1 2 9 0

UNIVERSIDADE D
COIMBRA

Jessica Pastilha Santos

Métodos Quimiométricos em Ciências do Desporto

Aplicação nas modalidades de natação e futebol

Dissertação no âmbito do Mestrado em Química – Controlo da Qualidade e Ambiente,
orientada pelo Professor Doutor Alberto António Caria Canelas Pais,
coorientada pelo Professor Doutor Luís Manuel Pinto Lopes Rama
e apresentada ao Departamento de Química da Faculdade de Ciências e Tecnologia
da Universidade de Coimbra.

Junho de 2020

Faculdade de Ciências e Tecnologia da

Universidade de Coimbra

Métodos Quimiométricos em Ciências do Desporto

Aplicação nas modalidades de natação e futebol

Jessica Pastilha Santos

Dissertação no âmbito do Mestrado em Química – Controlo da Qualidade e Ambiente, orientada pelo Professor Doutor Alberto António Caria Canelas Pais, coorientada pelo Professor Doutor Luís Manuel Pinto Lopes Rama e apresentada ao Departamento de Química da Faculdade de Ciências e Tecnologias da Universidade de Coimbra.

Junho de 2020

1 2 9 0

UNIVERSIDADE D
COIMBRA

Agradecimentos

Termina um lindo percurso de 5 anos no Departamento de Química da Universidade de Coimbra. Cidade do estudante que me fez crescer a nível pessoal e estudantil. Chego ao fim com o sentimento de dever cumprido e com uma enorme gratidão pelas pessoas que me acompanharam nesta caminhada.

Em primeiro lugar gostava de agradecer ao Professor Doutor Alberto Canelas Pais, por toda a orientação, apoio, ensinamentos, conhecimento, disponibilidade ao longo deste percurso e também por ter proporcionado momentos de descontração.

Ao Professor Doutor Luís Rama da Faculdade Ciências do Desporto e Educação Física da Universidade de Coimbra, por ter aceite o convite de participar neste projeto e por ter sido incansável com o seu conhecimento, disponibilidade e ajuda.

Às Doutoras Tânia Cova e Sandra Nunes, que me acolheram de forma exemplar no grupo de investigação e por terem estado sempre disponíveis para me ajudar.

Aos estudantes que efetuaram as recolhas dos dados nos seus projetos doutorais e laboratório integral pertencentes à Faculdade de Ciências do Desporto e Educação Física da Universidade de Coimbra.

Ao meu grupo de amigas "Agora é que é" e à casa do Túlio, que no meio dos estudos e dos meus treinos de futebol, conseguiram estar sempre comigo, proporcionando bons momentos de alegria, diversão e partilha. Elas foram também um grande pilar de incentivo e força.

Por fim e não menos importante, agradecer aos meus pais e à minha irmã, que sempre acreditaram em mim e me incentivaram a seguir os meus sonhos. Eles de facto são o grande motor de todas as minhas realizações.

A todos, muito obrigada!

Resumo

A quimiometria é uma área da química cujas técnicas e metodologias têm uma aplicação muito ampla. O tratamento de dados através dos métodos quimiométricos permite uma análise que combina profundidade com clareza, após uma seleção judiciosa dos métodos a aplicar.

Técnicas como a análise hierárquica de agrupamentos (HCA) e análise de componentes principais (PCA) são extremamente úteis no tratamento de dados multivariados, permitindo revelar a sua estrutura, muitas vezes recorrendo a representações gráficas adequadas, e sendo adequados mesmo quando a informação é muito vasta.

Nesta dissertação, utilizaremos técnicas quimiométricas para a análise de conjuntos de dados bioquímicos e cineantropométricos relativos a atletas de escalões jovens nas modalidades de natação e futebol.

Os resultados mostram que a conjugação dos métodos PCA e HCA permitem, entre vários outros aspetos, mapear o desempenho em função de características cineantropométricas dos atletas da modalidade de natação, e promover uma representação gráfica muito útil em que o conjunto de todos os atletas pode surgir a duas dimensões. Fez-se também a análise de parâmetros bioquímicos em relacionamento com cargas de treino aplicadas a atletas desta mesma modalidade, tendo sido patente a dificuldade de apresentar padrões bem definidos ou relações claras. Em contraste, atletas de escalões jovens do futebol, mostram perfis de indicadores bioquímicos mais regulares que permitem estabelecer, com facilidade, as caraterísticas e comportamentos em cada escalão.

Palavras Chave: Quimiometria, HCA, PCA, cineantropometria, marcadores biológicos salivares, natação, futebol.

Abstract

Chemometrics is an area of chemistry whose techniques and methodologies have a very wide application. The treatment of data through chemometric methods allows an analysis that combines depth with clarity, through a judicious selection of methods to be applied.

Techniques such as HCA and PCA are extremely useful in the treatment of multivariate data, allowing to reveal their structure, often using appropriate graphical representations, and being adequate even when the information is very wide.

In this dissertation, we will use chemometrics techniques for the analysis of biochemical and kinanthropometry data sets related to young athletes in swimming and football.

The results show that the combination of PCA and HCA methods allows, among several other aspects, to map the performance in function of kinanthropometry characteristics from the athletes that practice swimming as modality, and to promote a very useful graphical representation in which the set of all athletes can appear in two dimensions. The analysis of biochemical parameters was also carried out in relation to training the overload applied to athletes of this same modality, and the difficulty of presenting well-defined patterns or clear relationships was evident. In contrast, young footballers show more regular biochemical indicator profiles that allowed to easily establish the characteristics and behaviours in each class.

Keywords: Chemometric, HCA, PCA, Kinanthropometry, bio markers, swimming, football.

Abreviaturas

AS – Altura Sentado

CM – Comprimento da Mão

CP – Comprimento do Pé

CMI – Comprimento do Membro Inferior

CMS – Comprimento do Membro Superior

DBC – Diâmetro Biacromial

DBC – Diâmetro Bicristal

Desl - Deslise

DT – Diâmetro Toracosagital

Env – Envergadura

Est - Estatura

ExtOm – Extensão do Ombro

ExtPl – Extensão Plantar

ExtTr – Extensão do Tronco

FlexOm – Flexão do Ombro

FlexPl – Flexão Plantar

FlexTr – Flexão do Tronco

HCA – Análise Hierárquica de Agrupamentos

Id - Idade

Ig – Imunoglobulina

IgA – Imunoglobulina A

IMD – Índice Morfodiamico

INT30 – Índice de nado T30

ITRS – Infeções do Trato Respiratório Superior

MS – Massa Corporal

PC – Principal Component

PCA – Principal Component Analysis

PrMD – Força de Preensão Mão Direita

PrME – Força de Preensão Mão Esquerda

SK – Somas das 6 Pregas Subcutâneas

TMB – Tetrametilbensidina

Velmax – Velocidade Máxima

VelT30 – Velocidade T30

Índice

	Pág./Page
Agradecimentos	iii
Resumo	v
Abstract	iiv
Abreviaturas	ix
Índice	xi
Índice de Figuras	xiii
Índice de Tabelas	xvi
Capítulo I - Introdução	1
Capítulo II – Fundamentação Teórica	7
2.1 Quimiometria	9
2.2 Métodos Quimiométricos	9
2.2.1 Análise Multivariada	9
2.2.2 Análise de Hierárquica Agrupamento (HCA)	10
2.2.3 Análise de Componentes Principais (PCA)	12
2.2.4 Aplicação do PCA	13
2.2.4.1 Escolha da Matriz	14
2.2.4.2 *Scores*	16
2.2.4.3 Critério de seleção do número de componentes principais	16
Capítulo III – Aspetos Práticos e de Software	19
3.1 RStudio	21
3.2 Implementação das linhas de código	23
Capítulo IV – Conjunto de Dados	33
4.1 Base de Dados	35
4.1.1 Características Cineantropométricas em Nadadores	36
4.1.2 Marcadores Biológicos Salivares em Nadadores	38
4.1.3 Marcadores Biológicos Salivares em Futebolistas	38
4.2 Metodologia de Recolha de Dados	38
Capítulo V – Resultados e Discussão	41
5.1 Características Cineantropométricas em Jovens Nadadores (dados 1)	43
5.2 Marcadores Biológicos Salivares em Nadadores durante uma época competitiva (dados 2)	53
5.3 Marcadores Biológicos Salivares em jovens atletas de Futebol (dados 3)	59
Capítulo VI – Conclusão	69
Referências Bibliográficas	73
Anexos	83

Índice de Figura

	Pág.
Figura 1.1. Estrutura química do cortisol.	**4**
Figura 1.2. Estrutura química da testosterona.	**4**
Figura 1.3. Representação da Imunoglobulina A. (adaptado [26]).	**5**
Figura 2.1. Exemplo de um dendrograma.	**10**
Figura 3.1. Representação do painel de trabalho do RStudio.	**22**
Figura 3.2. Ilustração do funcionamento dos packages FactoMineR e factoextra.	**23**
Figura 3.3. Visualização gráfica do método *SreePlot*. Neste caso, deveriam ser selecionadas 5 componentes principais.	**27**
Figura 3.4. Representação gráfica da contribuição das variáveis nas componentes principais.	**28**
Figura 3.5. Exemplo da representação gráfica a duas dimensões dos resultados do PCA.	**29**
Figura 3.6. Exemplo da representação do dendrograma do package factoextra.	**30**
Figura 3.7. Três exemplos de apresentação do dendrograma: 1) formato retangular, 2) formato circular, 3) formato ramificado.	**31**
Figura 5.1. Representação gráfica do critério do *ScreePlot* aplicado aos dados 1.	**44**
Figura 5.2. Representação dos objetos (atletas) num sistema de eixos a duas dimensões. O eixo PC1 representa a primeira componente com recuperação de 44.3% e o eixo PC2 representa a segunda componente principal com recuperação de 9.5%.	**45**
Figura 5.3. Representação gráfica da posição das variáveis em relação às duas primeiras componentes principais.	**45**
Figura 5.4. Relação entre as variáveis originais e as duas primeiras componentes principais.	**46**

Figura 5.5. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento dos atletas de natação pura desportiva. As diferentes cores representam os diferentes grupos formados, com base na semelhança dos atletas. **47**

Figura 5.6. Representação gráfica dos grupos de atletas formados dendrograma num eixo de a duas dimensões aplicado aos nadadores, com uma recuperação de 53.8% da informação inicial. As cores dos grupos representam a partição da amostra nos mesmo grupos formados pelo HCA e representados no dendrograma da Figura 5.5. **49**

Figura 5.7. Representação gráfica do resultado do PCA num eixo de duas dimensões aplicado aos atletas de natação, com uma recuperação de 53.8% da informação inicial. As cores dos grupos representam o género dos atletas. **51**

Figura 5.8. Gráfico de barras para a média da carga de treino ao longo das 29 semanas. **53**

Figura 5.9. Valores médios semanais da carga de treino (cinzento) e do cortisol salivar (riscas vermelhas). **54**

Figura 5.10. Representação do número de ocorrências de ITRS por semana. **55**

Figura 5.11. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento da média da concentração dos marcadores biológicos (sC e sIgA) em cada semana. **56**

Figura 5.12. Representação dos atletas dos respetivos escalões num sistema de eixos a duas dimensões. O eixo PC1 apresenta uma recuperação de 25.7% e o eixo PC2 apresenta uma recuperação de 14.2%. **60**

Figura 5.13. Relação entre as variáveis originais e as duas primeiras componentes principais para o conjunto de dados do futebol. **61**

Figura 5.14. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento dos atletas de futebol para o marcador sC. **61**

Figura 5.15. Perfis do comportamento médio do sC dos grupos formandos pelo dendrograma da Figura 3. **62**

Figura 5.16. Perfis do comportamento médio do sC para cada escalão, sub-15, sub-17, sub-19. **62**

Figura 5.17. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento dos atletas de futebol para o marcador biológico hormonal sT. **63**

Figura 5.18. Perfis do comportamento médio da sT referentes aos grupos formandos pelo dendrograma da Figura 5.17 e os Perfis do comportamento médio da sT para cada escalão, sub-15, sub-17, sub-19. **64**

Figura 5.19. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento dos atletas de futebol para o marcador biológico imunitário sIgA. **65**

Figura 5.20. Perfis do comportamento médio da sT referentes aos grupos formandos pelo dendrograma da Figura 5 e os Perfis do comportamento médio da sIgA para cada escalão, sub-15, sub-17, sub-19. **65**

Figura 5.21. Variação sazonal do volume de treino e perfis do comportamento médio do sC e sT referentes a cada escalão, sub-15, sub-17, sub-19. **67**

Índice de Tabelas

	Pág.
Tabela 2.1. Representação dos vários tipos de ligação.	**12**
Tabela 4.1. Médias e desvios padrões referentes aos dados cineantropométricos e funcionais.	**29**
Tabela 5.1. Resultado do PCA referente aos valores próprios e percentagem de recuperação da informação para cada componente principal.	**37**
Tabela 5.2. Sumário do resultado dos grupos formados pelo dendrograma. Assinalam-se os maiores e menores valores médios para cada variável, respetivamente pelas cores verde e vermelha.	**48**
Tabela 5.3. Valores de correlação entre carga de treino e o cortisol salivar.	**55**
Tabela 5.4. Sumário do resultado dos grupos formados pelo dendrograma da Figura 5.11.	**57**
Tabela 5.5. Contagem da ocorrência de ITRS nos grupos formados por quartis de IgA no momento antes do começo de época.	**58**
Tabela 5.6. Resultado do PCA referente aos valores próprios e percentagem de recuperação da informação para cada componente principal.	**59**
Tabela 5.7. Registo de ocorrências de sintomas de ITRS durante a época desportiva de cada escalão.	**66**

Capítulo I
Introdução

Capítulo I

Introdução

Na atualidade, áreas como a Química, a Biologia, a Medicina e a Engenharia geram quantidades elevadas de dados, que necessitam de ferramentas poderosas para o seu tratamento. Muita da informação tem caracter multidimensional e, desta forma, a importância da análise multivariada pode ser demasiadamente realçada. [1]

Na área da Química, sempre muita associada ao reconhecimento de padrões, a quimiometria tem providenciado, historicamente e com desenvolvimentos recentes, metodologias desenvolvidas com caracter muito flexível, aplicando-se desde sistemas relativamente pequenos até ao "big data". [2]

A quimiometria é utilizada para diversas tarefas e em diversas fases dos projetos, como sejam a recolha e análise de dados, essencialmente de carácter extensamente multivariado, calibração, modelação de processos, reconhecimento e classificação de padrões, correção e compressão de sinais, e controlo estatístico de processos.[3] Uma das características dos dados submetidos tem a ver com o facto de, apesar do número elevado de variáveis, ser possível encarar essas variáveis como um conjunto. [4] Muitas vezes, elaboram-se modelos com o intuito de prever valores alvo, características desejadas ou comportamentos de interesse. [3]

As técnicas quimiométricas têm dado provas em áreas diversas e com resultados muito positivos. [5,6] A variedade de aplicações é patente em revistas especializadas como sejam o *Journal of Chemometrics* ou *Chemometrics and Intelligent Laboratories*. Refira-se, no entanto, que em aspetos relacionados com a área das Ciências do Desporto a aplicação da análise multivariada é escassa, quando comparada com a proveniente da utilização de instrumentos estatísticos tipicamente lineares, tais como análises comparativas e da variância, de correlação e regressão. [7,8,9,10,11]

Nesta dissertação, algumas das aplicações vão demonstrar que, também aqui, a utilização de técnicas quimiométricas clássicas é de grande utilidade na exploração e entendimento dos sistemas. Devido ao estímulo do exercício imposto, existe um alto nível de *stress* fisiológico e psicológico sobre os atletas, que importa

compreender e se possível antecipar. Para tal, é fundamental a monitorização de marcadores biológicos sensíveis a alterações homeostáticas.[12] Dos exemplos de marcadores biológicos destacam-se as hormonas cortisol salivar (sC) e testosterona salivar (sT) que pela sua participação nos processos catabólico e anabólico, têm sido associados à resposta adaptativa na resposta à carga de treino.[13]

O cortisol, Figura 1.1, é uma hormona corticosteroide, da família dos esteroides e é produzida pela glândula suprarrenal.[14] A vantagem que existe em monitorizar o sC está relacionada com o facto de se encontrar na forma livre ativa, o mesmo não se verifica com o existente no plasma.[15]

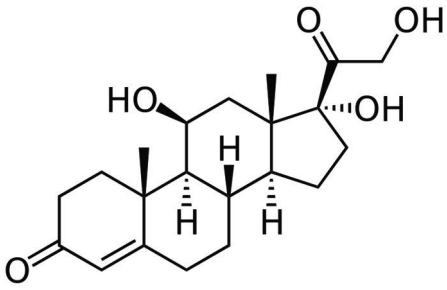

Figura 1.1. Estrutura química do cortisol.

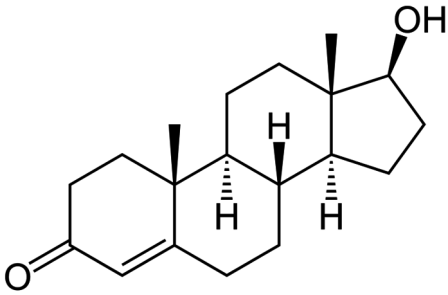

Figura 1.2. Estrutura química da testosterona.

A testosterona, Figura 1.2, é também um corticosteroide da família dos esteroides, que faz parte do grupo dos androgénios (grupo de hormonas com efeitos masculinizantes), e é sintetizada a partir do colesterol, ou, diretamente, a partir da Acetil-CoA.[14]

Estas hormonas surgem, assim, como sinalizadores das dificuldades de adaptação dos atletas ao treino, podendo interferir com a sua performance, e considerados como marcadores da fadiga. [16]

Situações de fadiga em atletas tem sido associada a valores baixos de sT. [17] Por outro lado, o sC é um marcador de stress por excelência e dependente da duração do esforço. Quando este excede 1 hora, existe um aumento dessa hormona. [18,19] Refira-se, ainda, que um aumento exponencial do cortisol tem efeitos nefastos sobre a ação imunitária.[20]

De facto, a adaptação dos atletas ao processo de treino está intimamente ligada com a resposta do sistema imunitário [21], sendo certo que a manutenção da saúde dos atletas é um fator importante no cumprimento dos planos de treino, bem como na prontidão para competir. [9] Assim sendo, vários autores relatam a importância de abordar a possibilidade de uma relação entre a contração de infeções, principalmente, no trato respiratório superior (ITRS) e a carga de treino. [22,23,24] Com este fim, é frequente recorrer ao controlo da imunoglobulina A salivar (sIgA). [25] As imunoglobulinas (Ig) são glicoproteínas produzidas e segregadas peças células B e plasmáticas. Encontram-se no soro e outros fluidos do corpo e dividem-se em cinco classes: IgG, IgA, IgM, IgD e IgE. A molécula básica de Ig apresenta uma forma de "Y" e é composta por quatro cadeias polipeptídicas, unidas por ligações de dissulfato. Esta torna-se importante no combate a agentes patogénicos que não são transportados pelo sangue, nomeadamente os causadores de ITRS.

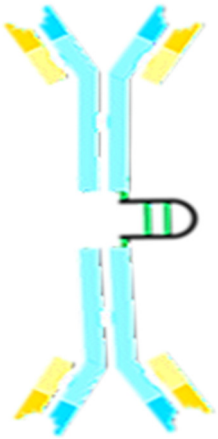

Figura 1.3. Representação da Imunoglobulina A. (adaptado [26]).

Embora não exista unanimidade nos estudos no que diz respeita à relação entre concentração de sIgA (Figura 1.3) e a predisposição para contração de ITRS. Certos autores referem que a baixa concentração de IgA pode estar associada a uma maior predisposição para a contração de ITRS [27], sendo a sIgA apontada como um bom indicador do risco deste tipo de infeções em atletas. [28]

Outra possibilidade prometedora para aplicação da quimiometria no âmbito da investigação em Ciências do Desporto passa pela identificação de potenciais variáveis discriminantes do talento desportivo, constituindo um processo prático e acessível para a análise dos dados provenientes da avaliação e controlo do treino em jovens atletas.[10] Neste contexto, as características cineantropométricas são um fator relevante na caracterização dos atletas em diversas modalidades.[29] Outro fator importante, está relacionado com a contribuição da energia aeróbia uma vez que a natação é considerada um desporto de grande resistência.[30]

No estudo constante desta dissertação, foram abordadas duas modalidades, Futebol e Natação. Relativamente à modalidade Futebol, existem fatores como a dinâmica de jogo, a velocidade de execução e a condição física que são determinantes no prognóstico do potencial de desenvolvimento do atleta. A conjugação destes fatores acarreta consequências fisiológicas que influenciam (e caracterizam) a performance.[31] Com o intuito de prever e analisar a sua influência foram analisados recorrendo novamente a técnicas quimiométricas, o sC e a sT, controlados regulamente ao longo de épocas desportivas.

Relativamente à Natação, o acompanhamento do atleta é um fator importante na elaboração do planeamento e periodização de treino, para evitar a exaustão e a fadiga que pode comprometer o seu desempenho [32] e, consequentemente, são utilizados métodos para avaliar a carga de treino, nomeadamente, componentes fisiológicas, indicadores de stress ou até análise subjetiva. [33] As características antropométricas, tem sido igualmente objeto de controlo dada a reconhecida influencia no sucesso dos nadadores. [11]

Considerando os pressupostos antes enunciados, o objetivo deste estudo, é o de interpretar e racionalizar os dados recolhidos em amostras de jovens atletas praticantes de natação e futebol recorrendo à visualização gráfica e reconhecimento de padrões. A análise centra-se sobre o comportamento dos atletas em resposta à aplicação das cargas de treino, bem como na avaliação de como as respetivas características físicas podem influenciar o desempenho, e consequentemente, promover uma melhoria da performance dos atletas.

