



Letícia Cruz
Costa Leite

Técnicas exploratórias na deteção de *outliers* em
dados composicionais



Letícia Cruz
Costa Leite

Técnicas exploratórias na deteção de *outliers* em dados composicionais

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Dedico plenamente este trabalho aos meus Pais, Jorge Manuel Leite e Maria Clarinda Leite, e ao meu querido Irmão, Jorge Filipe Leite, pelo apoio, paciência e incentivo que sempre manifestaram para que a concretização de uma das grandes etapas do meu percurso de vida fosse alcançada.

“Quem desiste nunca ganha e um vencedor nunca desiste.”

Autor desconhecido

o júri

presidente

Doutor Eugénio Alexandre Miguel Rocha

Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

vogais

Doutora Sónia Manuela Mendes Dias

Professora Adjunta da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viana do Castelo

Doutora Adelaide de Fátima Baptista Valente Freitas

Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

agradecimentos

O trabalho, a dedicação e o esforço que deposito em todas as tarefas que me são confiadas demonstram a pessoa que sou e fortalecem os resultados adquiridos. Durante o meu percurso académico, a coragem, a persistência e o apoio de todos os meus familiares, amigos e colegas tornaram possível a conquista de mais uma etapa com mérito próprio.

Um agradecimento reconhecido aos meus pais que sempre me apoiaram em tudo o que fosse necessário para garantir o meu bem-estar. Um agradecimento incondicional ao meu querido irmão, alguém que admiro desde sempre, pela forma como incessantemente me apoiou manifestando o seu profundo reconhecimento pelo meu trabalho.

Um notável agradecimento à minha orientadora, Professora Adelaide Freitas, pela prontidão, disponibilidade e flexibilidade demonstrada a fim de esclarecer todas as minhas dúvidas durante a elaboração desta dissertação. Agradeço-lhe também a confiança depositada em mim, a motivação e a sua boa disposição que sempre me animaram em dias menos bons. Alguém por quem tenho uma grata amizade e muito respeito.

Um especial agradecimento à Professora Cristina Gomes pela iniciativa em facultar os dados demográficos para a componente prática desta dissertação, o auxílio e a disponibilidade em clarificar as minhas questões na interpretação dos dados.

Um enorme agradecimento a todos os meus Professores, em especial à Professora Elsa Pomares, pela sabedoria e valores transmitidos contribuindo, indubitavelmente, para o meu crescimento pessoal.

Agradeço a seriedade na apreciação de todo o trabalho desenvolvido na presente dissertação. A todos os que genuinamente me apoiaram e aos que sempre me dirigiram uma palavra amiga de incentivo e força, um Bem-Haja a todos vós!

palavras-chave

dados composicionais, transformações log-razões, observações atípicas, distância de mahalanobis robusta, estimador MCD, biplot composicional robusto.

Resumo

Dados multivariados que representam descrições quantitativas positivas das partes de um todo como, por exemplo proporções, transmitindo informação relativa em vez de absoluta, são designados por dados composicionais e são o alvo fulcral de estudo da presente dissertação. Em particular, estudam-se e aplicam-se técnicas estatísticas numéricas, baseadas em transformações de log-razões, e técnicas estatísticas gráficas sobre os dados transformados na deteção de composições (observações) atípicas (*outliers*) às quais correspondem as observações multivariadas que, por algum motivo, diferem das restantes composições do conjunto de dados.

Os métodos estatísticos multivariados clássicos tendem a ignorar os *outliers*, tomando-os como observações “normais” e potenciando assim o enviesamento de resultados. Técnicas estatísticas robustas, que reduzem a influência de *outliers*, são de extrema importância para uma correta análise e interpretação dos dados.

Um dos métodos estatísticos mais usuais na identificação de observações multivariadas atípicas baseia-se na distância de Mahalanobis calculada com estimativas robustas da média e da matriz de covariância populacionais obtidas através do estimador MCD (*Minimum Covariance Determinant*). Gráficamente, o método biplot é uma ferramenta exploratória amplamente utilizada na visualização de observações multivariadas e, consequentemente, de *outliers*. Considerando o caso especial de dados composicionais, um dos propósitos do presente trabalho reside também em estudar propriedades da distância de Mahalanobis robusta e biplots robustos sobre este tipo de dados na deteção de composições *outliers*.

Como aplicação destas metodologias estatísticas exploram-se, sob o ponto de vista relativo (isto é, composicional), três conjuntos de dados demográficos, extraídos dos Censos de 2011, baseados na migração interna em Portugal. Esses conjuntos dizem respeito a todos os 308 municípios e, para cada município, têm-se contagens de residentes que afirmaram que no período de 2005 a 2011 mudaram de residência passando a habitar no município em causa. A contagem dos residentes que mudaram de município tem em conta o grupo etário, a habilitação académica e a situação profissional.

A análise estatística realizada conduziu à identificação de grupos distintos de municípios *outliers* entre os três conjuntos de dados. Relativamente à situação profissional as conclusões foram mais interpretáveis. Tendo em conta a distribuição do grupo etário, da habilitação académica e da situação profissional, este estudo denuncia a existência de municípios atípicos por serem mais ou menos atrativos. Usando cartogramas constata-se que muitos destes municípios *outliers* localizam-se em regiões do interior de Portugal Continental.

keywords

compositional data, logratio transformations, outliers, robust mahalanobis distance, MCD estimator, robust compositional biplot.

Abstract

Multivariate data of positive values which describe parts of a whole such as proportions, conveying relative rather than absolute information, are referred to as compositional data. This type of data is the main subject of the study of this dissertation. Numerical statistical techniques, based on log-ratios transformations, and graphical statistical techniques on transformed data in the detection of atypical compositions (multivariate observations outliers) are discussed. Outliers are observations that, for some reason, differ from the other observations belonging to the data set.

