	0,96	0,28	1,67	0,19
PC4	-0,36	0,05	1,94	0,16	2,77	0,08	0,82	0,07
PC5	-3,64	0,06	-8,28	0,09	-3,31	0,08	-4,51	0,07
PC6	0,60	0,17	0,17	0,10	-0,8	0,4	-0,27	0,04
PC7	1,35	0,23	-2,14	0,22	-5,0	0,4	-1,30	0,11
PC8	1,28	0,13	-1,42	0,38	-2,63	0,59	-0,51	0,23
PC9	0,61	0,14	1,56	0,13	0,002	0,013	0,58	0,11
PC10	1,08	0,08	0,70	0,11	-0,44	0,19	0,54	0,08
PC11	-0,09	0,09	-0,88	0,07	1,61	0,07	0,06	0,06
PC12	2,11	0,12	-0,21	0,12	-2,66	0,12	-0,10	0,03
PC13	2,43	0,04	-1,80	0,07	-1,70	0,07	0,035	0,024
PC14	1,55	0,20	-2,46	0,19	-2,33	0,27	-0,87	0,08
PC15	-1,36	0,29	0,8	0,4	0,50	0,46	-0,48	0,11
PC16	2,39	0,18	-1,67	0,22	-2,61	0,27	-0,01	0,04
PC17	0,75	0,04	-2,80	0,09	-3,08	0,06	-1,447	0,024
PC18	0,22	0,10	0,01	0,03	-0,68	0,09	-0,17	0,12
PC19	0,96	0,21	0,21	0,18	0,75	0,27	1,02	0,23
PC20	-0,63	0,09	0,30	0,10	0,12	0,19	-0,21	0,15
PC21	0,06	0,07	1,09	0,04	1,53	0,13	0,92	0,11
PC22	0,23	0,05	-0,71	0,21	0,06	0,24	0,07	0,18
PC23	0,054	0,015	0,64	0,13	0,55	0,11	0,41	0,08