Esta dissertação encontra-se estruturada do seguinte modo: no capítulo 2 é apresentada uma breve introdução à quimiometria, bem como dar conta da sua aplicabilidade da mesma. Ainda neste Capítulo são descritos métodos quimiométricos clássicos como a análise hierárquica de agrupamentos (HCA), análise de componentes principais (PCA) e correlação. No Capítulo 3 é apresentada a descrição da implementação dos métodos no software R e um guia prático da sua aplicação. A descrição e caracterização das bases de dados utilizadas nesta análise encontra-se no Capítulo 4. No Capítulo 5 são então apresentados os resultados bem como a discussão dos mesmos e, por fim, no Capítulo 6 encontram-se as considerações finais.

Capítulo II
Fundamentação Teórica

Capítulo II

Fundamentação Teórica

2.1 Quimiometria

Na década de 70, registava-se, na área da Química e noutras áreas, excessiva produção de dados face à capacidade de análise. Esta foi uma das motivações para que, entre 1972 e 1973, Bruce Kowalski e Charles Bender aplicarem uma mudança de rumo que marcou o início da quimiometria. [34,35]

No ano seguinte foi fundada a Sociedade Internacional de Quimiometria, sendo Bruce Kowalski e Svante Wold os seus primeiros membros. [36]

Assim, um dos objetivos da Quimiometria resulta da necessidade de solucionar dificuldades no tratamento de dados em Química, desenvolvendo e aplicando novos métodos estatísticos e matemáticos, a fim de processar quantidades cada vez mais maciças de dados produzidos pelos instrumentos analíticos modernos.[1]

O tratamento implica a utilização de técnicas diversas, passando pela otimização e validação de metodologias analíticas, planeamento experimental, técnicas várias de regressão, processamento de sinal, análise de fatores e calibração multivariada.

Na atualidade as aplicações mais frequentes são em áreas diversas da química como sejam a química analítica e forense, mas também a medicina, a tecnologia farmacêutica, e as áreas alimentar e ambiental. Todas estas áreas recorrem cada vez mais a técnicas semelhantes de tratamento estatístico com capacidade para tratar conjuntos de dados de grandes dimensões.

2.2 Métodos Quimiométricos

2.2.1 Análise Multivariada

Para analisar uma base de dados que contenha mais do que uma variável, é necessário recorrer a métodos multivariados, nos quais se incluem a análise de agrupamentos e a análise de componentes principais (PCA).

De uma forma sucinta, a análise de agrupamentos reúne os objetos com perfis semelhantes de acordo com um critério selecionado, e a análise de componentes principais permite enfatizar a informação mais importante do conjunto multivariado sob análise.

2.2.2 Análise Hierárquica de Agrupamento

A análise hierárquica de agrupamentos (do inglês *Hierarchical Cluster Analysis*, HCA), é um método cujo objetivo é agrupar objetos, i.e., pontos num espaço multidimensional descritos por variáveis comuns a todos eles. Esta agregação da informação pode ser realizada de forma iterativa, que dissocia (método divisivo) ou associa (método aglomerativo) os objetos a partir dos dados originais ou já transformados. Relativamente aos métodos divisivos, temos inicialmente um agrupamento global, sendo posteriormente feitas sucessivas divisões até que se encontre um só objeto por grupo. Por outro lado, os métodos aglomerativos são iniciados com os objetos separados, e de seguida é efetuada uma combinação de forma sequencial com o objetivo de reduzir o número de grupos, para finalmente produzir apenas um agrupamento com todos os objetos. [37,38,39]

A representação gráfica mais usual é o chamado dendrograma, onde os objetos (e grupos) são ligados por linhas com base nas semelhanças apresentadas entre si. Na Figura 2.1 apresenta-se um exemplo de um dendrograma.

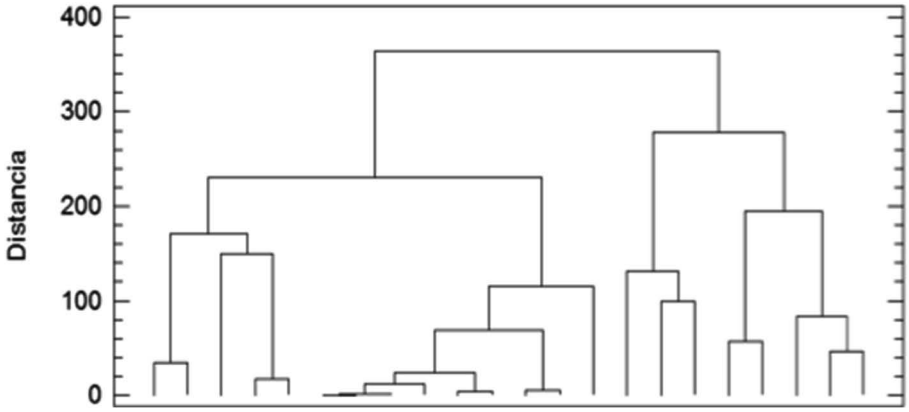

Figura 2.1. Exemplo de um dendrograma.

Para que os objetos sejam agrupados, é fundamental escolher o tipo de medida de similaridade entre objetos a ser utlizada, bem como o tipo de ligação. Inicialmente, promove-se a construção da matriz de distâncias, sendo a distância Euclidiana ("*Euclidean*"), a distância de Manhattan e a distância de Minkowski as mais utilizadas, existindo muitas outras. Estas distâncias são dadas por:

$$\mathbf{d}(\boldsymbol{i},\boldsymbol{j}) = \sqrt{\left|\boldsymbol{x_{i1}} - \boldsymbol{x_{j1}}\right|^{\mathbf{2}} + \left|\boldsymbol{x_{i2}} - \boldsymbol{x_{j2}}\right|^{\mathbf{2}} + \cdots + \left|\boldsymbol{x_{in}} - \boldsymbol{x_{jn}}\right|^{\mathbf{2}}} \tag{2.1}$$

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{in} - x_{jn}\right| \tag{2.2}$$

$$d(i,j) = \sqrt[q]{\left|x_{i1} - x_{j1}\right|^{q} + \left|x_{i2} - x_{j2}\right|^{q} + \cdots + \left|x_{in} - x_{jn}\right|^{q}} \tag{2.3}$$

Respetivamente, onde x_{ij} corresponde à coordenada j do objeto i no espaço multidimensional da representação.

Seguidamente procede-se à escolha do método de ligação dos objetos, i.e. o método de estabelecer a distância entre objeto e grupo ou entre grupos já formados. Tal como para a distância, também existem vários tipos de ligação como a ligação simples (*single linkage*) que define a distância entre os grupos como sendo a dos objetos mais próximos, a ligação completa (*complete linkage*) que define a distância entre grupos como sendo a distância entre os dois objetos mais afastados de cada grupo, a ligação média (*average linkage*) que é a média das distâncias entre todos os pares de objetos de cada grupo, o método do centróide (*centroid*) que representa um ponto no centro de cada grupo, o método da mediana (*median*) que é um caso particular do centróide ou o método de variância mínima (*Ward*) que associa os grupos que apresentam menor variância intra-grupo. Na Tabela 2.1 podemos visualizar como ocorrem estas ligações através dos esquemas representados.

Tabela 2.1. Representação dos vários tipos de ligação.

Critério de Ligação	Associação
Ligação Simples	
Ligação Completa	
Ligação Média	
Ward	

2.2.3 Análise de Componentes Principais (PCA)

A análise de componentes principais incide normalmente sobre um conjunto de dados que contém um elevado número de variáveis, sendo o seu principal objetivo reduzir o tamanho desse conjunto, preservando o máximo da informação original.

Quando aplicado o PCA, reduz-se a dimensão dado conjunto multivariado sob análise, transformando-se as variáveis originais noutras variáveis, preferencialmente em menor número e não correlacionadas entre si. Estas variáveis são combinações lineares das varáveis originais que são designadas por componentes principais (PC).[40]

Uma vez que as componentes principais devem refletir ao máximo as características dos dados originais, são então escolhidas as combinações lineares com variância máxima, de modo a explicarem uma grande parte da variação associada às variáveis originais. [41]

Posto isto, é possível reduzir o número de variáveis com o intuito de analisar o conjunto de dados de forma mais eficaz. É importante salientar que a informação mais relevante está contida nas primeiras PCs, pois estas retêm, em maior

quantidade, a variação presente nas variáveis originais, sendo, portanto, as mais informativas.

Após o procedimento PCA, os dados podem ser representados num gráfico de duas ou três dimensões, cujos eixos são as componentes principais com maior variância (as primeiras). Note-se que o significado da representação é tanto maior quanto maior for a informação recuperada por estas componentes.

Em suma, a análise de componentes principais tem como objetivo diminuir o conjunto de dados a analisar, reduzindo o número de variáveis e através da visualização gráfica, permitir identificar e interpretar relações entre variáveis, e variáveis e objetos.

2.2.4 Aplicação do método PCA

Para aplicar a Análise de Componentes Principais são necessárias várias etapas:

1. Escolher o software para execução do método, no caso vertente o RStudio;
2. Escolher as variáveis tendo em conta o objetivo do estudo;
3. Selecionar, usualmente, entre a matriz de covariância e a matriz de correlação, uma vez que os dados têm de sofrer esta transformação para permitir um cálculo de valores próprios. A matriz de covariância é muitas vezes utilizada se as unidades de medida forem as mesmas entre variáveis i.e, se as variáveis tiverem ordens de grandeza semelhantes. Por outro lado, a matriz de correlação é preferível quando as variáveis apresentam distintas ordens de grandeza;
4. Escolher o número de componentes principais a analisar, podendo-se recorrer a três critérios distintos Pearson, Kaiser e Scree Plot;
5. Analisar e interpretar as representações gráficas (muitas vezes assentando no biplot).

2.2.4.1 Escolha da matriz

Apresentamos, de seguida, algumas definições.

Para que seja encontrado um conjunto de dados com p variáveis não correlacionadas e de variância máxima, a matriz principal tem de sofrer uma rotação do sistema ortogonal de eixos associados às variâncias iniciais.

Posto isto, seja $X = [X_1 \dots X_p]'$ um vetor e Σ a respetiva matriz de variância/covariância,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix} \tag{2.4}$$

onde,

$$\sigma_{ij} = Cov\,(X_i\,, X_j) \tag{2.5}$$

Seja $X_{(n\times p)}$ uma matriz de dados de dimensão *n*, resultando de observações desta matriz.

$$\mathrm{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{2.6}$$

A matriz $X_{(n\times p)}$, após centragem, sofre uma rotação ortogonal associada às variáveis iniciais, resultando na matriz $Y_{(n\times p)}$. Assim as colunas desta nova matriz Y representam as componentes principais, passando estas ser as novas variáveis.

A formação das componentes principais advém de combinações lineares das *p* variáveis da matriz X

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = a_j'X, \quad j = 1, \cdots, p \tag{2.7}$$

onde, $j = 1, \cdots, p$ e $a_{ij}(i = 1, \cdots, p\,;\, j = 1, \cdots, p)$ são contantes.

Importante referir que os coeficientes destas combinações são determinados satisfazendo:

1. A variância da primeira componente é maior que a variância da segunda componente principal e assim sucessivamente:

$$Var\ (Y_1) \geq Var\ (Y_2) \geq \cdots \geq Var(Y_p) \tag{2.8}$$

2. Não há correlação entre duas componentes principais;
3. A soma dos quadrados dos coeficientes em qualquer componente principal é 1;

Deste modo, asseguramos que a componente principal com maior variância é Y_1, que Y_2 corresponde à segunda maior variância e assim sucessivamente. Importante ainda é o facto de que as componentes principais não são correlacionadas entre si.

Na equação 2.6 acima temos os p vetores próprios ($a_1, a_2, \cdots a_p$), associados aos p maiores valores próprios de $\Sigma\ (\lambda_1 > \lambda_2 > \cdots > \lambda_p)$ com $Var(Y_j) = \lambda_j$. Visto que, as componentes principais são calculadas de forma a não serem correlacionadas duas a duas, então temos uma covariância nula

$$Cov(Y_j\ , Y_{j'}) = a_j' \Sigma a_{j'} = a_j' \lambda_j a_{j'} = \lambda_j a_j' a_{j'} = 0 \tag{2.9}$$

i.e. $a_j' a_{j'} = 0$, o que significa que a_j' e $a_{j'}$ (com $j \neq j'$) são vetores ortogonais.
Nem sempre as variáveis estudadas são medidas na mesma unidade. Neste caso, surge a necessidade de uniformizar a base de dados, através da divisão de cada valor pelo desvio padrão de cada variável, devidamente centrada. Assim, todas as variáveis em estudo passam a ter a mesma uma variância unitária e valor médio nulo.

A matriz de covariância das "novas" variáveis é igual à matriz de correlação do conjunto de variáveis iniciais, dado que:

$$\boldsymbol{Cov}\left(\frac{X_i}{\sigma_i}, \frac{X_j}{\sigma_j}\right) = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j} = \boldsymbol{Corr}(X_i, X_j) \tag{2.10}$$

Posto isto, a matriz de correlação (P) irá possibilitar a análise das componentes principais, sendo

$$P = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \tag{2.11}$$

onde $p_{ij} = Corr\ (X_i\ , X_j)$

Estas componentes serão determinadas através dos valores e vetores próprios da matriz P. Naturalmente, os vetores próprios de P e as componentes principais não são iguais aos da matriz Σ.

2.2.4.2 *Scores*

A transformação descrita pela Equação 2.6, em que as componentes principais são combinações lineares da transformação das variáveis em estudo, pode também aplicar-se aos dados, ou seja, aos vetores de observações $x_1\ , x_2\ , \cdots , x_p$ (colunas da matriz de dados X) das variáveis $X_1\ , X_2\ , \cdots , X_p$ respetivamente.

Posto isto, obtemos uma nova matriz $Y_{(n \times r)}$, onde ij-ésimo elemento será igual ao *score* do i-ésimo objeto para j-ésima componente principal

$$y_{ij} = a_{1j}x_{i1} + a_{2j}x_{i2} + \cdots + a_{pj}x_{ip} \tag{2.12}$$

Assim, a matriz dos *scores* dos objetos é dada por:

$$\mathrm{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix} \tag{2.13}$$

E corresponde à representação dos objetos no novo sistema de eixos.

2.2.4.2 Critério de seleção do número de componentes principais

Como foi indicado**,** depois de selecionada a matriz de covariância ou de correlação dependendo do tipo de conjunto de dados a analisar, resolve-se uma equação de valores próprios e são formadas as novas variáveis, as componentes principais. Para que se selecionem as componentes principais com maior relevância, podem ser usados três critérios de seleção distintos, nomeadamente, critério de Kaiser $(\lambda > 1)$, critério de Pearson (ou regra dos 80%) e o Scree Plot:

I. Critério de Kaiser ($\lambda > 1$) [10]

Este critério é utilizado quando é selecionada a matriz de correlação. Assim, devem ser consideradas apenas as componentes com valor próprio superior a um, valor este que corresponde à média.

$$\bar{\lambda} = \frac{1}{p}\sum_{j=1}^{p} \lambda_j \tag{2.14}$$

II. Critério de Pearson (ou regra dos 80%)

Este é aplicado quando selecionada a opção referente à matriz de covariância, mas pode sê-lo também quando é utilizada a matriz de correlação. Deste modo, as componentes principais têm de apresentar uma recuperação da informação ou variabilidade total acima de 80%. Dito de outra forma, as componentes principais a serem consideradas devem ser as necessárias para que a percentagem de variância seja superior a 80%; como tal retêm-se a primeiras r componentes principais de modo a que:[42]

$$\sum_{j=1}^{r} \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 0.80 \tag{2.15}$$

III. Scree Plot

Por fim, este critério tem como base a visualização gráfica da representação dos pontos de abcissa *j* e ordenada igual ao j-ésimo valor próprio, ou a percentagem de variância explicita pela j-ésima componente principal. De acordo com este critério, as r componentes principais consideradas devem ser as que mais se destacam das restantes, apresentando uma elevada contribuição aquando comparadas com as outras.[41]

Capítulo III
Aspetos Práticos e de Software

Capítulo III

Aspetos Práticos e de Software

Com a finalidade de tornar mais clara a metodologia experimental desenvolvida nesta dissertação, que permitiu desenvolver o processo de tratamento de dados, surge este Capítulo com o intuito de apresentar o software usado, as linhas de código aplicadas, bem como a sua funcionalidade.

É importante realçar que as bases de dados analisadas neste estudo sofreram um pré-processamento antes de ser aplicada a metodologia que se segue. Este passo é fundamental em muitos tratamentos análogos e é algo que não deve ser menosprezado quer em termos de extensão temporal, quer em esforço.

No que se segue, iremos utilizar um formato semelhante ao de um tutorial para abordar os aspetos práticos e de software. Em anexo encontram-se scripts realizados em RStudio.

3.1 RStudio

R é um software desenvolvido por Ross Ihaka e Robert Gentleman na Universidade de Auckland na década de 90. Este software é grátis e de fácil acesso, utilizado para manipulação e visualização de dados, sendo empregue em áreas científicas de estatística computacional. [43]

A linguagem R é um ambiente de programação estatística e de código aberto com uma enorme variedade de algoritmos já implementados, aplicada no programa RStudio® que, por sua vez, é um ambiente de desenvolvimento integrado (IDE) desta linguagem.

De um modo geral, o RStudio® visualmente é dividido em quatro painéis (ver Figura 3.1.):

- No painel 1 insere-se o *script*, sendo este constituído por linhas de código R;
- O painel 2 apresenta o histórico dos dados inseridos, bem como as alterações efetuadas;
- O painel 3 consiste num terminal de *input* e *output*;

- O painel 4 é multifuncional, e nele surge a visualização gráfica.

Assim que é efetuada uma instrução no painel 1, esta é registada nos painéis 2 e 3, e, caso seja uma instrução gráfica, o painel 4 irá apresentá-la graficamente.

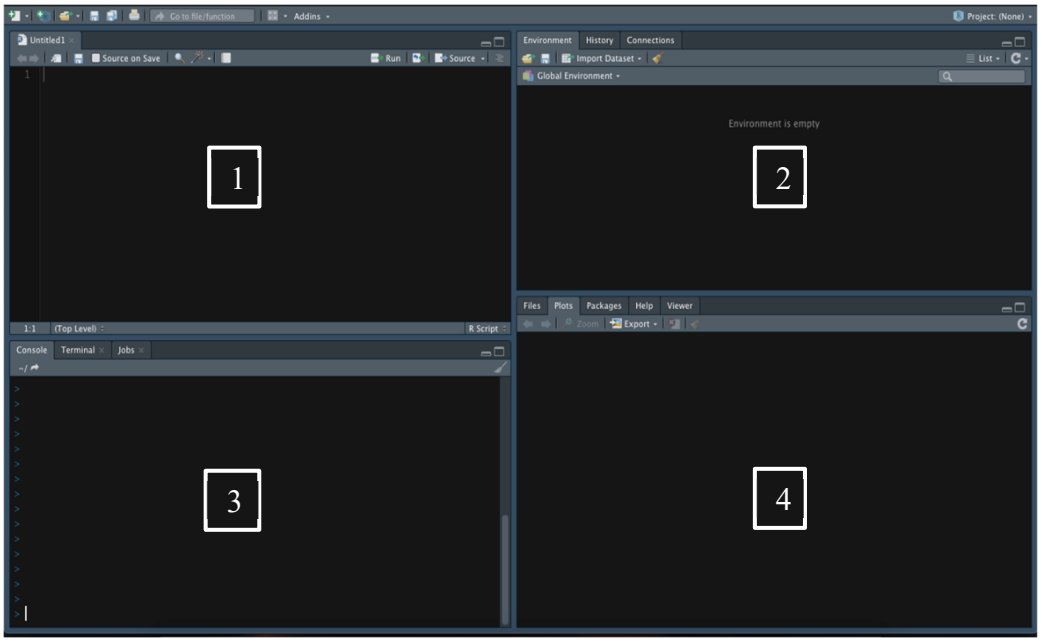

Figura 3.1. Representação do painel de trabalho do RStudio.

Todos os utilizadores do R têm à sua disposição uma vasta diversidade de pacotes (packages) que contêm conjuntos de dados e funções específicas com várias aplicações estatísticas. Estas packages por sua vez são elaborados por colaboradores que contribuem com o intuito de inovar e ajudar o utilizador.[44]

Para o desenvolvimento deste estudo foram usados, entre outras, packages desenvolvidos por Alboukadel Kassambara, cientista da área da saúde com uma vasta experiência em métodos estatísticos e computacionais, sendo por sua vez o autor de alguns packages de análise multivariada. [45] As packages utilizadas foram, a FactoMineR que serve para analisar e calcular as componentes principais e factoextra para extrair, visualizar e interpretar os resultados. Na Figura 3.2 podemos observar esquematicamente o processo utilizado.

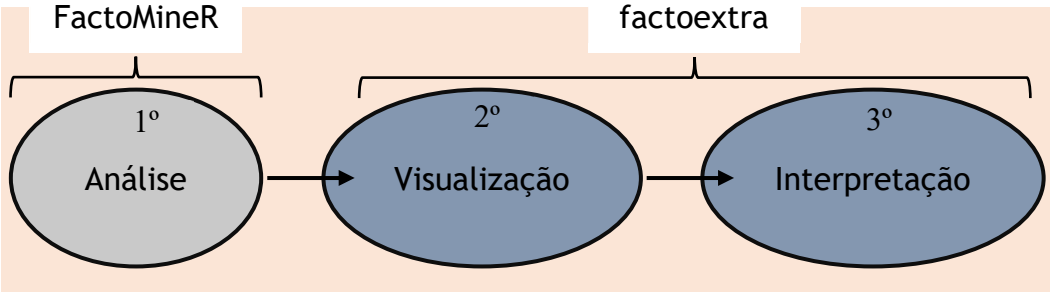

Figura 3.2. Ilustração do funcionamento dos packages FactoMineR e factoextra.