Classic multivariate statistical methods tend to ignore outliers which are taking as “normal” observations and can produce results biased. Hence, robust statistical techniques, which reduce the influence of outliers, are of extreme importance for proper analysis and interpretation of the data.

One of the most popular statistical methods for identifying outliers is based on the Mahalanobis distance calculated using robust estimates of the mean and covariance matrix obtained by the MCD (Minimum Covariance Determinant) estimator. On another hand, graphically, the biplot method is an exploratory tool widely used in the visualization of multivariate data and, consequently, outliers. Considering the special case of compositional data, properties of Mahalanobis robust distance and robust biplots on this type of data in the detection of outlier are also studied of this dissertation.

The applications of these statistical methodologies on three demographic data sets, extracted from the 2011 Census and based on internal migration in Portugal, are explored from a relative point of view (i.e., compositional). These data sets concern the total set of 308 municipalities of Portugal. For each municipality, there are counts of residents who stated that in the period from 2005 to 2011 they changed their residence and began to live in the municipality in question.

The count of the residents who changed the municipality considers the age group, the academic qualification and the occupational status. The statistical analysis performed led to the identification of distinct groups of outlier's municipalities among the three datasets. Concerning the occupational status the conclusions were more interpretable. Considering the age distribution, academic qualification and occupational status, this study denounces the existence of atypical municipalities because they are more or less attractive. Using cartograms, it is found that many of these outlier's municipalities are in regions of the interior of Portugal.

Conteúdo

Conteúdo	i
Lista de Figuras	iii
Lista de Tabelas	vii
Abreviaturas	ix
1 Introdução	1
1.1 Noção de dados composicionais	1
1.2 Noção de observações atípicas	3
1.3 Motivação para o tema	4
1.4 Objetivos e organização da dissertação	6
2 Metodologias Numéricas	9
2.1 Conceitos básicos de dados composicionais	9
2.2 Geometria de Aitchison no simplex	11
2.3 Princípios de uma análise composicional	14
2.4 Transformações de dados composicionais	15
2.4.1 Transformação alr	17
2.4.2 Transformação clr	18
2.4.3 Transformação ilr e Coordenadas Pivô	21
2.5 Base ortonormal no simplex	27
2.5.1 Generalização das coordenadas pivô	28
2.5.2 Partição Binária Sequencial	31
2.5.3 <i>Balances</i>	34
2.6 Detecção de <i>outliers</i> multivariados	35
2.6.1 Métodos para detecção de <i>outliers</i> multivariados	36
2.6.2 Estimador MCD	40
2.7 Propriedades das transformações na detecção de observações atípicas	41
2.8 Componentes irregulares	48
2.8.1 Zeros de contagem	49
2.8.2 Algoritmo k -NN para imputação	50
3 Metodologias Gráficas	53
3.1 Estatística descritiva de dados composicionais	53
3.2 Representação gráfica de dados composicionais	56

3.2.1	Diagrama ternário	56
3.2.2	Biplot	60
3.2.3	Biplot composicional	65
3.2.4	Biplot composicional robusto	69
4	Aplicação em dados demográficos	71
4.1	Matrizes de dados	71
4.2	Metodologias de análise dos dados	72
4.3	Análise e discussão dos resultados	75
4.3.1	Conjunto de dados por grupo etário	75
4.3.2	Conjunto de dados por habilitação acadêmica	83
4.3.3	Conjunto de dados por situação profissional	92
5	Conclusão e trabalho futuro	101
	Referências Bibliográficas	105
A	<i>Outliers: função mvoutlier.CoDa()</i>	109
A.1	Conjunto de dados por grupo etário	109
A.2	Conjunto de dados por habilitação acadêmica	110
A.3	Conjunto de dados por situação profissional	110
B	Poster	111

Lista de Figuras

1.1	(a) Representação de uma observação que se desvia visivelmente de um conjunto de dados, tratando-se de um dado atípico. (b) Representação de três <i>outliers</i> . (c) Representação de dois conjuntos de dados, A e B, onde a observação verde é atípica para o conjunto A mas pertence ao conjunto B.	3
1.2	Esquema representativo das partes de cada um dos três conjuntos de dados em estudo.	6
2.1	Representação das quatro elipses referentes aos quantis de ordem 0.25, 0.5 e 0.75 e ao valor de corte para detetar <i>outliers</i> , quantil de ordem 0.975 (Figura extraída de [33]).	39
3.1	Representação do simplex em \mathbb{R}^3	56
3.2	Diagrama ternário (adaptado de [2]).	57
3.3	Representação de um diagrama ternário com coordenadas cartesianas.	57
3.4	Representação em diagramas ternários dos padrões (i), (ii) e (iii), respetivamente.	58
3.5	Representação em diagramas ternários dos padrões (iv), (v) e (vi), respetivamente.	59
3.6	Diagrama ternário com os dados composicionais (a) iniciais (b) após o método de centralização.	60
3.7	Ilustração de um biplot de uma matriz de dados \mathbf{X} genérica com a seguinte representação dos símbolos: \bullet linhas (observações); \rightarrow colunas (variáveis); $--$ ligação (razão entre variáveis); \bullet origem do conjunto de dados (centro O) (adaptado de [7]).	66
4.1	Comparação gráfica da raiz quadrada da distância de Mahalanobis robusta (gráfico da esquerda) e clássica (gráfico da direita) para o conjunto de dados por grupo etário. As observações identificadas por números são <i>outliers</i>	75
4.2	Gráfico que exhibe os <i>outliers</i> (destacados a vermelho) de acordo com o quantil $\chi^2_{4;0.975}$ para o conjunto de dados por grupo etário.	76
4.3	Gráfico que compara as distâncias de Mahalanobis robusta versus clássica para o conjunto de dados por grupo etário. Os símbolos “+” no 1º quadrante destacam as observações atípicas que se localizam fora do quantil $\chi^2_{4;0.975}$	76
4.4	Biplots composicionais robustos (a) para visualização dos <i>outliers</i> bem como do comprimento dos <i>links</i> e (b) para visualização dos ângulos entre os <i>links</i> para o conjunto de dados por grupo etário.	77
4.5	Gráficos de dispersão univariados para o conjunto de dados por grupo etário.	79
4.6	Cartograma de Portugal (Continental) com 28 municípios <i>outliers</i> a cor verde, para o conjunto de dados por grupo etário.	81