3.2 Implementação das linhas de código

Antes de proceder à explicação das linhas de código aplicadas neste estudo, é importante referir que, previamente, é necessário estabelecer qual a análise que mais se adequa, para posteriormente aplicar ao conjunto de dados.

Como já referido anteriormente, neste tipo de análise existem vários procedimentos que devem ser cumpridos para que esta seja eficaz, existem processos que nos ajudam a perceber se ela é significativa.

O procedimento que se segue é referente à análise de componentes principais e análise hierárquica de agrupamentos. É importante realçar que houve uma combinação entre o PCA e o HCA, ou seja, primeiramente foi feita uma análise de componentes principais e, posteriormente é utilizando este resultado, foi calculada a análise hierárquica de agrupamentos.

A combinação destes dois métodos é benéfica se o conjunto de dados for multidimensional, englobando múltiplas variáveis contínuas. Através da análise de componentes principais é possivel reduzir a dimensão do sistema, utilizando-se um menor número de variáveis continuas contendo a informação mais relevante. Deste modo, a análise de agrupamentos hierarquica fornecerá um *clustering* mais estável.

O procedimento seguiu, assim, os seguintes passos:

1. <u>Seleção dos métodos a aplicar:</u>

Consoante a base de dados e o objetivo do estudo, é importante definir previamente quais os métodos quimiométricos a ser aplicados.

2. Escolha dos packages e instalação no Rstudio:

Em seguida segue-se a escolha dos packages que irão ser utilizados no estudo para que, caso não estejam já incorporados no software de base, possam ser instalados. Assim, as linhas de código representadas abaixo, indicam a instalação dos packages que irão ser usados neste guia prático e no estudo realizado.

```
#Instalção das packages
install.packages("FactoMineR")
install.packages("factoextra")
```

3. Importação da base de dados para o software:

Este comando irá depender do modo como o ficheiro que contém os dados, está armazenado e em que formato; no exemplo que se segue o ficheiro "UCoimbra" está gravado em *csv* (*read.csv2*). Neste caso, a primeira coluna do ficheiro não é contabilizada para os cálculos, uma vez que contém o título das variáveis (*header=True*) e as colunas estão separadas por ponto e virgula (*sep=";"*).

```
#Introdução da base de dados no sofware
UCoimbra=("UCoimbra.csv", # conjunto de dados
          header = TRUE, # indicação que as variáveis têm título
          sep = ";") # indicação que as colunas são separadas por ";"
```

4. Utilização da função PCA do package FactoMineR:

O código que segue leva cabo a análise das componentes principais, sobre os objetos e as variáveis. É recomendado que as variáveis sejam escaladas, ou seja, padronizadas, uma vez que podem ter unidades de medida distintas (*scale.unit=TRUE*). O objetivo desta ação é tornar as variáveis comparáveis, tendo geralmente desvio padrão 1 e média 0.

```
#Função para cálcular as componentes principais
PCA_UCoimbra<-PCA(UCoimbra, # ficheiro
                  graph=FALSE, # indicação para não ser produzido o gráfico
                  scale.unit =TRUE) # indiação p/ escalar os dados
```

5. Escolha das componentes principais (PC):

5.1 Valores próprios/ variâncias:

Os valores próprios medem a quantidade de variação que cada componente principal (PC) possuirá. As primeiras PC's apresentam os maiores valores próprios e os menores estão nas subsequentes PC's, como focado em Capítulo anterior. Conclui-se que as primeiras PC's correspondem à quantidade máxima de variação no conjunto de dados. Esta informação é extraída pela função *get_eigenvalue*.

```
#Cálculo dos Valores Próprios
valores_proprios<- get_eigenvalue(PCA_UCoimbra)
```

É através desta avaliação que conseguimos determinar o número de PC's a ter em conta. Neste caso específico foram calculadas 14 componentes principais.

```
#Resultado do Cálculo dos Valores Próprios
        eigenvalue    variance.percent        cumulative.variance.percent
Dim.1  6.110267819      43.64477014                    43.64477
Dim.2  2.082358732      14.87399095                    58.51876
Dim.3  1.511605486      10.79718204                    69.31594
Dim.4  1.110364740       7.93117671                    77.24712
Dim.5  0.808687481       5.77633915                    83.02346
Dim.6  0.536817871       3.83441336                    86.85787
Dim.7  0.459149299       3.27963785                    90.13751
Dim.8  0.419192186       2.99422990                    93.13174
Dim.9  0.336507034       2.40362167                    95.53536
Dim.10 0.278856180       1.99182985                    97.52719
Dim.11 0.151199958       1.07999970                    98.60719
Dim.12 0.113161461       0.80829615                    99.41549
Dim.13 0.077481439       0.55343885                    99.96893
Dim.14 0.004350312       0.03107366                   100.00000
```

A segunda coluna é referente à variação explicada por cada valor próprio. Por exemplo, a primeira PC (*Dim.1*) apresenta uma percentagem de variação de 43.64% e a segunda PC (*Dim.2*) 14.87%. Somando estas duas componentes temos apenas em duas componentes principais uma recuperação da informação total de 58.51%. Tipicamente, tal recuperação corresponde a uma representação gráfica que permite já uma interpretação fundamentada.

No Capítulo 2 são descritos os critérios teóricos da seleção das componentes principais. Segundo Kaiser [10], os valores próprios podem ser utilizados para determinar o número de PC's a reter depois de aplicada a função PCA.

- O valor próprio > 1 indica que as PC's são contabilizadas ou não por uma das variáveis originais dos dados padronizados. Isto só se aplica quando os dados são normalizados.
- Também se pode limitar o número de componentes à soma da variância total. Por exemplo, se for considerado suficiente 70% da variância total explicada, então usa-se o número de componentes até atingir essa variância. No caso acima descrito ir-se-ia usar as duas primeiras componentes.

Geralmente, a tendência é escolher as primeiras componentes principais com o objetivo de encontrar padrões de interesse nos dados. Mesmo assim, na escolha das mesmas, não existe uma forma objetiva bem aceite de decidir a quantidade suficiente a considerar. Esta escolha pode estar dependente da área de aplicação do estudo e do conjunto de dados. Neste exemplo temos as duas primeiras componentes a explicar 58.51% da variação, o que é considerado uma percentagem aceitável, como acima foi referido.

5.2 *Scree Plot*:

O método *Scree Plot* é efetuado pela função abaixo:

```
#Função Scree Plot
fviz_eig(PCA_UCoimbra, addlabels = TRUE)
```

Refere-se a um método alternativo para determinar o número de componentes principais, considerando a parcela de valores próprios ordenados do maior para o menor (Figura 3.3). O número de PC's é determinado no ponto a partir do qual os restantes valores próprios são todos relativamente pequenos e de dimensão comparável. [42]

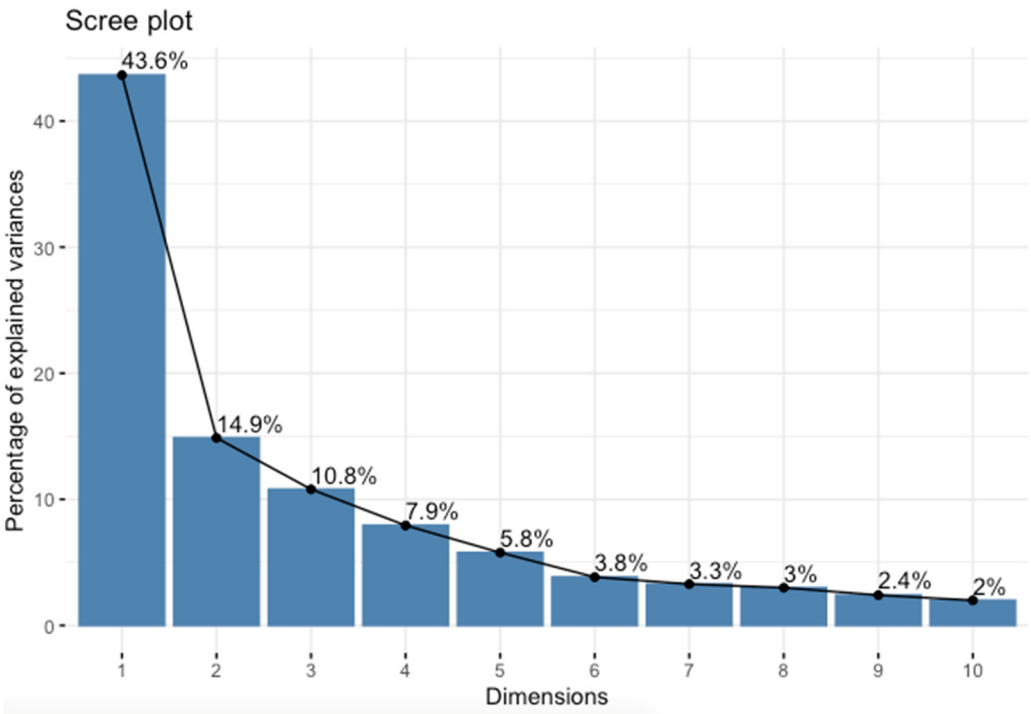

Figura 3.3. Visualização gráfica do método *SreePlot*. Neste caso, deveriam ser selecionadas 5 componentes principais.

6. <u>Resultado das variáveis:</u>

Através da função *get_pca_var* é fornecida uma matriz com os resultados para as variáveis ativas (coordenadas de variáveis para criar um gráfico de dispersão, *$coord*, correlação entre variáveis e eixos, *$cor*, co-seno quadrado representa a qualidade de representação das variáveis no mapa de fatores, *$cos²* e contribuições (em percentagem) das variáveis para as componentes principais, *$contrib*).

```
#Matriz das Variáveis Ativas
var_UCoimbra <- get_pca_var(PCA_UCoimbra)
```

Através da função *fviz_contrib* observa-se graficamente a contribuição/peso que as variáveis apresentam nas componentes principais.

```
#Visualização gráfica da contribuição das variáveis nas PC's
fviz_contrib(PCA_UCoimbra, choice = "var", axes = 1, top = 10) # Primeira Componente
fviz_contrib(PCA_UCoimbra, choice = "var", axes = 2, top = 10) # Segunda Componente
```

Pode-se observar os gráficos dessas mesmas contribuições na Figura 3.4.

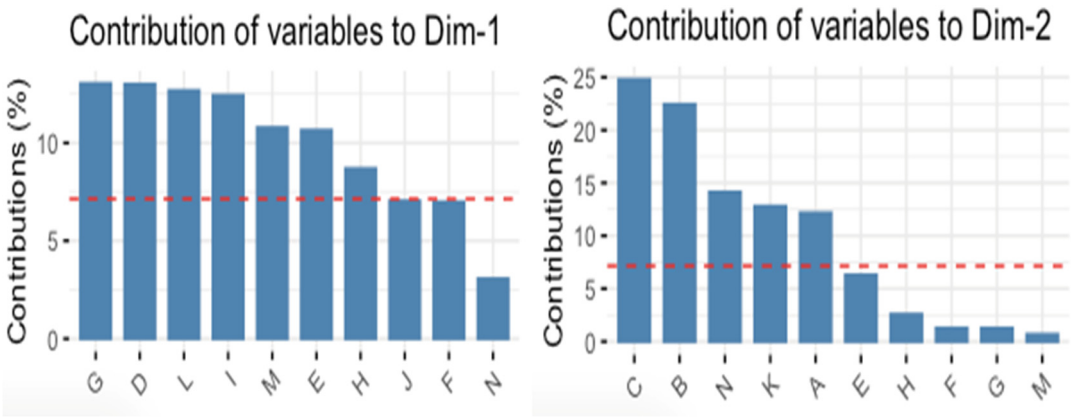

Figura 3.4. Representação gráfica da contribuição das variáveis nas componentes principais.

7. <u>Visualização gráfica do PCA:</u>

Por último temos a visualização gráfica do PCA, neste caso foi escolhido que os objetos fossem representados por pontos (*geom.ind="point"*), com um determinado tamanho (*pointshape=20 , pointsize=3*). Por sua vez, os objetos foram definidos em cor de laranja e as variáveis com a cor preta (*col.ind=("darckorchid3"), col.var=("black")*), sendo esta seleção deixada ao critério do utilizador. Existem mais parâmetros que se podem acrescentar para controlar o *biplot*, como sejam legenda, título, e nome dos eixos, entre vários outros.

```
#Visualização gráfica do resultado do PCA
fviz_pca_biplot(PCA_UCoimbra,
                geom.ind = "point", # apresentação do objeto (ponto)
                pointshape=20 , pointsize=3, # tamanho do objeto
                col.ind =("darkorchid3"), # cor dos objetos
                col.var="black") # cor das variáveis
```

No exemplo apresentado na Figura 7, temos uma recuperação da informação de cerca de 58.5%, sendo esta obtida nas primeira e segunda componentes principais. Os pontos representam os objetos e as setas as variáveis do conjunto de dados inicial.

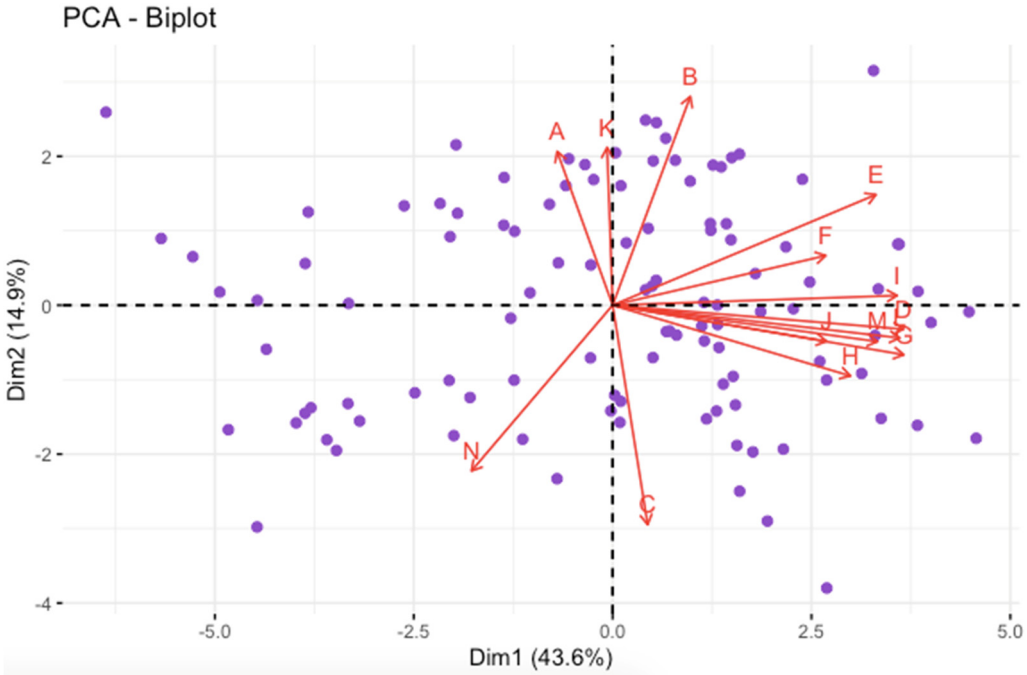

Figura 3.5. Exemplo da representação gráfica a duas dimensões dos resultados do PCA.

De um modo geral, a interpretação do biplot passa por registar se um objeto (representação dos pontos da Figura 3.5) se encontra no sentido crescente de uma dada variável ou do lado oposto. Caso esteja colocada no sentido crescente, significa que tem um valor elevado para uma essa variável; se estiver do lado oposto, esse valor será baixo para essa variável.

8. <u>Função HCPC:</u>

A função que se segue é usada para efetuar o agrupamento hierárquico utilizando o critério de Ward (ver Capitulo 2, seção 2.2.2) sobre as componentes principais selecionadas. Este critério baseia-se em informação ligada à variância multidimensional do conjunto de dados.

```
#Função para cálcular o agrupamento hierárquico
HCA_UCoimbra=HCPC(PCA_UCoimbra, nb.clust = 0, graph = FALSE)
```

É importante realçar que o ficheiro *PCA_UCoimbra* inserido na função *HCPC* foi calculado anteriormente pela função *PCA*. O parâmetro *nb.clust* especifica o número de grupos que são considerados como formados. Os valores possíveis para *nb.clust* são 0, que indica que o corte será onde o utilizador assinalar, 1, que indica que o

corte é realizado automaticamente e qualquer outro número inteiro positivo levará a um corte com esse número de agrupamentos. Esta função possibilida ao utilizador a escolha o número de componentes principais que deseja considerar; sem outra indicação a função considera 5 PC's.

9. Visualização gráfica da função HCPC:

Com a finalidade de visualizar o dendrograma gerado pela função acima explicada, é utilizada a função *fviz_dend* do package *factoextra*.

```
#Visualização gráfica do resultado do HCPC
fviz_dend(HCPC)
```

A função acima é a função simplificada e, por definição, apresenta-nos o dendrograma da Figura 3.6.

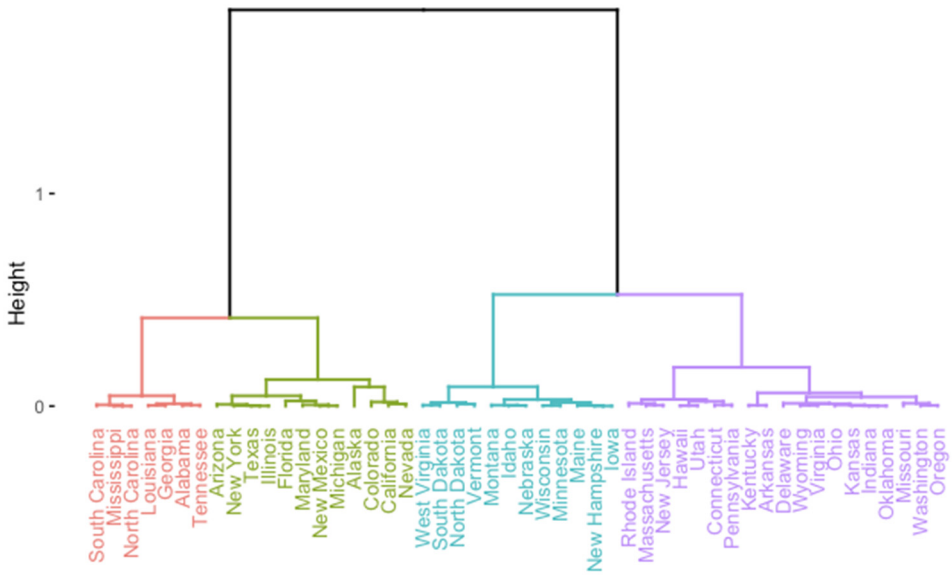

Figura 3.6. Exemplo da representação do dendrograma do package factoextra.

Esta função tem várias parâmetros disponíveis para controlar a visualização e, neste caso específicoforam colocados alguns comandos para melhorar a representação.

```
fviz_dend(HCPC, #ficheiro
          palette = "jco", #cor
          rect = TRUE , rect_fill = TRUE, #adicionar rectângulos
          rect_border = "jco", #cor dos retângulos
          show_labels = FALSE) #omitir o nome dos objetos
```

Na Figura 3.7 , podemos observar várias formas de apresentar o dendrograma de um mesmo conjunto de dados. Na imagem 1 temos o formato retângular, na imagem 2 temos o formato em círculo e na imagem 3 temos o formato ramificado. O utilizador pode escolher o formato mais adequado e, se inserir o comando type = ("rectangle") o formato será o retângular, type = ("circular") apresenta um formato circular e type = ("phylogenic") apresenta um formato ramificado.

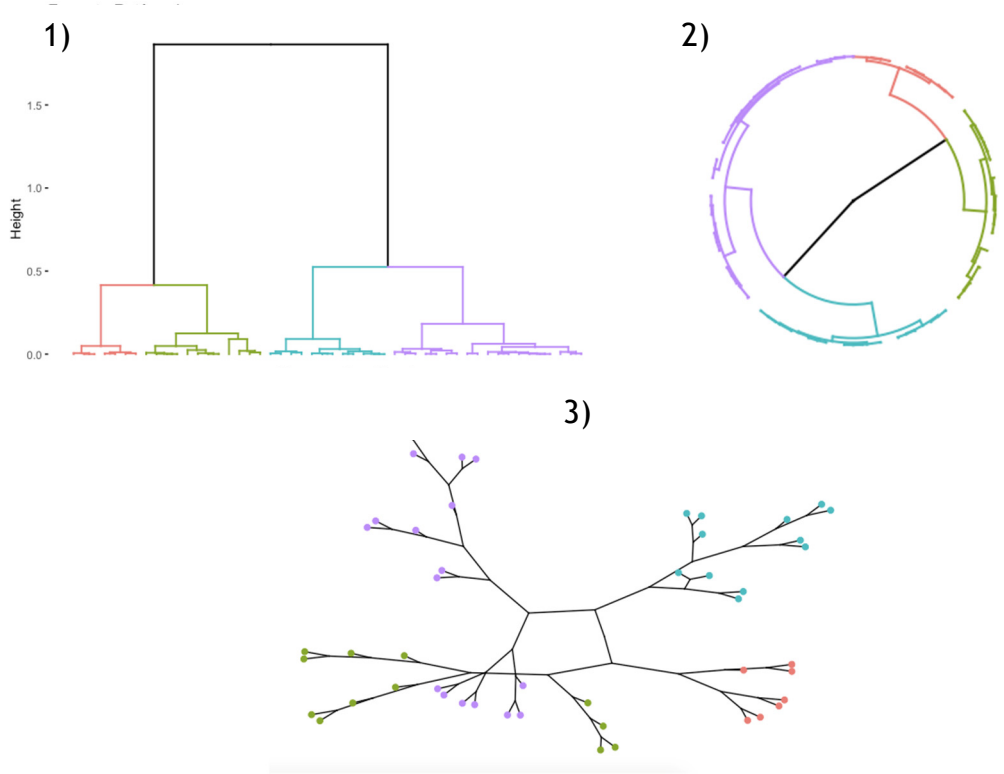

Figura 3.7. Três exemplos de apresentação do dendrograma: 1) formato retangular, 2) formato circular, 3) formato ramificado.