4.7	Cartograma do Arquipélago da Madeira com identificação do município <i>outlier</i> de Porto Moniz a cor verde, para o conjunto de dados por grupo etário.	82
4.8	Cartograma do Arquipélago dos Açores com identificação do município <i>outlier</i> de Lajes das Flores a cor verde, no grupo Ocidental, para o conjunto de dados por grupo etário.	82
4.9	Comparação gráfica da raiz quadrada da distância de Mahalanobis robusta (gráfico da esquerda) e clássica (gráfico da direita) para o conjunto de dados por habilitação académica. As observações identificadas por números são <i>outliers</i> . 83	
4.10	Gráfico que exhibe os <i>outliers</i> (destacados a vermelho) de acordo com o quantil $\chi^2_{9;0.975}$ para o conjunto de dados por habilitação académica.	84
4.11	Gráfico que compara as distâncias de Mahalanobis robusta e clássica para o conjunto de dados por habilitação académica. Os símbolos “+” no 1º quadrante destacam as observações atípicas que se localizam fora do quantil $\chi^2_{9;0.975}$	84
4.12	Biplots composicionais robustos (a) para visualização dos <i>outliers</i> bem como do comprimento dos <i>links</i> e (b) para visualização dos ângulos entre os <i>links</i> para o conjunto de dados por habilitação académica.	85
4.13	Gráficos de dispersão univariados para o conjunto de dados por habilitação académica.	87
4.14	Cartograma de Portugal (Continental) com 54 municípios <i>outliers</i> a cor verde, para o conjunto de dados por habilitação académica.	90
4.15	Cartograma do Arquipélago da Madeira com identificação dos municípios <i>outliers</i> de Porto Moniz e Ponta do Sol a cor verde, para o conjunto de dados por habilitação académica.	91
4.16	Cartograma do Arquipélago dos Açores com identificação a cor verde dos municípios <i>outliers</i> de Corvo e Lajes das Flores no grupo Ocidental, do município Lajes do Pico no grupo Central e dos municípios Nordeste, Povoação e Vila Franca do Campo no grupo Oriental, para o conjunto de dados por habilitação académica.	91
4.17	Comparação gráfica da raiz quadrada da distância de Mahalanobis robusta (gráfico da esquerda) e clássica (gráfico da direita) para o conjunto de dados por situação profissional. As observações identificadas por números são <i>outliers</i> . 92	
4.18	Gráfico que exhibe os <i>outliers</i> (destacados a vermelho) de acordo com o quantil $\chi^2_{2;0.975}$ para o conjunto de dados por situação profissional.	93
4.19	Gráfico que compara as distâncias de Mahalanobis robusta e clássica para o conjunto de dados por situação profissional. Os símbolos “+” no 1º quadrante destacam as observações atípicas que se localizam fora do quantil $\chi^2_{2;0.975}$	93
4.20	Biplots composicionais robustos (a) para visualização dos <i>outliers</i> bem como do comprimento dos <i>links</i> e (b) para visualização dos ângulos entre os <i>links</i> para o conjunto de dados por situação profissional.	94
4.21	Gráficos de dispersão univariados para o conjunto de dados por situação profissional.	96
4.22	Diagrama ternário para o conjunto de dados por situação profissional.	97
4.23	Cartograma de Portugal (Continental) com 9 municípios <i>outliers</i> a cor verde, para o conjunto de dados por situação profissional.	99
4.24	Cartograma do Arquipélago da Madeira com identificação do município <i>outlier</i> de Porto Santo a cor verde, para o conjunto de dados por situação profissional. 100	

LISTA DE FIGURAS

4.25	Cartograma do Arquipélago dos Açores com identificação do município <i>outlier</i> do Corvo a cor verde, no grupo Ocidental, para o conjunto de dados por situação profissional.	100
B.1	Poster apresentado nas XXVI Jornadas de Classificação e Análise de Dados, na Escola Superior de Tecnologia e Gestão de Viseu, em Abril de 2019.	111

Lista de Tabelas

1.1	Número de habitantes do município de Aguiar da Beira agrupados por grupo etário (em anos).	4
1.2	Parte inicial do conjunto de dados por situação profissional em composições absolutas.	6
1.3	Parte inicial do conjunto de dados por situação profissional em composições relativas.	7
2.1	Codificação de uma PBS para construção de uma base ortonormal onde a primeira partição confronta a parte $x_1^{(\ell)}$ com as restantes partes da composição. . .	31
2.2	Processo para obtenção da matriz de contrastes resultante da PBS obtida por partição.	32
2.3	PBS de uma composição de 5 partes.	32
2.4	Processo para obtenção da matriz de contrastes resultante da PBS de uma composição de 5 partes.	33
2.5	PBS de uma composição de 5 partes e equilíbrio entre grupos de partes do município de Anadia.	35
2.6	Proporção de habitantes que se deslocaram em cada município com Doutora-mento.	49
2.7	Alguns municípios que possuem zeros de contagem (NA's) nas diferentes partes das suas composições.	51
2.8	Resultado do algoritmo k -NN para imputação dos zeros de contagem para os diferentes municípios assinalados pela sigla NA na Tabela 2.7.	52
4.1	Tabela representativa da matriz inicial do conjunto de dados do grupo etário. .	71
4.2	Tabela representativa da matriz do conjunto de dados composicional (aqui re-ferenciada como matriz composicional) do grupo etário.	72
4.3	Matriz de variação composicional entre as variáveis dos grupos etários (a negrito destacam-se os valores do maior e do menor <i>link</i>).	78
4.4	Uma submatriz de correlações entre log-razões e partes no grupo etário. . . .	78
4.5	Síntese dos resultados obtidos pelas funções das metodologias numéricas e gráficas. 80	
4.6	Matriz de variação composicional entre as variáveis das habilitações académicas (a negrito destacam-se os valores do maior e do menor <i>link</i>).	86
4.7	Uma submatriz de correlações entre log-razões e partes na habilitação académica. 86	
4.8	Síntese dos resultados obtidos pelas funções das metodologias numéricas e gráficas. 88	
4.9	Matriz de variação composicional entre as variáveis da situação profissional (a negrito destacam-se os valores do menor e maior <i>link</i> , respetivamente).	94

4.10	Uma submatriz de correlações entre log-razões e partes na situação profissional.	95
4.11	Síntese dos resultados obtidos pelas funções das metodologias numéricas e gráficas.	98
A.1	Identificação dos <i>outliers</i> detetados pela função <i>mvoutlier.CoDa()</i> para o conjunto de dados por grupo etário.	109
A.2	Identificação dos <i>outliers</i> detetados pela função <i>mvoutlier.CoDa()</i> para o conjunto de dados por habilitação académica.	110
A.3	Identificação dos <i>outliers</i> detetados pela função <i>mvoutlier.CoDa()</i> para o conjunto de dados por situação profissional.	110

Abreviaturas

ACP Análise de Componentes Principais. 26, 63, 64, 69, 70

D-D Plot Gráfico de Distância versus Distância (*Distance-Distance Plot*). 73, 76, 84, 93

DVS Decomposição em Valores Singulares (*Singular Value Decomposition*). 61, 62, 69, 70

MCD (estimador da) Matriz de Covariâncias de Determinante Mínimo (*Minimum Covariance Determinant (estimator)*). 7, 40, 47, 48, 69, 72–77, 80, 88, 98, 101

PBS Partição Binária Sequencial. 7, 28, 31–35, 72, 75, 83, 92

SQRT Raiz quadrada (*Square Root*). 63

Capítulo 1

Introdução

1.1 Noção de dados composicionais

No decorrer dos tempos, nas mais diversas áreas de investigação, Biologia, Geologia, Química, Medicina, entre tantas outras, surgem conjuntos de dados que possuem uma estrutura na forma de composição [1]. Por exemplo, na área da Geologia quando se analisa a textura dos solos, as frações de areia, silte e argila formam uma composição com três elementos dando uma propriedade ao solo; na área da Química, os elementos que compõem o vinho formam uma composição e determinam que tipo de vinho podem descrever. Estes exemplos de dados podem ser abordados como uma composição, cujos elementos, no seu total, irão formar todo o conjunto de dados.

Um determinado conjunto de dados designa-se composicional se apresenta partes de um todo, dando como exemplo as proporções, percentagens ou concentrações. Nestes casos, o conjunto só pode ser considerado composição se apresentar pelo menos duas componentes. Caso contrário, não há uma parte do todo. Este pormenor é de extrema importância, pois revela que os dados composicionais (*compositional data*) são multivariados (no mínimo bivariados).

Os dados composicionais são observações multivariadas que possuem certas particularidades que os distinguem de outros conjuntos de dados, por exemplo, os dados absolutos. Os dados composicionais contêm apenas informação relativa. Cada elemento de uma observação composicional é designado por componente e deve ser um número real estritamente positivo. Essa componente pode ou não representar a proporção de um todo. Pelo facto dos dados serem relativos, a soma de todas as componentes deve ser igual a uma constante.

Segundo [2], qualquer vetor $\mathbf{x} = (x_1, x_2, \dots, x_D)$ com elementos não negativos, representando proporções de um todo está sujeito à seguinte restrição,

$$x_1 + x_2 + \dots + x_D = k$$

sendo $k > 0$ um número real. De um modo geral tem-se $k = 1$ quando os dados representam proporções, ou $k = 100$ para medições feitas em percentagens.

Os dados composicionais têm várias características e propriedades importantes com consequências para qualquer análise estatística [3]. Na maior parte dos casos, um aspeto bastante comum na análise destes dados é a sua interpretação ser feita através da aplicação de tradicionais técnicas destinadas a dados multivariados reais após transformações convenientes nos

dados originais. As primeiras recomendações que abordam a análise estatística de dados composicionais remete a um artigo de Karl Pearson de 1897 sobre correlações espúrias. O artigo expõe problemas decorrentes do uso de métodos estatísticos tradicionais em proporções. No entanto, as suas advertências foram ignoradas até por volta de 1960 quando o geólogo Felix Chayes também alertou contra a aplicação da análise multivariada tradicional para dados composicionais, a fim de evitar inconsistências pela restrição da soma unitária [1].

A primeira proposta metodológica consistente da análise de dados composicionais surgiu, apenas, nos inícios dos anos 80, em 1982 e 1986, com os trabalhos de John Aitchison, um estatístico escocês e professor na Universidade de Hong Kong [4, 1]. Dos trabalhos de Aitchison, o principal aspeto centra-se na análise estatística de log-razões (*logratios*) entre as componentes de um vetor composicional e os princípios de uma análise de dados composicionais [5].

Dos trabalhos de Aitchison de 1986, este concluiu que uma completa análise das partes que compõem um todo poderia ser realizada em termos de razões das partes da composição, uma vez que as composições apenas contêm informação relativa entre as componentes. Além disso, propôs, ainda, metodologias baseadas em vários tipos de transformações log-razões, visto que a transformação log-razão é uma correspondência biunívoca em \mathbb{R} e o uso matemático de um quociente é mais simples em termos do seu logaritmo. Estas transformações permitiram a aplicação de procedimentos standard da análise multivariada sobre os dados transformados transpondo as conclusões extraídas em termos de dados originais [6].