Capítulo IV
Conjunto de Dados

Capítulo IV

Conjunto de Dados

Em Capítulos anteriores foram detalhados os métodos utilizados para a elaboração da análise das bases de dados relativas a atletas de natação e futebol. Deste modo, no Capítulo que se segue encontram-se descritos três conjuntos de dados sujeitos análises. É de realçar que todos estes dados tratados foram submetidos à comissão de ética da Faculdade de Ciências do Desporto e Educação Física da Universidade de Coimbra.

4.1 Bases de Dados

As bases de dados analisadas neste trabalho foram disponibilizadas pelo Professor Luís Rama da Faculdade de Ciências do Desporto e Educação Física da Universidade de Coimbra, com objetivo de submetê-los à análise quimiométrica e identificar padrões entre atletas.

O primeiro conjunto de dados é referente a uma recolha de variáveis cineantropométricas e funcionais em particular da capacidade aeróbia, em atletas de elevado nível competitivo de natação de ambos os sexos, recolhidas durante a participação em estágios nacionais promovidos pela Federação Portuguesa de Natação destinados a jovens dos escalões pré juniores.

O segundo conjunto refere-se a uma análise de marcadores biológicos salivares hormonais: cortisol e testosterona, e imunitários, imunoglobulina A, recolhidos ao longo de um macro ciclo competitivo. Foram analisadas as amostras de atletas de elevado nível competitivo de ambos os sexos, tendo sido controlada a carga de treino imposta durante o período de estudo.

Por fim, o terceiro estudo corresponde à análise do comportamento de vários marcadores biológicos salivares, cortisol, testosterona e imunoglobulina A em jovens do sexo masculinos pertencente a três escalões de futebol: sub-15, sub-17 e sub-19, durante uma época de competição.

É de salientar que este processo teve vários obstáculos uma vez que a recolha não foi feita pelo analista. Por exemplo, durante a recolha da primeira base

de dados, houve medições novas que foram surgindo no decorrer do estudo, o que levou a atletas terem sido submetidos a protocolos de avaliação diferentes para avaliação da mesma variável, ou de se terem adicionado ou excluído variáveis. Assim existem registos cujo número de sujeitos é variável, dada a necessidade de excluir todos os elementos que não cumpriram os mesmos protocolos de avaliação. Esta situação, além de promover limitação no poder estatístico origina a ocorrência de *missing data*. Foram então selecionadas as variáveis mais relevantes de acordo com a literatura, para o estudo em questão e que reuniam a amostra mais elevada possível.

Ressalte-se que o tratamento prévio resultante destas dificuldades é um processo demoroso e que requer cuidado especial.

4.1.1 Características Cineantropométricas em Nadadores

O primeiro conjunto de dados é constituído por 190 atletas do sexo feminino, com idades compreendidas entre os 12 e os 14 anos e 265 atletas do sexo masculino, com idades entre os 13 e os 16 anos.

Como já referido anteriormente, este conjunto de dados foi submetido a um pré-processamento e a amostra inicial foi reduzida, tanto no número de variáveis como no número de atletas, passando para 20 atletas do sexo feminino e 85 atletas do sexo masculino.

Estes atletas foram sujeitos ao controlo de variáveis agrupadas de acordo com as seguintes dimensões:

a) Cineantropometria: foram controladas medidas caracterizadoras da morfologia e composição corporal. Dimensões corporais e pregas subcutâneas.
b) Avaliação funcional geral: todas as flexões e extensões (plantar, ombro e tronco) e força de preensão;
c) Avaliação funcional específica: Velocidade máxima, Deslize, Velocidade T30 e In T30.

Facilitando a compreensão e a leitura, na Tabela 4.1 estão descritas as variáveis consideradas, com média e desvio padrão respetivamente.

Este primeiro conjunto de dados, é assim, referente a 105 atletas e 28 variáveis distintas.

Tabela 4.1. Médias e desvios padrões referentes aos dados cineantropométricos e funcionais.

Variáveis	Abv	Masculino Média ± sd	Feminino Média ± sd
Idade (Anos)	Id	15. ± 0.5	13. ± 0.4
Somas de 6 Pregas subcutâneas (mm)	SK	60 ± 15	83 ± 23
Massa Corporal (Kg)	MS	63 ± 7	50 ± 4
Estatura (cm)	Est	172 ± 6	160 ± 5
Altura Sentado (cm)	AS	88 ± 3	82 ± 2
Envergadura (cm)	Env	177 ± 6	163 ± 5
Índice de Morfodimânico	IMD	4.4 ± 0.3	4.7 ± 0.3
Comprimento da Mão (cm)	CM	19.3 ± 0.9	17.3 ± 0.3
Largura da Mão (cm)	LM	8.1 ± 0.4	7. ± 0.3
Comprimento do Pé (cm)	CP	26 ± 1	23.1 ± 0.9
Largura do Pé (cm)	LP	9.5 ± 0.6	8.3 ± 0.4
Comprimento do Membro Superior (cm)	CMS	69 ± 2	65 ± 3
Comprimento do Membro Inferior (cm)	CMI	84 ± 4	78 ± 4
Diâmetro Biacromial (cm)	DBA	39 ± 2	34 ± 2
Diâmetro Bicristal (cm)	DBC	26 ± 2	23 ± 1
Diâmetro Toracosagital (cm)	DT	20. ± 2	17 ± 1
Flexão plantar (º)	FlexPl	30 ± 10	34 ± 6
Extensão plantar (º)	ExtPl	27 ± 10	17 ± 3
Flexão do Ombro (º)	FlexOm	16 ± 16	24 ± 28
Extensão do Ombro (º)	ExtOm	85 ± 23	90 ± 25
Flexão do Tronco (cm)	FlexTr	9 ± 7	11 ± 8
Extensão do Tronco (cm)	ExtTr	39 ± 11	38 ± 7
Força de Preensão Mão Direita (N)	PrMD	42 ± 7	27 ± 4
Força de Preensão Mão Esquerda (N)	PrME	40 ± 6	26 ± 4
Velocidade Máxima (m/s)	Velmax	1.5 ± 0.1	1.4 ± 0.1
Deslise (m)	Desl	7 ± 1	7 ± 1
Velocidade T30 (m/s)	VelT30	1.31 ± 0.09	1.18 ± 0.09
In T30 (m/s²)	INT30	2.3 ± 0.3	2.1 ± 0.2

Sk: subescapular, tricipital, suprailíaca, abdominal, crural e geminal; IMD=estatura/diâmetro biacromial;

4.1.2 Marcadores Biológicos Salivares na Natação

Relativamente a este segundo estudo, 16 atletas de ambos os sexos foram submetidos a uma análise durante 29 semanas de treinos, onde foram recolhidas amostras de saliva para monitorizar a resposta ao treino dos marcadores biológicos já referidos anteriormente. Esta recolha foi feita sempre nas mesmas condições, nomeadamente, antes do primeiro treino semanal e no mesmo período horário (17h-19h).

Toda a carga de treino realizada foi monitorizada através do volume (metros nadados em cada sessão de treino) e a intensidade determinada pela ponderação do volume cumprido por unidade de treino.

4.1.2 Marcadores Biológicos Salivares em Futebolistas

O terceiro conjunto de dados, está relacionado com a modalidade de futebol, onde foram recolhidas mensalmente amostras de saliva para avaliar o comportamento do sC, do sT e sIgA em três escalões distintos durante os 8 meses correspondentes à época desportiva em curso. No total foram avaliados 57 atletas do sexo masculino, os quais 17 referentes ao escalão sub-15, 22 atletas do escalão sub-17 e 18 do escalão sub-19, é de realçar ainda que esta recolha decorreu durante a época.

4.2 Metodologia de Recolha de Dados

Como já referido anteriormente, a recolha destes três conjuntos de dados descritos acima, foi efetuada por investigadores da Faculdade de Ciências do Desporto da Universidade de Coimbra. Como seria de esperar foram seguidos os procedimentos adequados na realização dos protocolos e seguidas metodologias apropriadas na recolha dos dados, tendo sempre presente o objetivo de minimizar os erros entre recolhas (erro técnico de medida).

O conjunto de dados das características cineantropométricas regeu-se por quatro protocolos diferentes. Relativamente à cineantropometria e á avaliação funcional foi seguido o protocolo descrito Carter & Ackland (1994) e Marfell-Jones em 2006. [46,47] No que diz respeito à avaliação funcional específica a avaliação do

deslize seguiu o protocolo de Carzola em 1993 [48], a velocidade no ensaio de 30 minuto de nado seguiu o protocolo de Olbrecht em 2000 [49] e, por fim, o índice de nado o protocolo de Costill em 1985. [50]

Por outro lado, os conjuntos de dados referentes à recolha da saliva dos atletas da modalidade de natação e de futebol foram sujeitos a outro tipo de cuidados, uma vez que se trata de uma recolha de saliva que irá posteriormente para análise laboratorial:

Cuidados como:

- Não fazer uma refeição 1 hora antes da recolha;
- Evitar consumir álcool nas 24 horas antes da recolha;
- Não ingerir alimentos, mastigar pastilhas elásticas ou rebuçados nos períodos de 30 a 45 minutos antes da recolha;
- Não lavar os dentes com pasta dentífricas 2 horas antes das recolhas, sendo apenas permitido bochechar a boca com água 10 minutos antes da recolha;
- Documentar a prescrição médica que esteja a seguir;
- Cada indivíduo deverá permanecer sentado durante o processo de recolha de saliva. É recolhida toda a quantidade possível, sendo ela produzida de forma passiva e colocada num recipiente apropriado, durante 2 minutos.

São cuidados que o atleta deve ter em consideração e cumprir, para que a recolha não seja contaminada. Os frascos utilizados são de policarbonato com tampa de enroscar de 7.0 ml de capacidade previamente pesados no laboratório.[51] Assim que esgotado o tempo de recolha, o recipiente é fechado e colocado numa mala térmica, assegurando que não é sujeito a alterações de temperatura importantes até ser transportado ao laboratório. [52] No laboratório, a quantidade de saliva é medida para a quantificação do volume e fluxo salivar. Posteriormente, é centrifugada e distribuída por dois tubos Eppendorf ® antes de ser congelada a -20ºC, para análise posterior. É importante referir ainda, que todas as amostras do mesmo sujeito são analisadas no mesmo ensaio de forma a minimizar o erro de medida.

No momento antecedente à análise, as amostras de saliva são descongeladas à temperatura ambiente e centrifugadas (3000 rpm) antes da realização da mesma.

Posto isto, o método utilizado foi a ELISA competitiva [53] seguindo os procedimentos sugeridos pelo fabricante.

Relativamente à análise do cortisol salivar, cada placa é revestida com anticorpos, anti-cortisol. Uma vez que, o cortisol está ligado a uma peroxidase, o cortisol *standards*, bem como o existente nas amostras competem com ele através das ligações ao anticorpo. A quantidade de cortisol peroxidase é medida pela reação da peroxidase no substrato tetrametilbensidina (TMB). É formada uma cor azul devido à reação que posteriormente passa a amarela assim que adicionada a solução *"stop"* (ácido sulfúrico). A densidade ótica foi realizada num leitor ELISA (Elx 800) a 450 nm.

Na determinação da testosterona salivar, foi usado o *Expanded Range Salaviry Testosterone Enzime Immunoassay Kit as Salimetrics* ®, USA. Esta segue os procedimentos descritos acima para o cortisol salivar, considerando os aspetos específicos dos reagentes e dos *standards*. Seguindo a mesma linha de ação do cortisol salivar, uma placa de 96 poços é coberta com anticorpos de anti-testosterona humana. A testosterona *standards* e amostra de testosterona competem com a testosterona ligada à peroxidase na adesão ao anticorpo. A ligação testosterona-peroxidase é medida pela reação da peroxidase com o substrato TMB, formando uma cor azul e assim que adicionada a solução "*stop"* (ácido sulfúrico) a cor passa a amarela.

De um modo idêntico ao referido anteriormente, na análise da imunoglobulina A (IgA), a placa é revestida por anticorpos anti-IgA humano. Posto isto, ocorre a competição entre a IgA standards e amostra de IgA devido à ligação à peroxidase na adesão ao anticorpo. Assim, devido à reação da peroxidase com o substrato TMB, a ligação IgA-peroxidase forma uma cor azul na reação, passando para amarela assim que adicionada a solução "*stop"*.

Em todos os processos acima referidos, é importante realçar que a quantidade de peroxidase detetada é inversamente proporcional à quantidade de cortisol, testosterona e IgA presentes nas amostras, e o cálculo da concentração dos três marcadores biológicos das amostras foi feito através da interpolação dos valores das respetivas densidades óticas.

Capítulo V
Resultados e Discussão

Capítulo V

Resultados e Discussão

Este Capítulo está dividido em três secções, referentes aos conjuntos de dados 1, 2 e 3, respetivamente dados cineantropométricos e funcionais, marcadores biológicos salivares imunitários (sIga) e hormonais (sC e sT) em jovens atletas da modalidade de natação e jovens atletas da modalidade futebol, respetivamente. Nessas secções são apresentados os resultados da análise quimiométrica e a respetiva interpretação e discussão.

5.1 Características Cineantropométricas em Jovens Nadadores (dados 1)

O ponto alvo desta análise está diretamente relacionado com a identificação de padrões nas características cineantropométricas em atletas de elevado nível competitivo de natação pura desportiva de ambos os sexos, quando consideradas as variáveis funcionais. Este conjunto de dados corresponde a 105 atletas, dos quais 20 atletas do sexo feminino e 85 do sexo masculino. As variáveis demográficas idade, e de desempenho: capacidade de deslize subaquático após impulso na parede, velocidade máxima numa distância de 15m sem impulsão, velocidade no teste de 30 minutos de nado contínuo (T30) e o índice de nado no teste T30 são consideradas variáveis de resposta (ou de etiquetagem) e não são incluídas na análise de componentes principais.

Na Tabela 5.1 são apresentados os resultados relativos às variáveis do conjunto de dados com recuperação excedendo 80%, após aplicação da análise de componentes principais. Podemos observar que as duas primeiras componentes, PC1 e PC2, apresentam uma recuperação da informação de 53.8% o que é considerado aceitável para representação gráfica, com mencionado na secção 3.2.

Tabela 5.1. Resultado do PCA referente aos valores próprios e percentagem de recuperação da informação para cada componente principal.

#	Valor Próprio (λ_i)	Variância explicada (%)	Variância explicada cumulativa (%)
PC1	**10.182**	44.3	**44.3**
PC2	2.196	9.5	**53.8**
PC3	2.003	8.7	62.5
PC4	1.852	8.1	70.6
PC5	1.015	4.4	75.0
PC6	0.964	4.2	79.2
PC7	0.785	3.4	82.6

Relativamente ao critério de $\lambda > 1$, abordado no Capítulo 2, secção 2.2.4.2, podemos concluir que seriam necessárias 5 componentes para descrever os dados. Com base no *screeplot*, a primeira componente principal (PC1) com $\lambda > 10.182$ seria suficiente; a Figura 5.1 apresenta a respetiva visualização gráfica.

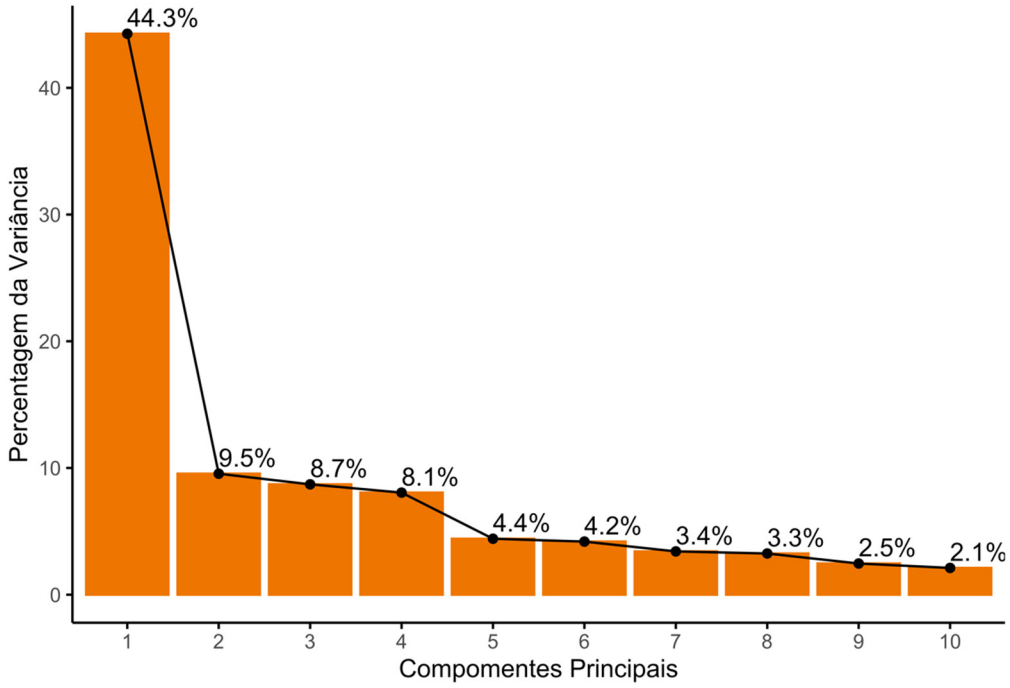

Figura 5.1. Representação gráfica do critério *ScreePlot*.

Tendo em consideração estes resultados, foram consideradas as primeiras duas PC's para as representações que se seguem. A Figura 5.2 posiciona-nos os atletas no novo sistema de eixos.

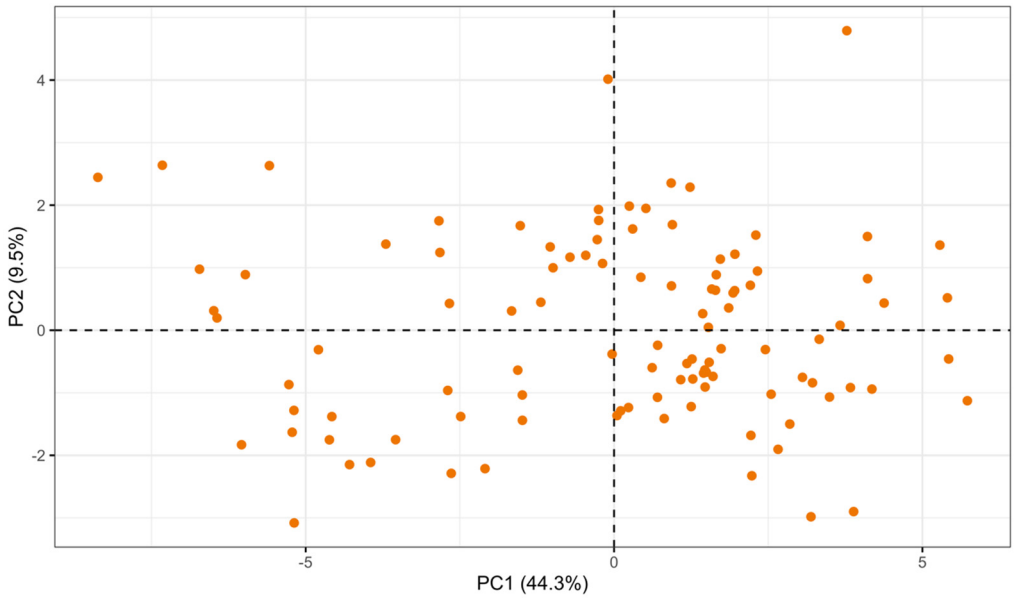

Figura 5.2. Representação dos objetos (atletas) num sistema de eixos a duas dimensões. O eixo PC1 representa a primeira componente com recuperação de 44.3% e o eixo PC2 representa a segunda componente principal com recuperação de 9.5%.

Antes de se proceder a análise de agrupamentos, verificou-se qual era o impacto e dependência de cada variável original nas componentes principais selecionadas, ou seja, PC1 e PC2 (Figura 5.3). Este procedimento corresponde à denominada análise das *loadings*.

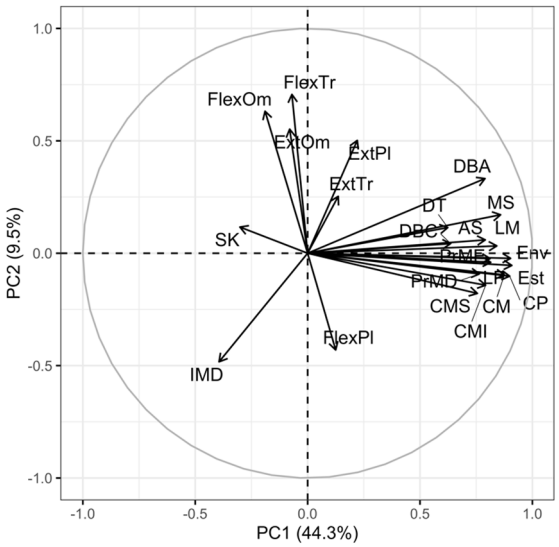

Figura 5.3. Representação gráfica da posição das variáveis em relação às duas primeiras componentes principais.