Apesar das vantagens que as técnicas baseadas em transformações log-razões contribuem para a análise de dados composicionais, tais técnicas não tiveram o sucesso que se esperava obter no seio dos estatísticos. Tal facto pode, talvez, ser explicado devido à tendência habitual de interpretar e analisar resultados em termos absolutos e, consequentemente, a uma menor fluidez no raciocínio numa perspetiva relativa, necessitando que o pensamento seja direcionado em termos de razões [4]. Para uma melhor análise de dados composicionais, na década de 2000, foram publicadas várias contribuições que permitiram uma melhor abordagem sistemática dos métodos propostos por John Aitchison. Tais contribuições podem ser vistas, por exemplo, nas seguintes referências: Aitchison *et al* 2002 e 2005 ([7, 8]), Pawlowsky *et al* 2001 ([9]), Egozcue *et al* 2003 ([10]) e Filzmoser *et al* 2009 ([11]). Atualmente, a análise de dados composicionais pode ser, principalmente, dividida em três etapas:

1. Representação de dados em coordenadas log-razões;
2. Uso de técnicas de análise estatística multivariada sobre os dados em coordenadas log-razões transformadas;
3. Interpretação dos resultados em dois contextos:
 - ✓ coordenadas transformadas;
 - ✓ coordenadas originais.

1.2 Noção de observações atípicas

Na análise estatística de qualquer categoria de dados, por diversas vezes, existem observações que se destacam das restantes por apresentarem comportamentos fora do padrão normal, designando-se por observações atípicas (*outliers*). Não existe uma definição formal definitiva do termo *outlier*, pois existem inúmeras designações associadas, por exemplo, podem ser denominados por: desvios, anomalias, exceções, dados atípicos, entre outras.

O problema dos *outliers* é um dos mais antigos na análise estatística, e durante os últimos anos o interesse nesta área aumentou progressivamente [12]. Qualquer estatístico que analise vários conjuntos de dados, provavelmente, deparou-se com *outliers*. Em 1978, Barnett e Lewis definiram *outlier* como sendo “uma observação (ou subconjunto de observações) que parece ser inconsistente com o restante daquele conjunto de dados” [13]. Segundo estes autores, as principais causas que levam ao aparecimento de *outliers* provêm da variabilidade inerente da população, de erros de medição, e de erros de execução.

Assim, de um modo simples e preciso, pode-se considerar uma observação atípica como sendo qualquer dado que se “desvie” de um “padrão” do conjunto de dados em que se insere, (Figura 1.1 (a)) estando dependente de como se mede esse desvio e como se define esse padrão. É, fundamental, destacar que um *outlier* será sempre um elemento do conjunto, podendo existir mais do que um dado atípico nesse mesmo conjunto (Figura 1.1 (b)), e, além disso, a observação é considerada atípica em relação a um determinado padrão A, pois poderá não ser em relação a um padrão B (Figura 1.1 (c)).

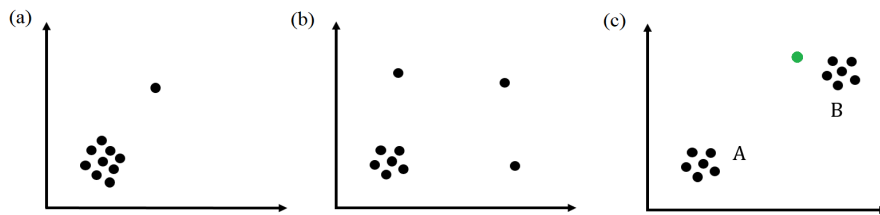


Figura 1.1: (a) Representação de uma observação que se desvia visivelmente de um conjunto de dados, tratando-se de um dado atípico. (b) Representação de três *outliers*. (c) Representação de dois conjuntos de dados, A e B, onde a observação verde é atípica para o conjunto A mas pertence ao conjunto B.

As observações atípicas, por apresentarem comportamentos diferentes dos restantes dados, requerem uma atenção redobrada com um cuidado especial a ter em conta. Numa primeira análise, ao ignorar a existência de tais observações pode induzir a que os resultados obtidos sejam inconsistentes. Na maior parte das situações, leva a que a interpretação de resultados provenientes da aplicação de técnicas estatísticas tradicionais possa ser prejudicada pela presença de *outliers*.

Uma forma de lidar com a existência de *outliers* é a utilização de técnicas estatísticas robustas que mantêm a adequação do modelo aos dados mesmo na presença de *outliers* [14]. Esta questão é, particularmente, de grande foco quando o conjunto de dados contém observações multivariadas, como é o caso dos dados composicionais.

1.3 Motivação para o tema

♣ Informação absoluta ou relativa?

Dois tipos de informação estão subjacentes a qualquer conjunto de dados:

✓ Informação **absoluta**: refere-se aos dados originais nas suas unidades de medidas concretas, como contagens, unidades monetárias, peso, altura, entre outras. Por qualquer redimensionamento das observações originais de uma variável, por exemplo percentagens, o seu valor informativo é afetado ou até mesmo perdido.

✓ Informação **relativa**: refere-se a uma representação de contribuições quantitativamente descritas sobre um todo, cujo valor absoluto é irrelevante. As unidades dos dados são geralmente proporções ou percentagens. Também qualquer tipo de quantidade positiva vista em termos de partes de um todo, por exemplo concentrações de elementos químicos em ppm ou mg/kg, são candidatos a observações com informação relativa.

☆ **Questão primordial**: “Que tipo de informação se está interessado em analisar?”

Na maior parte das vezes, numa análise de dados composicionais, o interesse baseia-se em investigar a variação relativa entre componentes em vez da variação absoluta. Ou seja, pretende-se avaliar os rácios e não as diferenças entre componentes. Considere-se o exemplo apresentado na Tabela 1.1.

Tabela 1.1: Número de habitantes do município de Aguiar da Beira agrupados por grupo etário (em anos).

Grupo etário	0-14	15-24	25-39	40-64	65+
Aguiar da Beira	5	28	195	143	43

Numa escala absoluta, a diferença entre os dois grupos etários 0-14 e 15-24 é de 23 ($28 - 5$). Por outro lado, a diferença entre os grupos 15-24 e 40-64 é de 115 ($143 - 28$). Logo, a diferença do número de habitantes entre os dois grupos mais novos é menor do que nos outros dois grupos etários (15-24 e 40-64). Ao ser comparado numa escala relativa o resultado torna-se totalmente diferente. Nos grupos 0-14 e 15-24, o rácio resulta em $\frac{28}{5} = 5.6$, e nos grupos 15-24 e 40-64 é bastante similar, $\frac{143}{28} \approx 5.1$. Logo, o rácio entre os dois primeiros grupos e entre os outros dois grupos etários é igual naquele município. O número de habitantes no grupo etário 15-24 é cerca de 5 vezes maior do que no grupo dos mais novos, assim como o grupo dos habitantes 40-64 é cerca de 5 vezes maior ao grupo dos 15-24 anos. Portanto, verifica-se uma clara distinção de conclusões destas duas escalas sendo de particular interesse para os dados composicionais avaliar as relações entre variáveis do ponto de vista de uma escala relativa.

⇒ **Conclusão**: Informações relativas, em vez de absolutas, são relevantes para a análise dos dados composicionais.

✿ Detecção de *outliers*

A detecção de *outliers* é uma tarefa importante numa análise estatística de dados multivariados. A presença de *outliers* permite tirar conclusões sobre a qualidade dos dados bem como fenómenos atípicos que possam surgir [15]. O uso de métodos estatísticos robustos deve ser de extrema importância para uma correta análise dos resultados. Nesta dissertação serão estudados **métodos estatísticos robustos** para a detecção de *outliers*, e os desvios das observações atípicas serão detetados em contexto analítico e gráfico.

✿ Representação gráfica

Na representação gráfica, o **diagrama ternário** e o **biplot composicional robusto** serão dois gráficos exploratórios para visualização e identificação dos *outliers*. Em geral, perante uma análise estatística composicional, os dados necessitam de ser transformados. Contudo para a construção de diagramas ternários tais transformações não precisam de ser efetuadas. Estes diagramas têm a particularidade de que as observações só podem ser visualizadas com três componentes de cada vez.

Um biplot permite a visualização de dados multivariados num espaço reduzido. Com este tipo de representação gráfica podem ser detetadas relações entre variáveis e/ou existência de grupos de indivíduos [16]. A sua aplicação em dados composicionais, para detecção de *outliers*, tem sido uma das ferramentas mais utilizadas. Representar os dados composicionais num biplot composicional robusto requer que se faça uso de transformações adequadas.

✿ Agregação dos 3 itens

A informação relativa proveniente dos dados composicionais, a detecção de *outliers* e as representações gráficas são os três itens para os quais esta dissertação se foca. Estes itens complementam-se na medida em que seguem um encadeamento lógico para que a análise aos dados demográficos possa ser realizada de forma apropriada.

Dado que se analisará os dados demográficos em termos composicionais é de todo o interesse que se consiga perceber a razão pela qual se opta por escolher a análise baseada na informação relativa e coloca-se de parte a perspetiva absoluta. O analista deve perceber que vantagens é que advêm de uma análise estatística composicional. A detecção de *outliers* surge como a principal tarefa para a análise dos dados demográficos sendo importante entender quais os melhores métodos estatísticos que permitem a descoberta de dados atípicos. A representação gráfica sucede da necessidade de se conseguir visualizar as observações atípicas, por isso, os gráficos concedem tal visualização.

Como aplicação prática destes três itens no contexto dos dados composicionais, a presente dissertação faz uma análise de dados demográficos reais no contexto de uma análise composicional. Após a identificação das observações atípicas, pelos dois gráficos exploratórios enunciados anteriormente, serão ainda utilizados **cartogramas** para uma melhor visualização geográfica das observações a fim de identificar regiões de Portugal Continental e dos Arquipélagos da Madeira e dos Açores onde se detetam os *outliers* sobressaindo, em termos composicionais, o fluxo migratório dos indivíduos.

1.4 Objetivos e organização da dissertação

A presente dissertação tem como principal objetivo estudar e utilizar técnicas exploratórias na deteção de *outliers* em dados composicionais. Estas serão aplicadas a dados reais. Os dados a analisar correspondem a dados demográficos, provenientes dos Censos de 2011, e baseiam-se nos fluxos migratórios internos em Portugal.

Na decorrente análise estatística a apresentar pretende-se analisar três conjuntos de dados relativos ao número de residentes que mudaram de município entre 2005 e 2011, nos 308 municípios que constituem Portugal (incluindo as Regiões Autónomas da Madeira e dos Açores). Os três conjuntos descrevem os residentes que mudaram de município quanto ao grupo etário, habilitação académica e situação profissional. Cada conjunto é constituído por 308 observações (municípios) sendo cada um descrito por partes, nomeadamente 5, 10 e 3, respetivamente. No esquema da Figura 1.2 encontram-se listadas as partes que descrevem cada conjunto.