A representação gráfica (Figura 5.4) obtida mostra-nos que as variáveis cineantropométricas (MS, Est, As, Env, CM, LM, CP, LP, CMS, CMI, DBA, DBC, DT) e a força de preensão (PrMD, PrME) são explicadas pela PC1. Por outro lado, medidas de avaliação funcional como a amplitude na flexão e extensão (FlexPl, FlexOm, FlexTr, ExtPl, ExtOm, ExtTr) e a massa adiposa subcutânea (SK) apresentam maior peso explicativo na 2ª componente.

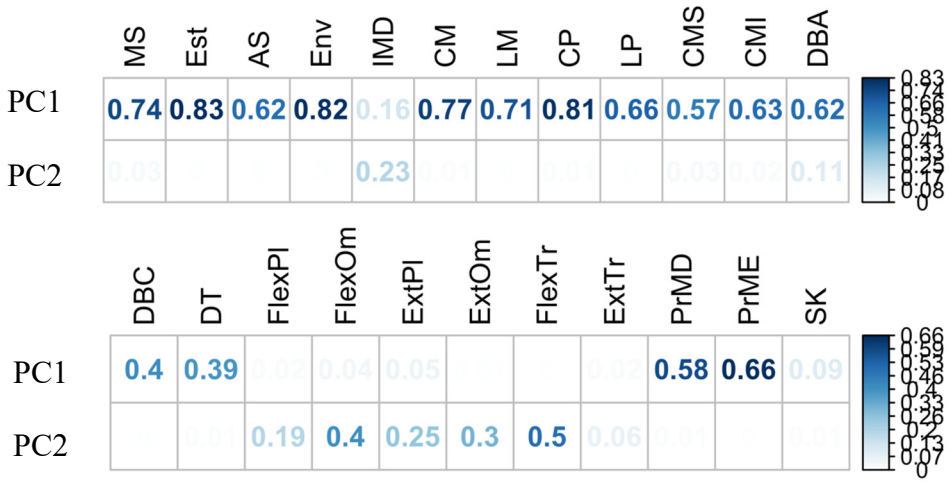

Figura 5.4. Relação entre as variáveis originais e as duas primeiras componentes principais.

Adicionalmente, foi realizada a análise de agrupamentos, que permite agrupar os atletas pelas suas semelhanças, permitindo assim identificar diferentes grupos com características distintas. É importante referir que foi usada a distância euclidiana e a ligação *Ward* para representar a dissimilaridade (distância) entre os atletas.

O dendrograma relaciona, então, os 105 nadadores, considerando todas as variáveis exceto a idade, o deslize, a velocidade máxima, a velocidade T30 e o índice de nado T30 por terem sido consideradas variáveis de resposta, servindo para caracterizar, posteriormente, os grupos. Na Figura 5.5 encontra-se a representação gráfica da formação dos grupos através desta análise HCA. A inspeção da Figura, torna evidente a formação de 4 grupos bem distintos.

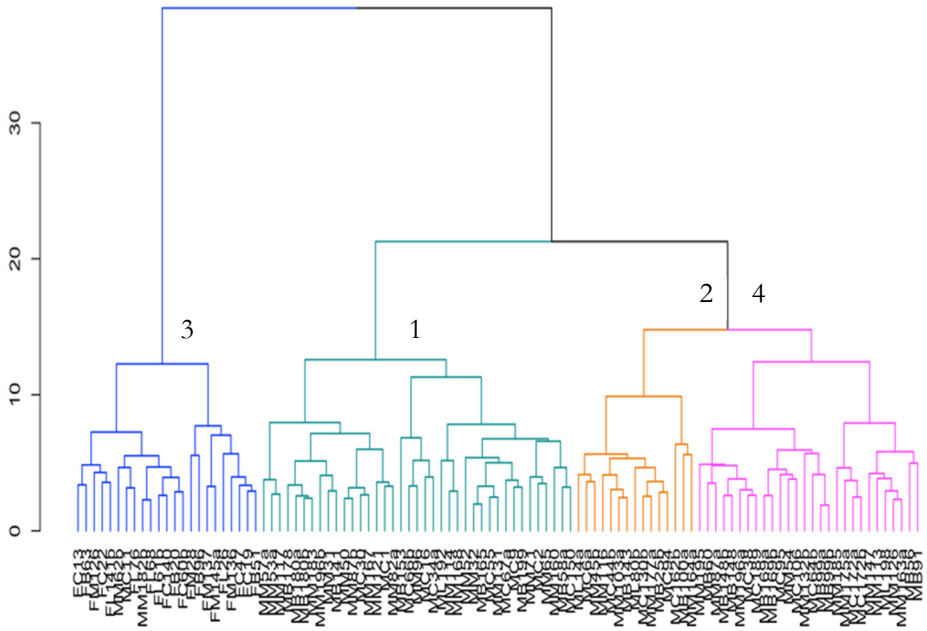

Figura 5.5. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento dos atletas de natação pura desportiva. As diferentes cores representam os diferentes grupos formados, com base na semelhança dos atletas.

A Tabela 5.2 apresenta um resumo da informação respeitante a cada grupo, apresentando-se as médias para cada um deles, permitindo uma melhor compreensão das respetivas características. O grupo 1 é constituído por 39 atletas, o grupo 2 por 15 atletas, o grupo 3 por 23 atletas e o grupo 4 por 28 atletas. Com uma análise mais detalhada, observamos que o grupo 2 é o grupo que apresenta a velocidade média mais elevada no protocolo T30. Relativamente às outras variáveis, é observável que estes atletas se destacam nas variáveis flexão do ombro (FlexOm), extensão plantar (ExtPl), extensão do ombro (ExtOm) e flexão do tronco (FlexTr) apresentando valores médios superiores aos restantes grupos e em sentido oposto, apresenta uma média inferior na flexão plantar (FlexPl). Podemos concluir que estas medidas neste conjunto de dados foram diferenciadoras para que os atletas que constituem o grupo 2 se destacassem por um maior rendimento neste ensaio T30.

Tabela 5.2. Sumário do resultado dos grupos formados pelo dendrograma. Assinalam-se os maiores e menores valores médios para cada variável, respetivamente pelas cores verde e vermelha.

Variáveis	Grupo 1	Grupo 2	Grupo 3	Grupo 4
N (Total)	39	15	23	28
Id	**15.2 ± 0.5**	14.8 ± 0.7	**13.1 ± 0.6**	15 ± 0.3
MS	**68 ± 6**	61 ± 5	**51 ± 5**	59 ± 5
Est	**176 ± 5**	173 ± 4	**161 ± 5**	169 ± 4
SK	66 ± 16	56 ± 12	**80 ± 23**	**54 ± 12**
AS	**90 ± 2**	88 ± 3	**83 ± 3**	86 ± 3
Env	**181 ± 6**	178 ± 5	**164 ± 5**	173 ± 4
IMD	4.4 ± 0.3	4.4 ± 0.1	**4.7 ± 0.3**	4.4 ± 0.2
CM	**19.7 ± 0.8**	18.9 ± 0.6	**17.3 ± 0.6**	19.2 ± 0.9
LM	**8.2 ± 0.4**	8. ± 0.5	**7.1 ± 0.3**	8.1 ± 0.4
CP	**26.2 ± 0.8**	25 ± 1	**23 ± 1**	25 ± 1
LP	**9.7 ± 0.4**	9.3 ± 0.5	**8.4 ± 0.5**	9.3 ± 0.6
CMS	**71 ± 3**	70. ± 2	**65 ± 3**	67 ± 2
CMI	**86 ± 4**	84 ± 2	**78 ± 4**	83 ± 3
DBA	**40 ± 3**	39 ± 1	**34 ± 2**	39 ± 2
DBC	**27 ± 2**	25 ± 1	**23 ± 2**	24 ± 2
DT	**21± 2**	19 ± 1	**17 ± 1**	20 ± 2
FlexPl	37 ± 9	**22 ± 5**	33 ± 6	26 ± 8
FlexOm	11 ± 9	**42 ± 20**	23 ± 26	9 ± 4
ExtPl	23 ± 7	**42 ± 8**	**17 ± 3**	27 ± 9
ExtOm	86 ± 17	**104 ± 32**	89 ± 24	**74 ± 16**
FlexTr	11 ± 7	**14 ± 7**	11 ± 9	**6 ± 7**
ExtTr	**45 ± 12**	40 ± 7	37 ± 8	**32 ± 9**
PrMD	**44 ± 6**	36 ± 6	**28 ± 5**	43 ± 7
PrME	**43 ± 6**	37 ± 6	**28 ± 4**	40 ± 7
Velmáx	1.7 ± 0.1	**1.73 ± 0.06**	**1.57 ± 0.07**	1.67 ± 0.08
Desl	7.1 ± 0.7	**8 ± 1**	**7 ± 1**	6.8 ± 0.9
VelT30	1.32 ± 0.08	**1.35 ± 0.09**	**1.19 ± 0.09**	1.29 ± 0.09
INT30	**2.4 ± 0.3**	2.3 ± 0.3	**2.1 ± 0.2**	2.3 ± 0.3

A representação a duas dimensões dos grupos formados pelo dendrograma encontra-se na Figura 5.6. Esta visualização apresenta os grupos de atletas bem como as variáveis sob análise.

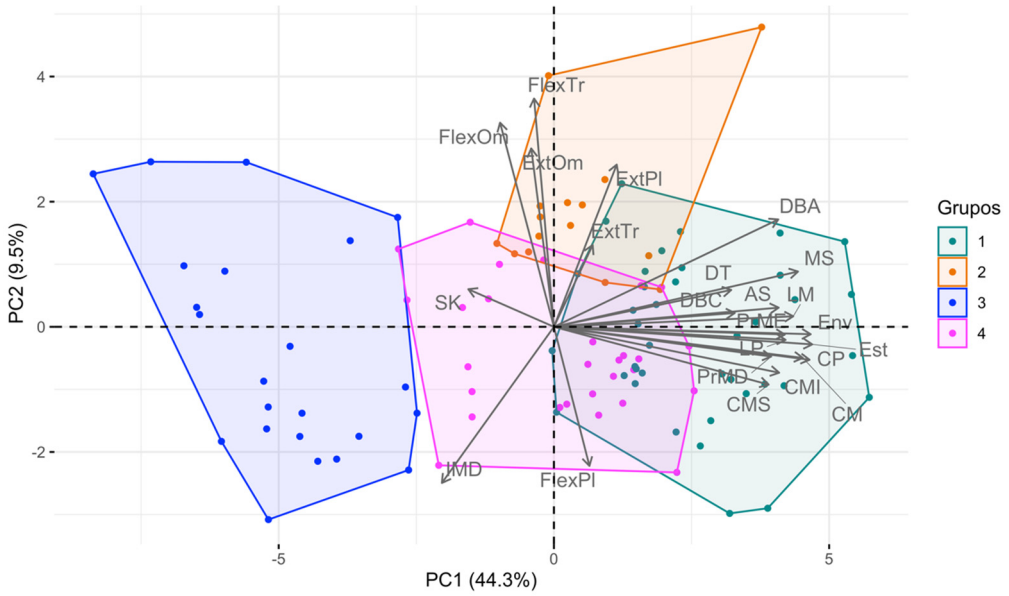

Figura 5.6. Representação gráfica dos grupos de atletas formados dendrograma num eixo de a duas dimensões aplicado aos nadadores, com uma recuperação de 53.8% da informação inicial. As cores dos grupos representam a partição da amostra nos mesmo grupos formados pelo HCA e representados no dendrograma da Figura 5.5.

De um modo geral, a Figura 5.6 permite uma visualização gráfica de tudo o que foi referido anteriormente, isto é, representação das variáveis com os grupos de atletas formados pela análise de agrupamentos.

O grupo 2 (laranja) destaca-se relativamente aos outros em vários aspetos. Este grupo é o que apresenta maior velocidade no teste T30 e na velocidade máxima, e quando inspecionamos o conjunto das variáveis, observamos que se destacam pelos valores superiores do ombro (FlexOm), extensão plantar (ExtPl), extensão do ombro (ExtOm) e flexão do tronco (FlexTr). Recentemente [54] foi reportado que valores favoráveis de flexibilidade, com repercussão no alinhamento corporal favorecem a menor resistência hidrodinâmica, ou seja, valores superiores na mobilidade do ombro e da articulação tibiotársica, melhoram a postura do atleta, reduzindo o impacto das forças que resistem ao deslocamento no meio aquático.

Consequentemente, quanto menor for a resistência que o nadador oferecer em relação à água, maior economia e velocidade de nado. Ainda sobre este grupo de atletas, podemos observar que a flexão plantar (FlexPl) apresenta um valor inferior quando comparado com os outros grupos. Esta relação corrobora vários estudos, nomeadamente um [55] que relaciona a amplitude da mobilidade da articulação tibiotársica e da força muscular, no desempenho da ação propulsiva gerada pelos membros inferiores, nomeadamente no nado de crol. Neste estudo a velocidade diminuiu significativamente quando perante a limitação da flexão plantar durante a ação descendente do movimento dos membros inferiores. Outro estudo [56] revelou uma correlação significativa entre a flexibilidade de flexão plantar do tornozelo e a velocidade na ação dos membros inferiores, não tendo, no entanto, encontrado associação significativa entre a mobilidade da tibiotársica e a velocidade em 50m. Este tema tem sido muito debatido, pois vários treinadores acreditam que um nadador com maior flexibilidade na articulação do tornozelo apresenta vantagem biomecânica na ação das pernas na geração de apoio propulsivo, conduzindo a deslocamentos mais velozes velocidade. [56,57,58] Este aspeto parece constituir um requisito técnico para a obtenção de elevado desempenho em natação desportiva.

Relações entre a flexão do ombro e tronco e extensão do ombro e plantar levam-nos a concluir que apresentam um impacto positivo na velocidade final no teste T30. Por outro lado, um valor mais reduzido de flexão plantar aparentemente não foi suficiente para influenciar a velocidade neste protocolo T30. Tal pode ser explicado parcialmente por um trabalho [59] que cinge a cerca de 10% a contribuição da pernada para a propulsão do nadador em crol.

O atleta no teste T30, como o nome indica, está 30 minutos em nado contínuo, e a característica determinante para este teste é a sua capacidade aeróbica. Uma vez que o teste da velocidade máxima na distância de 15 m é igual ao T30, podemos afirmar que o teste T30 descreve muito bem a capacidade do atleta. Embora a importância da contribuição aeróbica seja bem aceite em competições de longa distância, a antropometria e a capacidade aeróbica foram recentemente destacadas nos jovens nadadores como variáveis determinantes do desempenho em provas de curta distância. [60]

Comparando os resultados obtidos entre os grupos 1, 3 e 4 observamos que o destaque da velocidade no protocolo T30 está no grupo 1, sendo este o grupo com as características cineantropométricas mais elevadas, ou seja, é o grupo com os atletas que apresentam maiores dimensões corporais. Na Figura 5.4 reparamos que a variável estatura era a que tinha mais peso na componente principal 1 sendo a característica mais importante e determinante no desempenho dos atletas neste grupo. Muitas outras correlacionam diretamente com esta. Note-se que vários estudos referem que os nadadores são, em média, maiores em comparação com a média da população [61] e têm uma taxa crescimento superior à média. [62] Um estudo em nadadores olímpicos de elite conclui que são mais altos do que os nadadores de nível competitivo inferior. Neste estudo [63], a estatura, a altura sentado, o diâmetro do tronco e a relação entre a altura do tronco e o comprimento do tronco apresentaram valores mais elevados no grupo olímpico, tanto em homens como em mulheres. Isto conduziu ao conceito de que os nadadores de elite são altos, de ombros largos e quadris estreitos, quando comparados com a população normal.

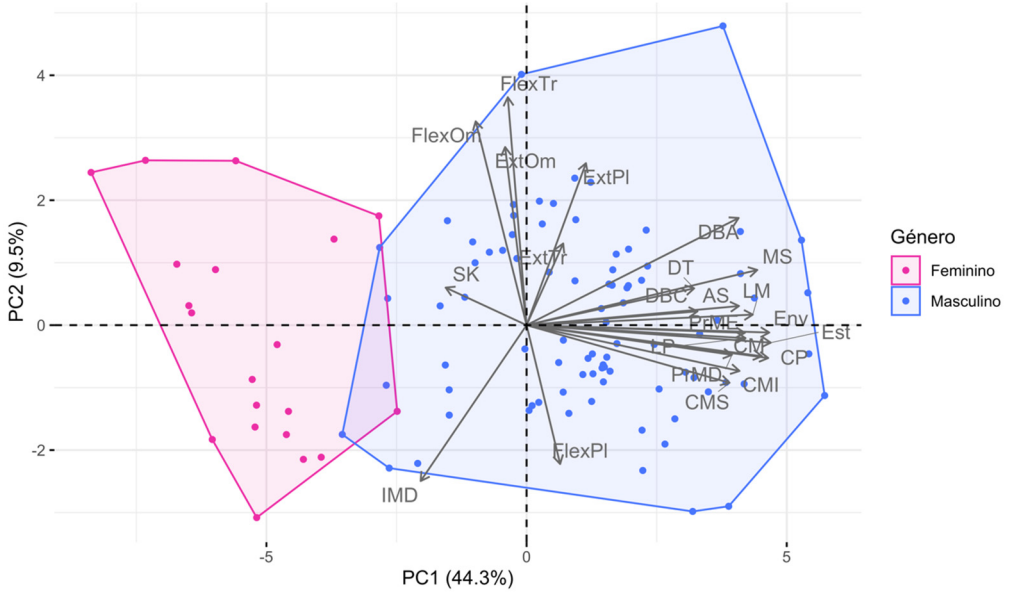

Figura 5.7. Representação gráfica do resultado do PCA num eixo de duas dimensões aplicado aos atletas de natação, com uma recuperação de 53.8% da informação inicial. As cores dos grupos representam o género dos atletas.

A Figura 5.7, faz a divisão entre os atletas do sexo masculino e os atletas do sexo feminino sendo visível que os atletas do sexo feminino apresentam

características cineantropométricas que apontam para menores dimensões quando comparados com o sexo masculino. Note-se que o grupo 3 formado pelo dendrograma é o que contém todas as 20 atletas femininas contendo apenas 3 nadadores masculinos. Prestes et al (2009) relatam que características como massa corporal, estatura e o diâmetro bicristal apresentaram valores mais altos em nadadores do sexo masculino quando comparados com nadadores do sexo femininos. [64] Interessante neste âmbito é o apresentado por um estudo que reporta que o treino físico intensivo pode perturbar o funcionamento hormonal influenciando o crescimento. [61]

5.2 Marcadores Biológicos Salivares em Nadadores durante uma época competitiva (dados 2)

Como referido anteriormente, o conjunto de dados 2 está relacionado com uma recolha de saliva a 16 atletas de natação pura desportiva num período de 29 semanas que corresponde a um macro ciclo competitivo (época de inverno). Os marcadores biológicos extraídos da saliva foram a sIgA, o sC e o sT. Aliada a esta recolha foi também monitorizada a carga de treino, quantificada pelo volume (total de metros nadados ao longo de cada semana).

Este conjunto de dados implicou uma abordagem diferente da anterior. Primeiramente verificámos qual foi a evolução da carga de treino durante as 29 semanas. Através do gráfico de barras da Figura 5.8, podemos constatar que existem três momentos em que a carga de treino foi mais elevada, nomeadamente o período da semana 1 à 8, da 13 à 18 e da 20 à 25.

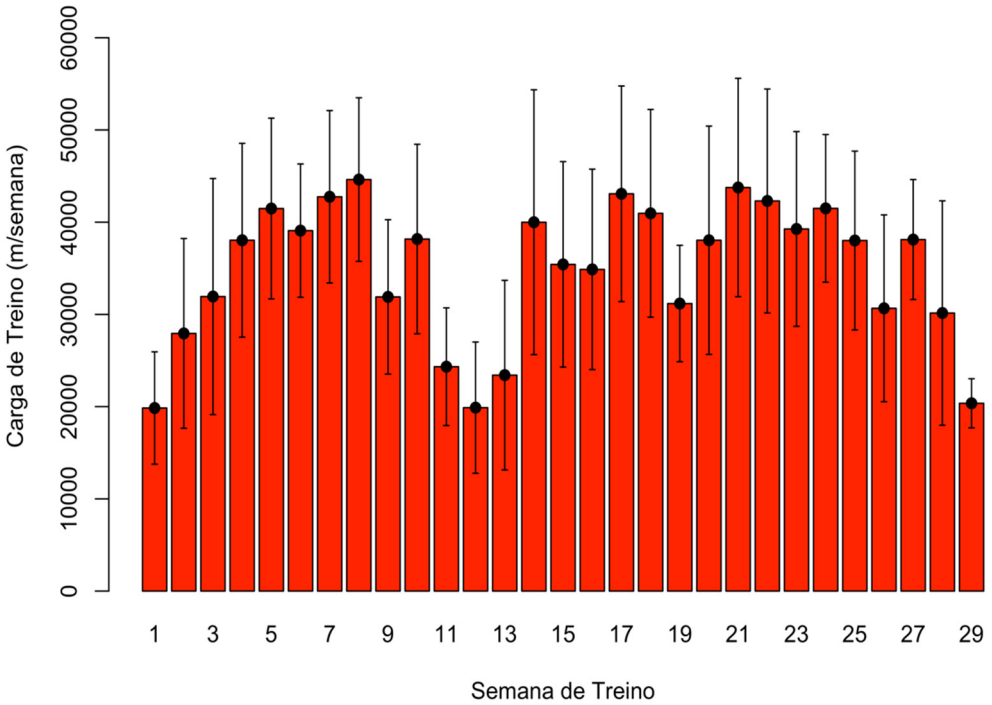

Figura 5.8. Gráfico de barras para a média da carga de treino ao longo das 29 semanas.