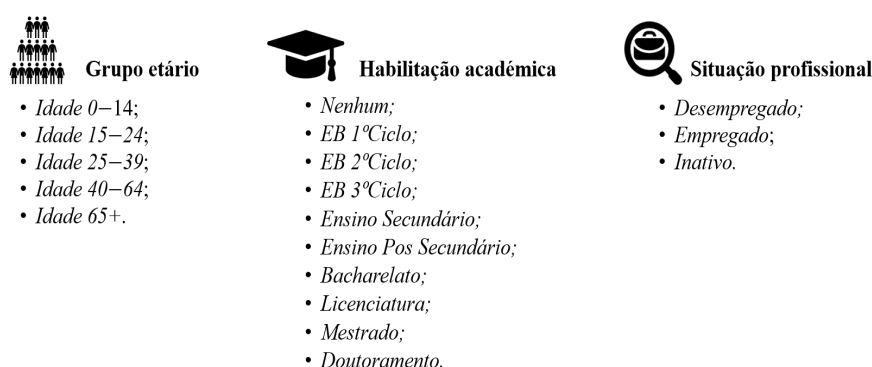


Figura 1.2: Esquema representativo das partes de cada um dos três conjuntos de dados em estudo.

Inicialmente, os conjuntos de dados contêm informação absoluta, ou seja, as composições das observações são referentes ao número de indivíduos que nos Censos de 2011 declararam ter mudado de residência por comparação à data de 2005 e estão caracterizados consoante o grupo etário, a habilitação académica e a situação profissional. A Tabela 1.2 ilustra um exemplo da informação absoluta do conjunto de dados do grupo etário para seis municípios de Portugal.

Tabela 1.2: Parte inicial do conjunto de dados por situação profissional em composições absolutas.

Municípios	<i>Desempregado</i>	<i>Empregado</i>	<i>Inativo</i>
Abrantes	233	876	1012
Águeda	270	1395	1071
Aguiar da Beira	27	120	267
Alandroal	22	101	138
Albergaria-a-Velha	155	984	671
Albufeira	783	3063	2207

Contudo, uma análise do ponto de vista composicional deve ser tida em consideração de forma a cumprir os objetivos. Assim, as frequências absolutas serão convertidas do ponto de vista composicional, isto é, serão consideradas as frequências relativas. Consequentemente, as composições das observações serão proporções referentes à distribuição dos habitantes que entraram em cada município de acordo com o grupo etário, a habilitação académica e a situação profissional. Serve de exemplo a Tabela 1.3 para ilustrar a informação relativa dos seis municípios apresentados na tabela anterior.

Tabela 1.3: Parte inicial do conjunto de dados por situação profissional em composições relativas.

Municípios	<i>Desempregado</i>	<i>Empregado</i>	<i>Inativo</i>
Abrantes	0.110	0.413	0.477
Águeda	0.099	0.510	0.391
Aguiar da Beira	0.065	0.290	0.645
Alandroal	0.084	0.387	0.529
Albergaria-a-Velha	0.086	0.544	0.371
Albufeira	0.129	0.506	0.365

Observação 1.4.1. *Os conjuntos de dados contendo informação absoluta e relativa serão usados para ilustrar conceitos ao longo da dissertação.*

Partindo de uma análise estatística multivariada robusta pretende-se identificar municípios que sejam considerados atípicos quanto a atração (ou não) de residentes, numa perspetiva composicional e em termos de grupo etário, habilitação académica e situação profissional, a fim de investigar a composição do fluxo migratório entre municípios de Portugal. Com os resultados obtidos em modo gráfico, pretende-se detetar as observações atípicas (isto é, municípios *outliers*) e fazer uso das devidas interpretações gráficas.

De forma a alcançar os objetivos mencionados anteriormente, esta dissertação é constituída, para além deste capítulo com uma breve introdução do tema em análise, por mais quatro capítulos e dois apêndices.

No Capítulo 2 apresentar-se-á as metodologias numéricas que serão utilizadas para identificar os *outliers* multivariados nos conjuntos de dados a analisar. Serão introduzidas nas primeiras quatro secções conceitos básicos de dados composicionais, a sua geometria, os princípios para uma correta análise estatística composicional bem como as três transformações log-razões usualmente referenciadas a este tipo de dados, *alr*, *clr* e *ilr*. Associada à transformação *ilr*, na quinta secção será introduzida a técnica da Partição Binária Sequencial (PBS) para construção de bases ortonormais no simplex. De seguida, na sexta secção introduz-se a principal métrica utilizada para identificação de *outliers* multivariados: a distância de Mahalanobis, sendo essencial o estimador MCD (*Minimum Covariance Determinant*) para obter estimativas robustas aplicadas a esta métrica. Na sétima secção serão expostas as propriedades inerentes às transformações log-razões fundamentais para uma adequada análise de dados composicionais. Por fim, na última secção, será introduzido o algoritmo *k*-NN para lidar com as componentes irregulares, nomeadamente os zeros de contagem.

No Capítulo 3 serão apresentadas as metodologias gráficas. Na primeira secção abordar-se-ão as estatísticas descritivas aplicadas aos dados composicionais, nomeadamente o centro e a matriz de variação que dizem respeito às medidas de localização e dispersão dos dados composicionais, respetivamente. Na segunda secção serão apresentadas as ferramentas gráficas, nomeadamente, o diagrama ternário, os biplots para dados multivariados (standard), os biplots para dados composicionais e, ainda, os biplots composicionais robustos.

No Capítulo 4 analisar-se-á, separadamente, os três conjuntos de dados demográficos aplicando as metodologias dos Capítulos 2 e 3. Iniciar-se-á por descrever em maior detalhe os conjuntos brevemente introduzidos no Capítulo 1 e, seguidamente, introduzir-se-á as técnicas utilizadas para análise dos dados. Posteriormente serão realizadas análises dos conjuntos e discutidos os resultados obtidos com exibição de gráficos e tabelas. Tendo em conta que o objetivo se prende com a deteção de *outliers* multivariados, uma abordagem robusta em termos de representação gráfica é indispensável pelo que esta última parte da análise incidirá apenas nos biplots composicionais robustos.