Estes incrementos estão relacionados com os momentos pré-competitivos, sendo que a competição mais relevante ocorre no final da 29ª semana. A Figura 5.9 ilustra o comportamento médio do marcador biológico sC (riscas vermelhas) coincidente com carga de treino realizada (barras cinzentas).

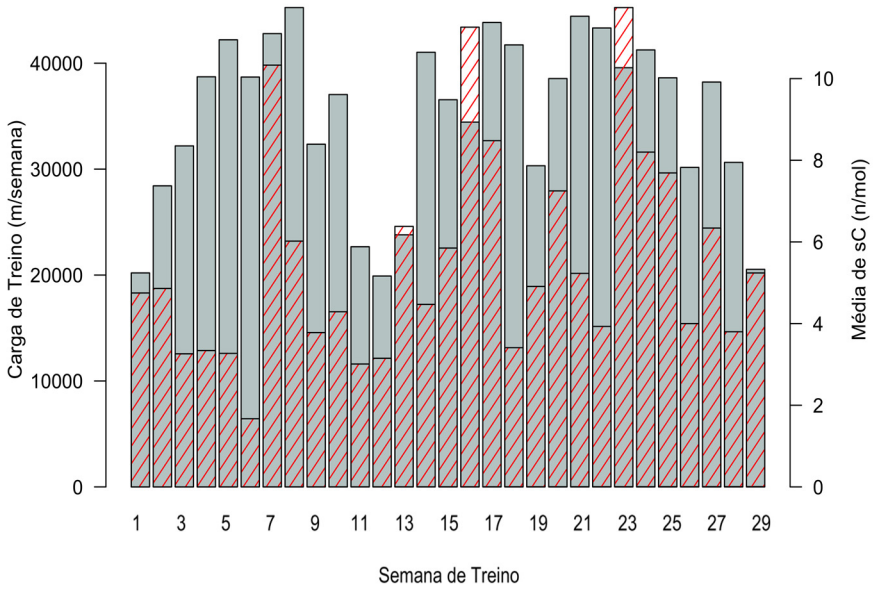

Figura 5.9. Valores médios semanais da carga de treino (cinzento) e do sC (riscas vermelhas)

Observamos a existência de três picos de sC, respetivamente nas semanas 7, 16 e 23. Tal geralmente acontece em resposta a uma elevada carga de treino, o que não é visível na Figura 5.9. Podemos e

- Será que a resposta do sC a
  ocorre com um desfasamento temporal?

Neste contexto, fomos verificar qual é a correlação existente entre a carga de treino e o sC, simulando desfasamento temporal. Na Tabela 5.3, são apresentados os valores das correlações entre a carga de treino e o sC, sem desfasamento, com 1 semana, 2 semanas e 3 semanas de desfasamento (i.e., para uma determinada carga de treino, o teor de sC corresponde ao medido 0-3 semanas

depois). Verifica-se que o desfasamento que maximiza a correlação corresponde a 2 semanas, com um valor apreciável desta última (0.424).

Tabela 5.3. Valores de correlação entre carga de treino e o sC.

Semanas de Desfasamento	Corelação
0	0.262
1	0.406
2	**0.424**
3	0.368

Deste modo, concluímos que os três picos de sC mencionados anteriormente estão positivamente associados com o incremento da carga observado nomeadamente as semanas, 5, 14 e 21, o que corresponde a um desfasamento temporal de aproximadamente 2 semanas. Um estudo pioneiro [1] na modalidade de natação indica que a concentração de sC apresenta valores superiores após a competição, ou seja, depois do atleta estar sujeito a uma carga de treino elevada (competição) os seus valores de sC aumentam. Desta forma, concluímos que no nosso estudo existe igualmente [9] uma correlação entre a carga de treino aplicada e a concentração do sC, com um desfasamento temporal de 1 a 2 semanas.

Um outro dado recolhido neste estudo passou pelo número de ocorrências de infeções do trato respiratório superior (ITRS). O gráfico da Figura 5.10 representa as 29 semanas do estudo e as ocorrências da respetiva semana.

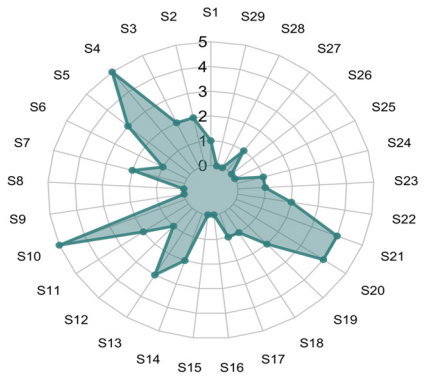

Figura 5.10. Representação do número de ocorrências de ITRS por semana.

Através da representação do número de ocorrências, destacam-se como sendo as que apresentam um maior número de ocorrências as semanas, 4, 5, 10, 13, 20 e 21. Ao verificar a relação entre estas semanas e a carga de treino, verificamos que as semanas 4, 5, 20 e 21 correspondem uma carga de treino elevada e contrariamente, as semanas 10 e 13 apresentam carga de treino mais baixa. Esta conclusão não é estatisticamente comprovada, mas sim decorrendo do observado na representação gráfica.

Com o objetivo de verificar se existem relações entre o número de ocorrências, a concentração de sC, concentração de sIgA e a carga de treino, efetuámos uma análise de agrupamento hierárquica e uma análise de componentes principais. Uma vez que foi demonstrado antes existir um desfasamento entre a carga de treino e o cortisol, foram removidas as cargas de treino relativas às duas últimas semanas para permitir a análise.

O conjunto de dados analisado contém a média das concentrações de sC e sIgA, apenas. Após efetuada a análise de agrupamentos, obtemos o dendrograma da Figura 5.11.

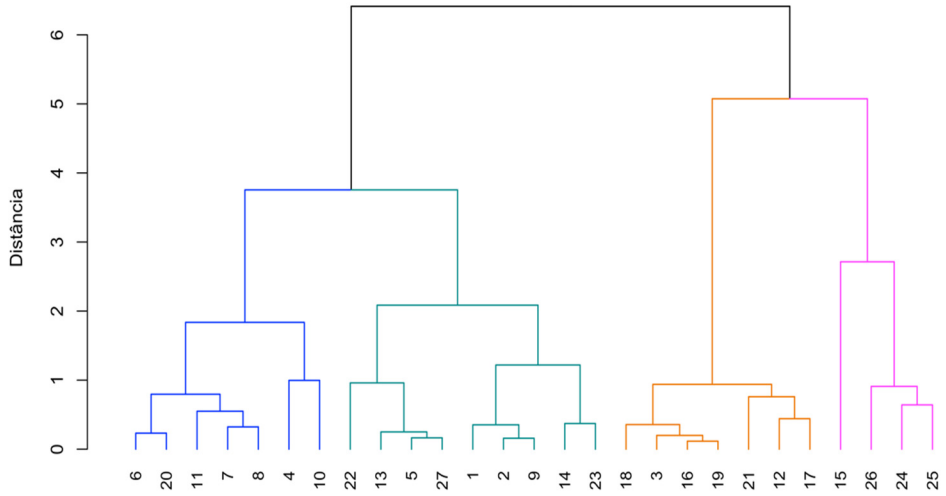

Figura 5.11. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento da média da concentração dos marcadores biológicos (sC e sIgA) em cada semana.

Através do dendrograma podemos observar a formação de quatro grupos bem definidos. A Tabela 5.4 apresenta a média e desvio padrão da concentração de sC e sIgA, a carga de treino, bem como o número de sintomas de cada grupo.

Tabela 5.4. Sumário do resultado dos grupos formados pelo dendrograma da Figura 5.11.

	Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**
Nº de Semanas	9	7	7	4
sIgA (mg/L)	238 ± 16	197 ± 9	256 ± 13	216 ± 32
sC (nmol/L)	3.0 ± 0.3	2.8 ± 0.2	2.1 ± 0.4	4.4 ± 0.3
Carga de treino (m)	33076 ± 8110	34922 ± 8543	37307 ± 6743	36230 ± 7522
Nº de Sintomas	15	11	19	1

Os resultados relativos aos grupos formados no dendrograma refletem algumas diferenças relativas aos marcadores. É possível observar que o grupo 3 corresponde ao conjunto de semanas onde ocorreram um maior número de sintomas de ITRS e que por sua vez é o que apresenta valor médio de sIgA e da carga de treino superiores, sugerindo que possa haver uma resposta imunológica à carga de treino induzindo uma maior ocorrência de infeções do trato respiratório superior. Vale apena referir que a observação de valores de sIgA elevados, poderá estar eventualmente associado à ação imune reativa, devido à existência de episódio de ITRS em curso. Os teores de sC variam entre 2.1 e 4.4 nmol/L, mas a explicação não é óbvia.

Foi efetuada uma divisão dos atletas por quartil relativamente aos valores recolhidos em repouso de sIgA no início da época desportiva, sendo o quartil 1 composto por os atletas com valores mais baixos de sIgA e assim sucessivamente. Ao contrário do que é reportado na literatura, observou-se que que os atletas que apresentaram valores mais baixos no momento antes de começar a época não corresponderam aqueles tiveram mais ocorrências de ITRS (Tabela 5.5).

Tabela 5.5. Contagem da ocorrência de ITRS nos grupos formados por quartis de sIgA no momento antes do começo de época.

	Quartil 1	**Quartil 2**	**Quartil 3**	**Quartil 4**
Nº de Sintomas	9	7	20	9

Tal vem contrariar vários estudos que indicam que os atletas que apresentam valores de sIgA mais baixos que os outros, adoecem mais. [66,67,68] É de salientar que por vezes, os episódios de ITRS, não correspondem a infeções, podendo ser confundidas com outro tipo de ocorrências, nomeadamente alergias ou outras repostas inflamatórias que podem vir a ser confundidas com ITRS. [69]

Existem vários fatores que podem influenciar a ocorrência de ITRS, como o fator ambiental, geográfico, social e até a responsabilidade do próprio atleta. Uma vez que se trata de respostas hormonais e imunitárias individualizadas, quando se analisa o conjunto em bruto podemos perder o caracter específico da resposta individual. Por estes motivos, estas análises nem sempre estão de acordo umas com as outras, visto que, existem vários fatores como os descritos anteriormente que podem influenciar estes resultados.

Apesar de não serem totalmente conclusivos, estes resultados podem sugerir certas indicações que nos permitirão realizar trabalho adicional ou prosseguir a investigação em certas direções mais prometedoras, nomeadamente com um enfoque mais detalhado em características individuais de cada atleta.

5.3 Marcadores Biológicos Salivares em Jovens Atletas de Futebol (dados 3)

O terceiro conjunto de dados é referente à recolha de saliva obtida em jovens praticantes de futebol, pertencentes à mesma academia de formação desportiva. Fizeram parte da amostra três escalões competitivos diferentes, sub-15 constituído por 17 jogadores, sub-17 com 22 jogadores e sub-19 com 18 jogadores. As recolhas foram efetuadas no final de cada mês de treino sempre à mesma hora (16h -17h). Foi sempre assegurado um intervalo mínimo de 36h entre a última sessão de treino ou jogo e a colheita de saliva. Os marcadores biológicos analisados foram os hormonais: sC e sT e imunitário, sIgA. É de salientar que a época começou em agosto para todos os escalões e terminou em março para os sub-15 e em abril para os sub-17 e sub-19.

Relativamente a este conjunto de dados, procuramos saber se as respostas dos marcadores biológicos são idênticas ou distintas entre os três escalões em estudo. Para tal elaborámos uma análise hierárquica de agrupamentos e análise de componentes principais, a fim de averiguar se existem diferenças tanto hormonais (sC e sT) como imunitárias (sIga) quando comparados os escalões.

Na Tabela 5.6 encontram-se os resultados da recuperação relativa a cada componente principal após a análise PCA. Assim, temos uma recuperação da informação de 39.9% (relativamente baixa), nas duas primeiras componentes principais, PC1 e PC2.

Tabela 5.6. Resultado do PCA referente aos valores próprios e percentagem de recuperação da informação para cada componente principal.

#	Valor Próprio (λ_i)	Variância explicada (%)	Variância explicada cumulativa (%)
PC1	6.178	25.741	25.741
PC2	3.409	14.203	39.944
PC3	2.292	9.549	49.493
PC4	1.743	7.262	56.755
PC5	1.595	6.646	63.401
PC6	1.301	5.421	68.822
PC7	1.104	4.600	73.422
PC8	0.945	3.937	77.359
PC9	0.844	3.518	80.877

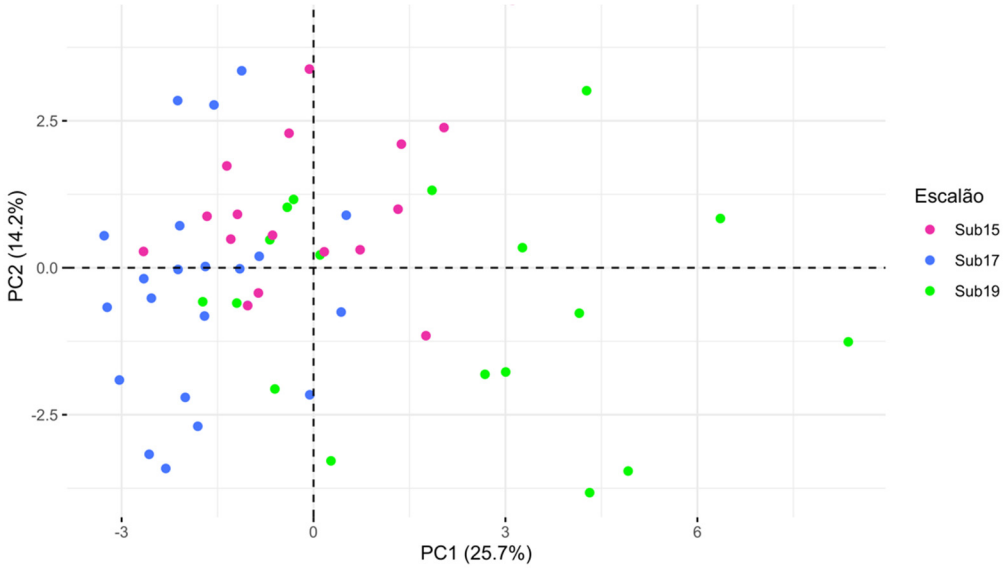

Figura 5.12. Representação dos atletas dos respetivos escalões num sistema de eixos a duas dimensões. O eixo PC1 apresenta uma recuperação de 25.7% e o eixo PC2 apresenta uma recuperação de 14.2%.

Pela posição espacial dos atletas representados por pontos, (Figura 5.12) correspondentes aos três escalões em análise, observamos que de um modo geral são formados 3 grupos. Os atletas do escalão sub-19 (verde) encontram-se distribuídos mais à direita, os atletas sub-17 (azul) encontram-se mais à esquerda e os atletas sub-15 (magenta) encontra-se ao centro e em cima. Uma vez que, os escalões dividem os atletas por idades, pode levar-nos a concluir que quando se estuda o conjunto completo de dados por inteiro, i.e, as variáveis hormonais e imunitárias mostram-se diferenciadores em função das idades.

Aliada à análise anterior, e partindo para mais detalhe verificamos que o marcador hormonal sC é o que manifesta maior impacto na primeira componente principal (Figura 5.13). A PC1 é a componente com maior percentagem de recuperação da informação e é maioritariamente explicada pela variável sC.

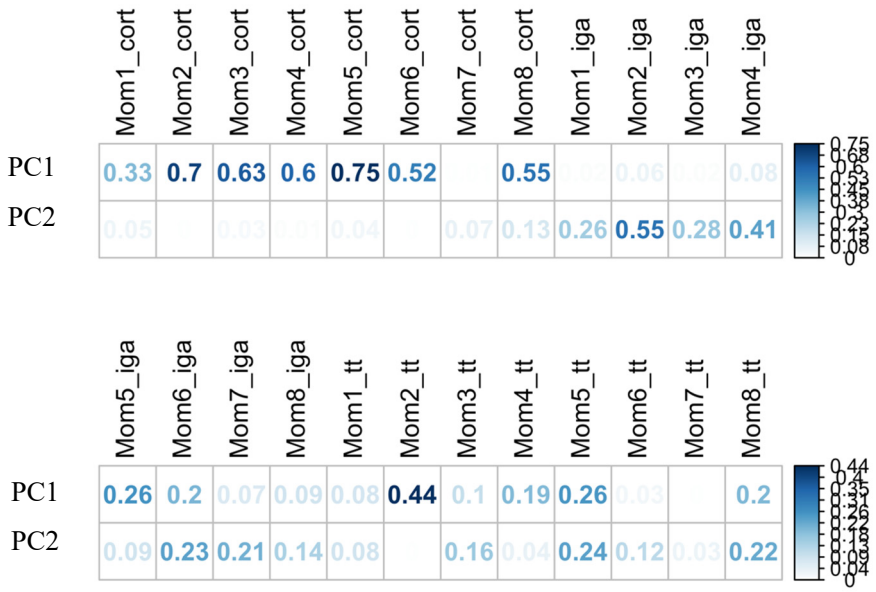

Figura 5.13. Relação entre as variáveis originais e as duas primeiras componentes principais para o conjunto de dados do futebol.

A representação gráfica do dendrograma (Figura 5.14) revela a formação de três grupos: em primeiro lugar temos a separação dos atletas em dois grupos, o grupo 1 e o grupo formado pelo grupo 2 e 3 e posteriormente temos também a separação e a formação do grupo 2 e 3.

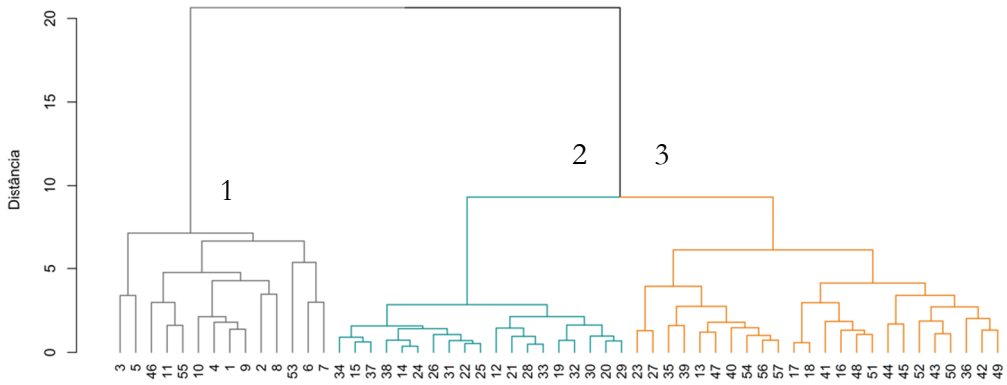

Figura 5.14. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento dos atletas de futebol para o marcador sC.

Através da análise HCA, fomos verificar o perfil médio comportamental do sC em cada grupo de atletas formado pelo dendrograma. Os perfis representados na Figura 5.15, demostram que o perfil do grupo 1 é distinto dos outros dois, observando-se semelhanças entre os perfis 2 e 3.

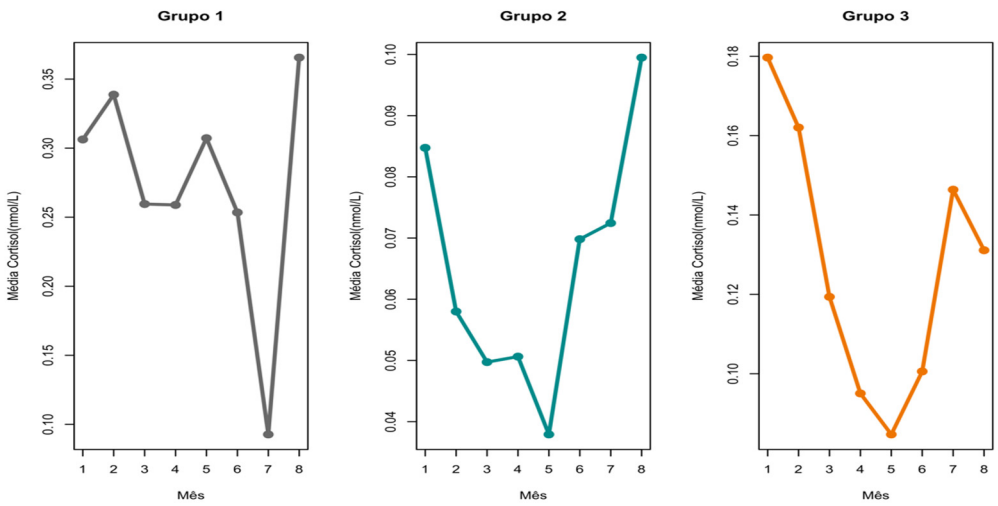

Figura 5.15. Perfis do comportamento médio do sC dos grupos formandos pelo dendrograma da Figura 5.14.

A Figura 5.16 representa o perfil comportamental médio do sC para cada escalão, sendo possível observar semelhanças entre os perfis formados pelos grupos do dendrograma (Figura 5.14). Tal leva-nos a concluir que os grupos formados no dendrograma correspondem a um comportamento típico do sC dos atletas de cada escalão uma vez que eles foram sujeitos a condições de stress diferentes. Noutros ternos, o grupo 1 corresponde a atletas do escalão sub-19, o grupo 2 a atletas do escalão sub-17 e o grupo 3 a atletas do escalão sub-15. Tal pode diretamente ser confirmado pela inspeção dos atletas constituintes de cada grupo.

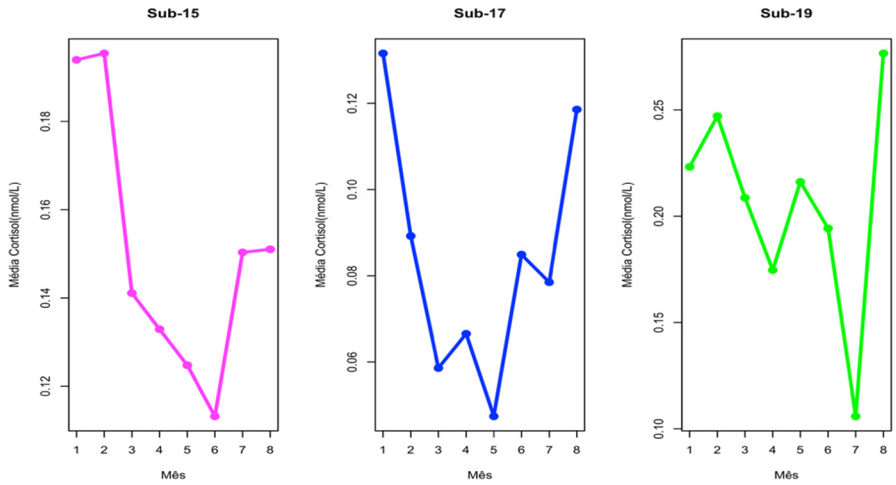

Figura 5.16. Perfis do comportamento médio do sC para cada escalão, sub-15, sub-17, sub-19.

Como referido anteriormente, e embora o processo de treino seja idêntico o envolvimento competitivo apresenta diferentes exigências induzindo igualmente níveis de stress diferenciados. Os atletas do escalão sub-15 no início do estudo que corresponde a agosto (1) e setembro (2) estão sujeitos à pressão da captação e de assegurar o seu lugar na equipa. Este condicionalismo parece elevar o nível de stress tal como é representado na Figura 5.16. Relativamente ao escalão sub-17, os atletas no final da época disputam o lugar de campeão com as restantes equipas e os sub-19 disputam a não despromoção do campeonato nacional. Nestes dois escalões é coincidente a existência de elevado stress no final da época, como é sustentado pelo valor médio do sC ser elevado, no fim da época, em ambas equipas. Esta relação entre o aumento do sC em momentos de stress ou elevada carga de treino está em consonância com o reportado na literatura. [70,65]

Com o intuito de avaliar se os restantes marcadores biológicos seguem um padrão de resposta identificável, como o do sC, analisamos também os marcadores sT e sIgA.

A análise da sT (Figura 5.17), apresenta o dendrograma da análise hierárquica de agrupamentos. Tal com o que ocorre no sC, regista-se a formação de 3 grupos distintos e bem separados.

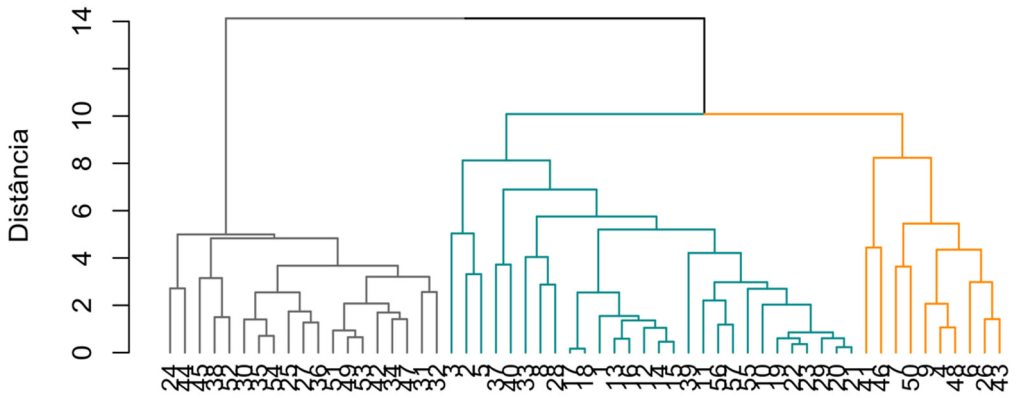

Figura 5.17. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamentos dos atletas de futebol para o marcador biológico hormonal sT.

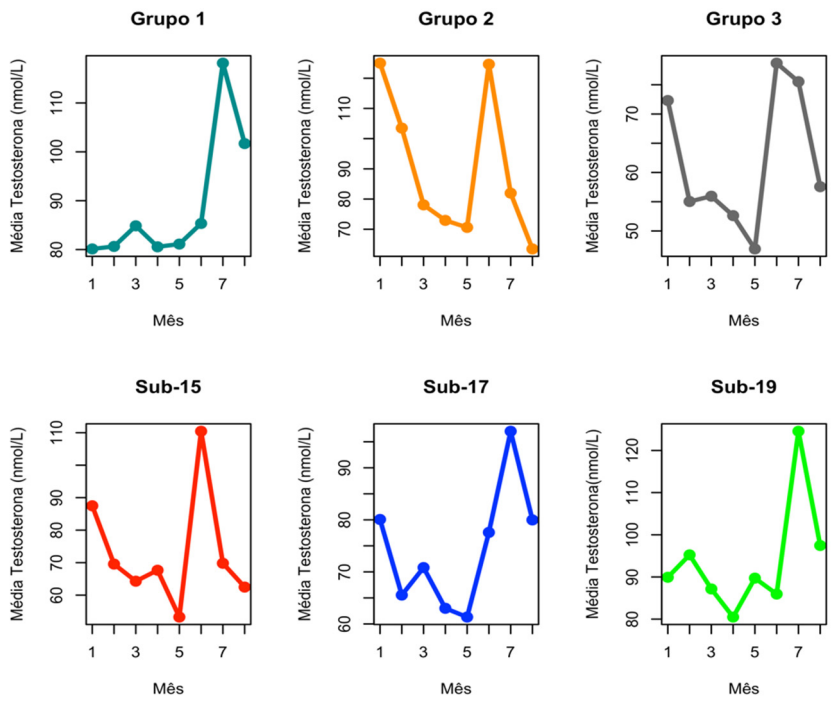

Figura 5.18. Perfis do comportamento médio da sT referentes aos grupos formandos pelo dendrograma da Figura 5.17 e os perfis do comportamento médio da sT para cada escalão, sub-15, sub-17, sub-19.

Comparando os perfis (Figura 5.18) formandos pelos grupos do dendrograma e os perfis comportamentais médios da sT para cada escalão, verificamos que embora existindo alguma semelhança, não nos permite a admissão de que os grupos correspondem aos escalões de competição. Note-se que os perfis dos escalões são muito idênticos, levando-nos a concluir que os níveis de sT nestas idades não são diferenciados. Estudos relatam que os jovens apresentam um período crítico de maturação na puberdade. Durante a restante adolescência, e cerca de um ano a ano e meio após esse período, os níveis sT tendem a não apresentar uma discrepância entre idades tão elevada. [71,72]

No que toca ao comportamento do sIgA, a análise de HCA representada pelo dendrograma (Figura 5.19) forma essencialmente dois grandes grupos, indicando não existem diferenças na resposta imunológica entre os escalões.

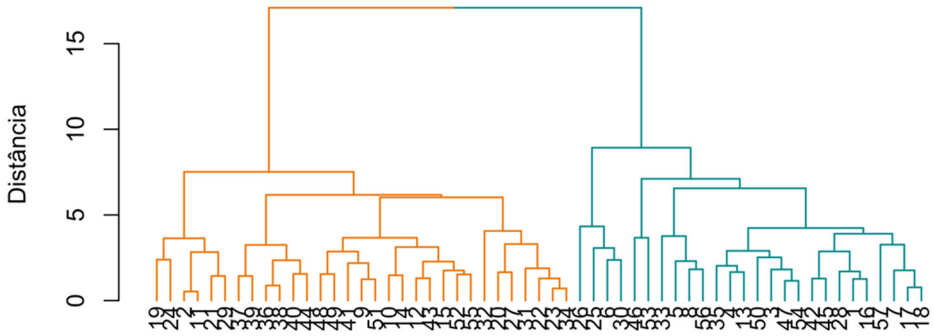

Figura 5.19. Representação gráfica do dendrograma através do método Ward da análise hierárquica de agrupamento dos atletas de futebol para o marcador biológico imunitário sIgA.

Quando analisados os perfis (Figura 5.20) do comportamento médio da sIgA nos grupos e nos escalões, observamos que de facto este comportamento não está relacionado com a idade, ou seja, é independente do escalão.

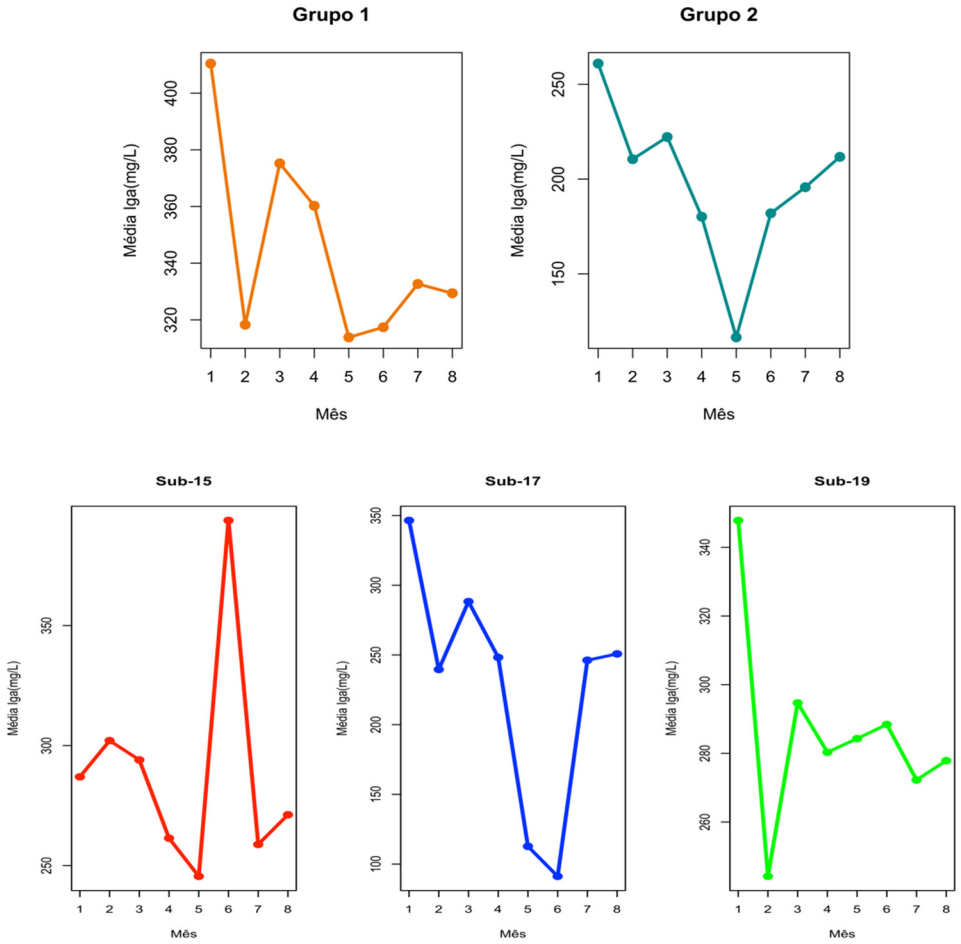

Figura 5.20. Perfis do comportamento médio da sT referentes aos grupos formandos pelo dendrograma da Figura 5.19 e os Perfis do comportamento médio da sIgA para cada escalão, sub-15, sub-17, sub-19.

Uma vez que este marcador imunitário tem sido relacionado com infeções do trato respiratório superior, a Tabela 5.7 apresenta o número de ocorrências de sintomas para cada escalão nos respetivos meses.

Tabela 5.7. Registo de ocorrências de sintomas de ITRS durante a época desportiva de cada escalão.

Mês	Nº de Sintomas		
	Sub-15	**Sub-17**	**Sub-19**
Agosto (1)	5	4	3
Setembro (2)	7	7	3
Outubro (3)	10	4	0
Novembro (4)	10	7	5
Dezembro (5)	12	9	5
Janeiro (6)	11	5	3
Fevereiro (7)	1	3	8
Março (8)	3	12	5
Abril (9)	----	3	4
Total	59	56	36

De um modo semelhante ao estudo referente ao segundo conjunto de dados, este fator imunitário não segue nenhum padrão que relacione a ocorrência de sintomas e a concentração média de sIgA. Uma das possíveis razões explicativas para não ser encontrado nenhum padrão de resposta neste estudo, pode estar relacionado com a recolha da frequência da recolha da amostra, dado que esta acontece no final do mês e pode coincidir ou não com um momento de resposta imunitária a uma ITRS. No entanto, esta relação tem sido muito discutida na literatura e parece ainda não existir consenso. [73]

Outra relação que despertou interesse quando analisada, está relacionada com o volume de treino e a relação que apresenta entre o sC e a sT. A Figura 5.21, engloba a variação do volume de treino, o perfil médio do comportamento do sC e da sT para os diferentes escalões.

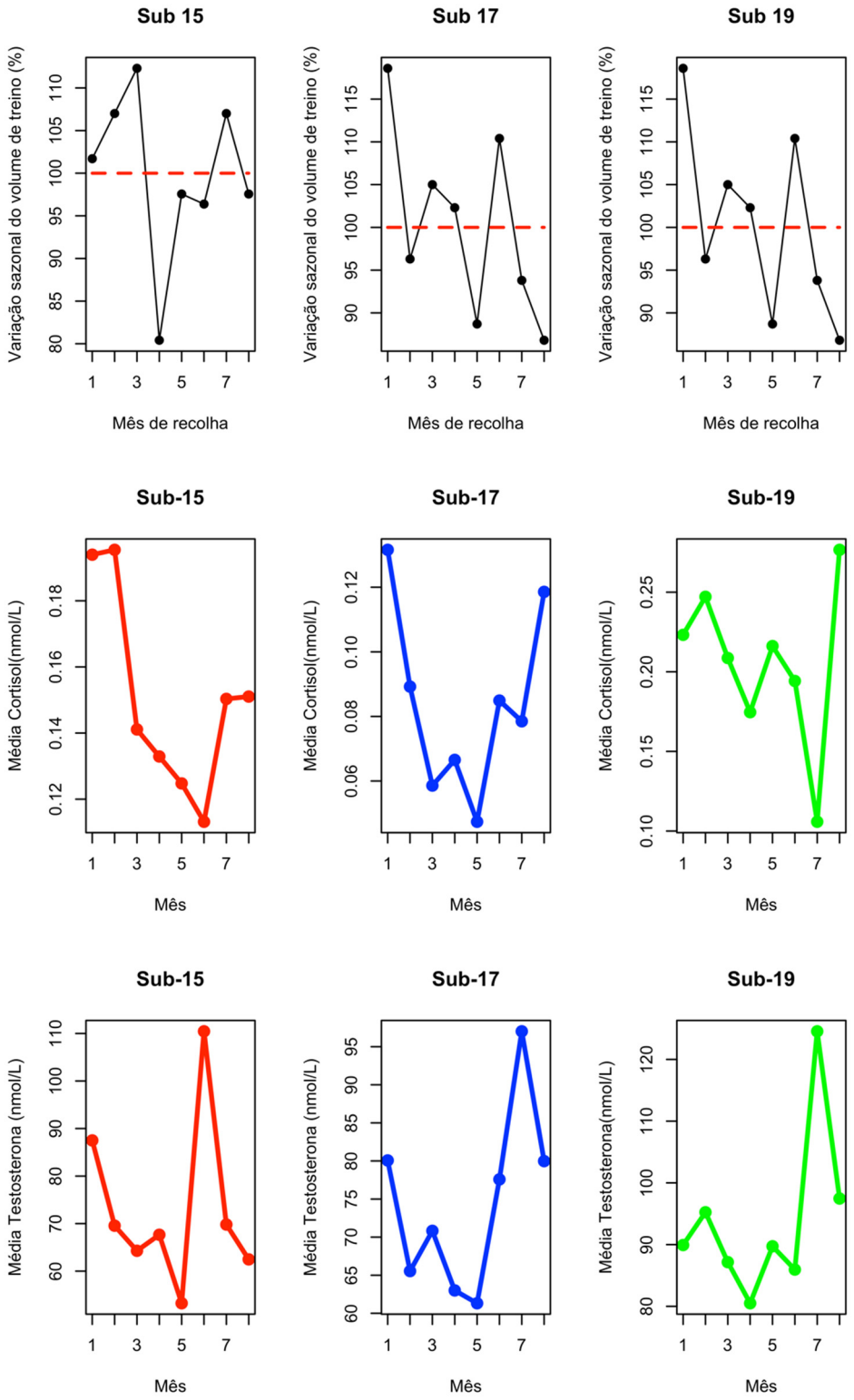

Figura 5.21. Variação sazonal do volume de treino e perfis do comportamento médio do sC e sT referentes a cada escalão, sub-15, sub-17, sub-19.

Analisando de forma individual (Figura. 5.21), no escalão sub-15, na passagem de outubro (3) para novembro (4) ocorre um decréscimo acentuado do volume de treino e ocorre também uma ligeira reação da sT. Tanto no escalão sub-17 como no sub-19 verifica-se que a diminuição do volume de treino em fevereiro (7) que acarreta um comportamento oposto entre o sC e a sT, ou seja, o teor sC diminui e a sT tem o seu pico máximo nesse mês. Este comportamento é esperado e está em concordância com a variação da razão entre dos processos anabólicos representados pela sT e catabólicos associados ao sC, e que são estimulados entre outros fatores pelo exercício e o treino. [74]

O presente estudo permite concluir, que por vezes uma análise global do conjunto de dados pode levar a conclusões menos sustentáveis. No entanto, conjugação dos métodos PCA e HCA levaram a identificar a variável descriminante como sendo o teor sC, enquanto as restantes não forneceram o mesmo resultado. Na verdade, as variáveis biológicas estão associadas a processos diferentes, diferenciando-se na amplitude e temporalmente na resposta ao estímulo, seja ele o stress induzido pelo treino, seja o desafio imunitário.

```
Total_cc=read.csv2("Totalcc.csv", header = TRUE, sep=";", row.names = 1)
Total_ii=read.csv2("Totalii.csv", header = TRUE, sep=";", row.names = 1)
Total_tt=read.csv2("Totaltt.csv", header = TRUE, sep=";", row.names = 1)

PCA<-PCA(Total_cc, graph=FALSE, scale.unit =TRUE)

group <- c(rep("sub19", times=18), rep("sub17", times=22),rep("sub15", times=17))

fviz_pca_biplot(PCA,
                #individuals
                col.ind=group,repel = TRUE,
                pointsize=2, geom.ind="text" ,pointshape=20,
                col.var="dimgray",
                palette=c("maroon2","royalblue1","green"),
                addEllipses=FALSE,habillage = "none",invisible="quali",
                legend.title="Genero",
                xlab="PC1 (44.3%)",ylab="PC2 (9.5%)")


Total_cc_s=scale(Total_cc)
matriz_dis=dist(as.matrix(Total_cc_s))
HCA=hclust(matriz_dis, method = "ward.D2")
plot(HCA,hang = -1, cex = 0.6)
aa=rect.hclust(HCA,k=3,border = 2:3)
grp<-cutree(HCA, 3)

library(dendextend)
dend <- Total_cc %>% # data
       scale %>% # Scale the data
       dist %>% # calculate a distance matrix,
       hclust(method = "ward.D2") %>% # Hierarchical
       as.dendrogram # Turn the object into a dendrog
plot(dend)

dend %>% set("branches_k_color", k=3, value = c("dimgray","cyan4","darkorange2","magenta")
            ,border = 6, lty = 5, lwd = 2) %>%
       set("branches_lwd", 1.2) %>%
       set("labels_cex", c(.9,.9)) %>%
       plot(main = "Customized colors", ylab = "Distância")

dev.print(tiff, "Dendrograma.tiff",
          compression="lzw", res=600, height=6, width=12,
          units="in")
```

Capítulo VI
Conclusão

Capítulo VI

Conclusão

Neste trabalho foi proposto uma metodologia capaz de interligar uma área da química, a quimiometria, com as Ciências do Desporto. Esta aplicação demonstrou que a quimiometria contém várias ferramentas competentes e habilitadas para fornecer um tratamento de dados simples, visual e eficaz.

De uma forma especifica e detalhada, foram tratados e discutidos três conjuntos de dados diferentes relacionados com as Ciências do Desporto, incluindo características cineantropométricas de nadadores, marcadores biológicos salivares de jovens de nadadores durante uma época desportiva e marcadores biológicos salivares de jovens atletas de futebol.

Os resultados alcançados nos diferentes tratamentos de dados, mostram que os métodos quimiométricos são bastante versáteis na resolução de diversos problemas.

Neste trabalho em concreto, no conjunto de dados que relaciona as características cineantropométricas com a avaliação funcional geral e específica demonstrou-se de um modo geral que a capacidade aeróbica e características cineantropométricas são determinantes no desempenho de jovens atletas, quer em provas de longa ou curta distância. Ainda neste conjunto de dados, o PCA fornece uma distinção clara entre atletas do sexo feminino e atletas do sexo masculino.

Relativamente ao segundo conjunto de dados a ser tratado através destes métodos foi possível verificar que pode existir um desfasamento entre a carga de treino do atleta e a resposta à mesma em termos do teor de cortisol salivar. Neste caso, não foram encontradas relações entre as ITRS e o teor de IgA salivar, podendo haver vários fatores externos que confundam a análise.

Por fim, no tratamento dos dados referente a jovens atletas de futebol, verificou-se a capacidade dos métodos para identificar características relevantes a partir da informação completa. Os métodos PCA e HCA tiveram um poder preponderante, mostrando diferenças de comportamento nos marcadores biológicos analisados como, (cortisol salivar, testosterona salivar e IgA salivar) relativamente aos perfis determinados para diversos escalões etários.

Desta maneira, os métodos clássicos da quimiometria como sejam o PCA e HCA revelaram-se muito eficazes em termos da dicotomia representação global/identificação de características específicas relevantes, para estas aplicações em Ciências do Desporto.

Finalmente, as ferramentas proporcionadas pela linguagem R mostraram-se muito adequadas para o desenvolvimento de programas cobrindo todas as fases da análise, incluindo a representação gráfica.

Bibliografia

Bibliografia

[1] Cova, T. F. G. (2011). *Aplicações em quimiometria. Do diagnóstico médico à cientometria.*

[2] Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control, 6*(6), 329-348.

[3] Singh, I., Juneja, P., Kaur, B., & Kumar, P. (2013). Pharmaceutical applications of chemometric techniques. *ISRN Analytical Chemistry, 2013*.

[4] HÉBERGER, K. (2008). Chemoinformatics—multivariate mathematical–statistical methods for data evaluation *Medical Applications of Mass Spectrometry* (pp. 141-169): Elsevier.

[5] Efenberger-Szmechtyk, M., Nowak, A., & Kregiel, D. (2018). Implementation of chemometrics in quality evaluation of food and beverages. *Critical reviews in food science and nutrition, 58*(10), 1747-1766.

[6] Siqueira, L. F., & Lima, K. M. (2016). MIR-biospectroscopy coupled with chemometrics in cancer studies. *Analyst, 141*(16), 4833-4847.

[7] Yustres, I., del Cerro, J. S., González-Mohíno, F., Peyrebrune, M., & González-Ravé, J. M. (2019). Analysis of World Championship swimmers using a performance progression model. Frontiers in Psychology, 10.

[8] Silva, B., Clemente, F. M., & Martins, F. M. (2018). Associations between functional movement screen scores and performance variables in surf athletes. The Journal of sports medicine and physical fitness, 58(5), 583-590.

[9] Rama, L. M. P. L. (2009). Variação de parâmetros fisiológicos, bioquímicos, hormonais e imunitários em nadadores e remadores numa época desportiva.

[10] Morais, J. E., Silva, A. J., Marinho, D. A., Seifert, L., & Barbosa, T. M. (2015). Cluster stability as a new method to assess changes in performance and its determinant factors over a season in young swimmers. International journal of sports physiology and performance, 10(2), 261-268.

[11] Nasirzade, A., Sadeghi, A., Sobhkhiz, A., Mohammadian, K., Nikouei, A., Baghaiyan, M., & Fattahi, A. (2015). Multivariate analysis of 200-m front crawl swimming performance in young male swimmers. Acta of Bioengineering and Biomechanics, 17(3).

[12] Costill, D. L., & Wilmore, J. H. (2001). Fisiologia do esporte e do exercício. São Paulo, ed. Manole, c, 2, 28-51.

[13] Rama, L., Alves, F., & Teixeira, A. (2010). Hormonal, Immune, Autonomic and Mood State Variation in the Initial Preparation Phase of a Winter Season. Portuguese Male Swimmers.

[14] Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A., & Updegraff, J. A. (2000). Biobehavioral responses to stress in females: tend-and-befriend, not fight-or-flight. *Psychological review, 107*(3), 411.

[15] CHATARD, J.-C. (2003). Suivi Scientifique et Biologique des Athlètes de Haut Niveau. *Lutter contre le dopage en gérant la récupération physique*, 119.

[16] Urhausen, A., Gabriel, H., & Kindermann, W. (1995). Blood hormones as markers of training stress and overtraining. Sports medicine, 20(4), 251-276.

[17] Maso, F., Lac, G., Filaire, E., Michaux, O., & Robert, A. (2004). Salivary testosterone and cortisol in rugby players: correlation with psychological overtraining items. British journal of sports medicine, 38(3), 260-263.

[18] Pedersen, B. K., & Hoffman-Goetz, L. (2000). Exercise and the immune system: regulation, integration, and adaptation. Physiological reviews, 80(3), 1055-1081.

[19] Ronsen, O. (2003). Immune, endocrine and metabolic changes related to exhaustive and repeated exercise sessions. Oslo: Unipub AS.

[20] Brenner, I., Shek, P., Zamecnik, J., & Shephard, R. (1998). Stress hormones and the immunological responses to heat and exercise. International journal of sports medicine, 19(02), 130-143.

[21] Nieman, D. C., Johanssen, L. M., Lee, J. W., & Arabatzis, K. (1990). Infectious episodes in runners before and after the Los Angeles Marathon. J Sports Med Phys Fitness, 30(3), 316-328.

[22] Gleeson, M. (2004). Immune function and exercise. European Journal of Sport Science, 4(3), 52-66.

[23] Reid, M., Drummond, P., & Mackinnon, L. (2001). The effect of moderate aerobic exercise and relaxation on secretory immunoglobulin A. International journal of sports medicine, 22(02), 132-137.

[24] Nieman, D. C. (1998). Exercise and resistance to infection. Canadian journal of physiology and pharmacology, 76(5), 573-580.

[25] Pyne, D. B., Mcdonald, W. A., Gleeson, M., Flanagan, A., Clancy, R. L., & Fricker, P. A. (2001). Mucosal immunity, respiratory illness, and competitive performance in elite swimmers. Medicine & Science in Sports & Exercise, 33(3), 348-353.

[26] Estrutura, tipos e funções de imunoglobulinas. (2020). Retrieved 21 March 2020, from https://pt.thpanorama.com/blog/ciencia/inmunoglobulinas-estructura-tipos-y-funciones.html.

[27] Callister, R., & Gleeson, M. (2007). The relevance of salivary IgA for the immunological management of athletes. Paper presented at the Conferences in Exercise Immunology. Coimbra

[28] Nieman, D. C. (2000). Exercise effects on systemic immunity. Immunology and Cell Biology, 78(5), 496-501.

[29] Zhao, K., Hohmann, A. A., Chang, Y., Zhang, B., Pion, J., & Gao, B. (2019). Physiological, anthropometric, and motor characteristics of elite Chinese youth athletes from six different sports. Frontiers in physiology, 10, 405

[30] Gastin, P. B. (2001). Energy system interaction and relative contribution during maximal exercise. *Sports Med, 31*(10), 725-741.

[31] Barbanti, V. J., Tricoli, V., & Ugrinowitsch, C. (2004). Relevância do conhecimento científico na prática do treinamento físico. Revista Paulista de Educação Física, 18(8), 101-109.

[32] Hopkins, W. G. (1991). Quantification of training in competitive sports. Sports medicine, 12(3), 161-183.

[33] Valdivieso, F. N., & Gaia, A. O. (2010). Planificación del entrenamiento y su control: Real Federación Española de Natación, Escuela Nacional de Entrenadores.

[34] Kowalski, B., & Bender, C. (1972). Pattern recognition. Powerful approach to interpreting chemical data. *Journal of the American Chemical Society, 94*(16), 5632-5639.

[35] Kowalski, B., & Bender, C. (1973). Pattern recognition. II. Linear and nonlinear methods for displaying chemical data. *Journal of the American Chemical Society*, 95(3), 686-693.

[36] Eriksson, L., Trygg, J., & Wold, S. (2014). A chemometrics toolbox based on projections and latent variables. *Journal of Chemometrics*, 28(5), 332-346.

[37] Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12), 1540-1542.

[38] Almeida, J., Barbosa, L., Pais, A., &. Formosinho, S. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 87(2), 208-217.

[39] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), 236-244.

[40] Drab, K., & Daszykowski, M. (2014). Clustering in analytical chemistry. *Journal of AOAC International*, 97(1), 29-38.

[41] Davies, A., & Fearn, T. (2004). Back to basics: the principles of principal component analysis. *Spectroscopy Europe*, 16(6), 20-23.

[42] Jolliffe, I. T. (2002). *Principal component analysis*. (2ª ed.). New York: Springer.

[43] rproject.org. (2020). Retrieved 15 May 2020, from https://cran.rproject.org/doc/contrib/OlszewskiR_for_biostatisticians.

[44] R Packages. (2020). Retrieved 16 April 2020, from https://rstudio.com/products/rpackages.

[45] Kassambara, A. (2017). *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra* (Vol. 2): STHDA.

[46] Marfell-Jones, M., Stewart, A., & Olds, T. (2006). *Kinanthropometry IX: Proceedings of the 9th International Conference of the International Society for the Advancement of Kinanthropometry*: Routledge.

[47] Carter, J., & Ackland, T. (1994). Kinanthrpometry in Aquatic Sports. *Champaign, IL: Human Kinetics*.

[48] Carzola, G. (1993). Tests Spécifiques d'évaluation du Nager. *Paris: Editions Vigot*.

[49] OLBRECHT, J. (2000). The Science of Winning: Planning, Periodizing and Optimizing: Swim.

[50] Costill, D., Kovaleski, J., Porter, D., Kirwan, J., Fielding, R., & King, D. (1985). Energy expenditure during front crawl swimming: predicting success in middle-distance events. *International journal of sports medicine, 6*(05), 266-270.

[51] Estrutura, tipos e funções de imunoglobulinas. (2020). Retrieved 21 March 2020, from https://pt.thpanorama.com/blog/ciencia/inmunoglobulinas-estructura-tipos-y-funciones.html

[52] Saunders, A., Feldman, H., Correia, C., & Weinstein, D. (2005). Clinical evaluation of a portable lactate meter in type I glycogen storage disease. *Journal of inherited metabolic disease, 28*(5), 695-701.

[53] Li, T., & Gleeson, M. (2004). The effect of single and repeated bouts of prolonged cycling and circadian variation on saliva flow rate, immunoglobulin a and-amylase responses. *J. Sports Sci, 22*(11-12), 1015-1024.

[54] Sanders, R. H. (2013). How do asymmetries affect swimming performance? *Journal of Swimming Research, 21*(1).

[55] Willems, T. M., Cornelis, J. A., De Deurwaerder, L. E., Roelandt, F., & De Mits, S. (2014). The effect of ankle muscle strength and flexibility on dolphin kick performance in competitive swimmers. *Human Movement Science, 36*, 167-176.

[56] McCullough, A. S., Kraemer, W. J., Volek, J. S., Solomon-Hill Jr, G. F., Hatfield, D. L., Vingren, J. L., . . . Häkkinen, K. (2009). Factors affecting flutter kicking speed in women who are competitive and recreational swimmers. *The Journal of Strength & Conditioning Research, 23*(7), 2130-2136.

[57] Hull, M. (1990). Flexible ankles: Faster swimming. *Swimming Technique, 27*(3), 23-24.

[58] Maglischo, E. W. (2003). Swimming fastest: *Human kinetics*.

[59] Watkins, J. (1983). The effects of leg action on performance in the sprint front crawl stroke. *Biomechanics and medicine in swimming*, 310-314.

[60] Demirkan, E., Can, S., Ozkadi, T., & Alagoz, I. (2019). Fifty-meter swimming performance in young swimmers: the effect of anthropometric and motoric values on performance. *Medicina Dello Sport, 72*(4), 488-497.

[61] Baxter-Jones, A. D., & Helms, P. J. (1996). Effects of training at a young age: a review of the training of young athletes (TOYA) study. *Pediatric Exercise Science, 8*(4), 310-327.

[62] Theintz, G. E., Howald, H., Weiss, U., & Sizonenko, P. (1993). Evidence for a reduction of growth potential in adolescent female gymnasts. *The Journal of pediatrics, 122*(2), 306-313.

[63] Hebbelinck, M., Carter, L., & De Garay, A. (1975). Body build and somatotype of Olympic swimmers, divers and water polo players. *Swimming II*, 285-305.

[64] Prestes, J., Leite, R. D., Leite, G. d. S., Donatto, F. F., Urtado, C. B., Bartolomeu Neto, J., & Dourado, A. C. (2006). Características antropométricas de jovens nadadores brasileiros do sexo masculino e feminino em diferentes categorias competitivas. *Rev Bras Cineantropom Desempenho Hum, 8*(4), 25-31.

[65] Aubets, J., & Segura, J. (1995). Salivary cortisol as a marker of competition related stress. *Science & sports, 10*(3), 149-154.

[66] Reid, V., Gleeson, M., Williams, N., & Clancy, R. (2004). Clinical investigation of athletes with persistent fatigue and/or recurrent infections. *British journal of sports medicine, 38*(1), 42-45.

[67] Nieman, D. C., & Pedersen, B. K. (1999). Exercise and immune function. *Sports medicine, 27*(2), 73-80.

[68] Gleeson, M., McDonald, W., Pyne, D., Clancy, R., Cripps, A., Francis, J., & Fricker, P. (2000). Immune status and respiratory illness for elite swimmers during a 12-week training cycle. *International journal of sports medicine, 21*(04), 302-307.

[69] Peters, E. (2004). Postrace upper respiratory tract 'infections' in ultramarathoners—infection, allergy or inflammation? *South African Journal of Sports Medicine, 16*(1), 3-9.

[70] Klentrou, P., Giannopoulou, A., McKinlay, B. J., Wallace, P., Muir, C., Falk, B., & Mack, D. (2016). Salivary cortisol and testosterone responses to resistance and plyometric exercise in 12-to 14-year-old boys. *Applied physiology, nutrition, and metabolism, 41*(7), 714-718.

[71] Mirwald, R. L., Baxter-Jones, A. D., Bailey, D. A., & BEUNEN, G. P. (2002). An assessment of maturity from anthropometric measurements. *Medicine & Science in Sports & Exercise, 34*(4), 689-694.

[72] Moore, S., Moore, M., Klentrou, P., Sullivan, P., & Falk, B. (2010). Maturity status in male child and adolescent athletes. *The Journal of sports medicine and physical fitness, 50*(4), 486-493.

[73] Pearson, D., Naughton, G. A., & Torode, M. (2006). Predictability of physiological testing and the role of maturation in talent identification for adolescent team sports. *Journal of science and medicine in sport, 9*(4), 277-287.

[74] Urhausen, A., Gabriel, H., & Kindermann, W. (1995). Blood hormones as markers of training stress and overtraining. *Sports medicine, 20*(4), 251-276.

Anexo

```r
Total_cc=read.csv2("Totalcc.csv", header = TRUE, sep=";", row.names = 1)
Total_ii=read.csv2("Totalii.csv", header = TRUE, sep=";", row.names = 1)
Total_tt=read.csv2("Totaltt.csv", header = TRUE, sep=";", row.names = 1)

PCA<-PCA(Total_cc, graph=FALSE, scale.unit =TRUE)

group <- c(rep("sub19", times=18), rep("sub17", times=22),rep("sub15", times=17))

fviz_pca_biplot(PCA,
                #individuals
                col.ind=group,repel = TRUE,
                pointsize=2, geom.ind="text" ,pointshape=20,
                col.var="dimgray",
                palette=c("maroon2","royalblue1","green"),
                addEllipses=FALSE,habillage = "none",invisible="quali",
                legend.title="Genero",
                xlab="PC1 (44.3%)",ylab="PC2 (9.5%)")


Total_cc_s=scale(Total_cc)
matriz_dis=dist(as.matrix(Total_cc_s))
HCA=hclust(matriz_dis, method = "ward.D2")
plot(HCA,hang = -1, cex = 0.6)
aa=rect.hclust(HCA,k=3,border = 2:3)
grp<-cutree(HCA, 3)

library(dendextend)
dend <- Total_cc %>% # data
        scale %>% # Scale the data
        dist %>% # calculate a distance matrix,
        hclust(method = "ward.D2") %>% # Hierarchical clust
        as.dendrogram # Turn the object into a dendrogram.
plot(dend)

dend %>% set("branches_k_color", k=3, value = c("dimgray","cyan4","darkorange2","magenta")
             ,border = 6, lty = 5, lwd = 2) %>%
        set("branches_lwd", 1.2) %>%
        set("labels_cex", c(.9,.9)) %>%
        plot(main = "Customized colors", ylab = "Distancia")

dev.print(tiff, "Dendrograma.tiff",
          compression="lzw", res=600, height=6, width=12,
          units="in")
```

Anexo

Nesta secção é apresentado um exemplo de script efetuado em RStudio para um conjunto de dados. É importante referir que todas as bases de dados foram sujeitas a um tratamento prévio antes do tratamento apresentado a seguir.

```
rm(list=ls())
setwd("~/Desktop/Mestrado/2º Ano/R_Caracteristicas_Cineantropometricas")

y=read.csv2("Analise_Fem.csv",header = TRUE,sep = ";", row.names = 1)
x=read.csv2("Analise_Masc.csv",header = TRUE,sep = ";", row.names = 1)

Total=rbind(x,y)
#Nome das variáveis
names(Total)=c("Id","MS","Est","AS",
               "Env","IMD", "CM","LM",
               "CP","LP", "CMS", "CMI",
               "DBA","DBC", "DT", "FlexPl",
               "FlexOm", "ExtPl", "ExtOm","FlexTr", "ExtTr",
               "Desl", "VelT30","INT30", "PrMD",
               "PrME","SK")

Total_s=scale(Total_1) #normalização

PCA<-PCA(Total_1, graph=FALSE, scale.unit =TRUE)

######### Valores Próprios ############
valores_proprios<- get_eigenvalue(PCA)

valores_proprios<-data.frame(valores_proprios)

fviz_eig(PCA, addlabels = TRUE, barcolor = "darkorange2",barfill = "darkorange2",
         xlab="Compomentes Principais",ylab = "Percentagem da Variância",
         ggtheme = theme_classic(),main = "")
dev.print(tiff, "ScreePlot.tiff",
          compression="lzw", res=800, height=4.5, width=6.2,
          units="in") #guardar gráfico

########### Variáveis ###############
var <- get_pca_var(PCA)
fviz_pca_var(PCA, col.var = "black", repel = TRUE ,
             ggtheme = theme_bw(),
             xlab="PC1 (44.3%)",ylab="PC2 (9.5%)")
dev.print(tiff, "Var1.tiff",
          compression="lzw", res=600, height=5, width=6,
          units="in")

pvlibrary("corrplot")
corrplot(t(var1), is.corr=FALSE, method = c("number"), hclust.method = "ward.D2",tl.col = "black")
corrplot(t(var2), is.corr=FALSE, method = c("number"),hclust.method = "ward",tl.col = "black")

```

```
########### PCA só com objetos ########
fviz_pca_ind(PCA,
             #Individuals
             geom.ind = "point",
             pointshape=20 , pointsize=3,
             col.ind = "darkorange2",
             col.var="red",
             addEllipses=,habillage = "none",invisible="quali",
             repel = TRUE,legend.title="Groups",
             xlab="PC1 (44.3%)",ylab="PC2 (9.5%)",
             ggtheme = theme_bw())
############ Dendrograma ###########3
library(dendextend)
dend <- Total_1 %>% # data
  scale %>% # Scale the data
  dist %>% # calculate a distance matrix,
  hclust(method = "ward.D2") %>% # Hierarchical clustering
  as.dendrogram # Turn the object into a dendrogram.
plot(dend)

dend %>% set("branches_k_color", k=4, value = c("blue","cyan4","darkorange2","magenta")
             ,border = 6, lty = 5, lwd = 2) %>%
  set("branches_lwd", 1.2) %>%
  set("labels_cex", c(.9,.9)) %>%
  plot(main = "Customized colors", ylab = "Distância")

grp=as.character(cutree(HCA,4))
G1_hcpc<-apply(Total_3[which(grp==1),],2,mean)
G2_hcpc<-apply(Total_3[which(grp==2),],2,mean)
G3_hcpc<-apply(Total_3[which(grp==3),],2,mean)
G4_hcpc<-apply(Total_3[which(grp==4),],2,mean)
G1_hcpc1<-apply(Total_3[which(grp==1),],2,sd)
G2_hcpc1<-apply(Total_3[which(grp==2),],2,sd)
G3_hcpc1<-apply(Total_3[which(grp==3),],2,sd)
G4_hcpc1<-apply(Total_3[which(grp==4),],2,sd)
dim(Total_3[which(grp==1),])
dim(Total_3[which(grp==2),])
dim(Total_3[which(grp==3),])
dim(Total_3[which(grp==4),])

mean_grp_hcpc=data.frame(cbind(G1_hcpc,G2_hcpc,G3_hcpc,G4_hcpc))
write.csv2(mean_grp_hcpc,"Medias_Grp_Final.csv")

sd_grp_hcpc=data.frame(cbind(G1_hcpc1,G2_hcpc1,G3_hcpc1,G4_hcpc1))
write.csv2(sd_grp_hcpc,"Sd_Grp_Final.csv")
############ Separação por género ############
group <- c(rep("Masculino", times=85), rep("Feminino", times=20))

fviz_pca_biplot(PCA,
                #Individuals
                col.ind=group,repel = TRUE,
                pointsize=2, geom.ind="point" ,pointshape=20,
                col.var="dimgray",
                palette=c("maroon2","royalblue1"),
                addEllipses=TRUE,habillage = "none",invisible="quali",
                legend.title="Género",ellipse.type="convex",
                xlab="PC1 (44.3%)",ylab="PC2 (9.5%)")
```