No Capítulo 5 serão apresentadas as principais conclusões do estudo aplicado elaborado no capítulo anterior e, ainda, referindo possíveis trabalhos de investigação que poderão ser desenvolvidos futuramente.

Para completar o trabalho serão incluídos dois apêndices onde constam os *outliers* identificados segundo uma das funções utilizadas no *RStudio* para as metodologias gráficas (Apêndice A) e o poster apresentado nas XXVI Jornadas de Classificação e Análise de Dados, na Escola Superior de Tecnologia e Gestão de Viseu, em Abril de 2019, fruto do presente trabalho (Apêndice B). O trabalho exibido no poster é relativo, somente, aos resultados obtidos pelos biplots composicionais robustos dos conjuntos de dados por habilitação académica e por situação profissional. As aplicações práticas foram todas realizadas com recurso ao software estatístico *RStudio* [17]. O *script* do código realizado para a análise dos dados demográficos e deteção de *outliers* desta dissertação encontra-se publicado online no site GitHub [18].

Capítulo 2

Metodologias Numéricas

2.1 Conceitos básicos de dados composicionais

Em qualquer análise estatística de dados, é importante reconhecer o espaço amostral na qual os dados estão inseridos. A maioria dos dados insere-se no espaço Euclidiano e técnicas estatísticas clássicas fornecem um vasto conjunto de ferramentas para lidar com esses dados [3]. No entanto, os dados composicionais pertencem a uma parte restrita do espaço real e operações associadas a esse espaço desempenham um papel fundamental nos dados composicionais [2].

O espaço amostral dos dados composicionais é representado por classes de equivalência de vetores proporcionais [19]. Todos os vetores com componentes positivas proporcionais representam a mesma composição, uma vez que a multiplicação de um vetor de componentes positivas por uma constante positiva não altera a razão entre as componentes, sugerindo que as composições podem ser vistas como classes de equivalência de vetores proporcionais, isto é, contendo a mesma informação [4].

Definição 2.1.1 (Composições como classes de equivalência). *Dois vetores $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$, onde \mathbb{R}_+^D denota o espaço real D -dimensional ($x_d, y_d > 0, \forall d = 1, 2, \dots, D$) são composicionalmente equivalentes se existe um escalar $\lambda \in \mathbb{R}_+$ tal que $\mathbf{x} = \lambda \mathbf{y}$, ou seja, as composições $\mathbf{x} = (\lambda y_1, \lambda y_2, \dots, \lambda y_D)$ e $\mathbf{y} = (y_1, y_2, \dots, y_D)$ contêm essencialmente a mesma informação relativa, $\forall \lambda \in \mathbb{R}_+$.*

Seja qual for o vetor escolhido de uma classe de equivalência, todos eles podem ser usados para representar essa mesma classe. Desta forma, qualquer composição pode ser explícita em proporções utilizando-se um fator de escala apropriado [4]. De modo a simplificar a análise, é adequado selecionar-se um representante da classe de equivalência, pela normalização dos vetores, para que a soma das componentes seja igual a uma dada constante k , podendo ser 1, 100, 1000 ou qualquer outra constante positiva. A reconstrução de composições pode ser formalizada pela operação de fecho (*Closure operator*).

Definição 2.1.2 (Operação de fecho). *Seja $\mathbf{x} = (x_1, x_2, \dots, x_D) \in \mathbb{R}_+^D$ um vetor de componentes reais e estritamente positivas. O fecho de \mathbf{x} para uma dada constante positiva $k \in \mathbb{R}_+$ é definido por:*

$$\mathcal{C}_k(\mathbf{x}) = \left(\frac{k \cdot x_1}{\sum_{d=1}^D x_d}, \frac{k \cdot x_2}{\sum_{d=1}^D x_d}, \dots, \frac{k \cdot x_D}{\sum_{d=1}^D x_d} \right).$$

Assim, ao aplicar a operação de fecho a uma composição \mathbf{x} origina-se um novo vetor composicional com o mesmo número de elementos, ou seja, esta operação transforma uma composição \mathbf{x} noutra composição equivalente, $\mathcal{C}_k(\mathbf{x})$ [19]. Como consequência, a reestruturação do vetor inicial permite que a soma das suas componentes transformadas seja k . Assim, dois vetores \mathbf{x}, \mathbf{y} em \mathbb{R}_+ são equivalentes se $\mathcal{C}_k(\mathbf{x}) = \mathcal{C}_k(\mathbf{y})$, para qualquer constante $k \in \mathbb{R}_+$ [6].

Exemplo 2.1.1. Considere-se o conjunto da situação profissional cujas composições são definidas por 3 partes, $x_1 = \text{Desempregado}$, $x_2 = \text{Empregado}$ e $x_3 = \text{Inativo}$. Seja $\mathbf{x} = (233, 876, 1012)$ a composição absoluta referente ao município de Abrantes. Sendo $x_1 + x_2 + x_3 = 2121$, então o fecho da composição \mathbf{x} para $k = 1$ será:

$$\begin{aligned} \mathbf{y} &= \mathcal{C}(\mathbf{x}) \\ &= \left(\frac{1 \times 233}{2121}, \frac{1 \times 876}{2121}, \frac{1 \times 1012}{2121} \right) \\ &= (0.110, 0.413, 0.477) \end{aligned}$$

onde a composição resultante \mathbf{y} satisfaz $y_1 + y_2 + y_3 = 1$. Além do mais, \mathbf{x} e \mathbf{y} são composições equivalentes uma vez que se pode escrever $\mathbf{x} = 2121\mathbf{y}$.

A restrição de soma constante concede aos dados composicionais particularidades que os distinguem dos dados no espaço multidimensional. Torna-se necessário definir o espaço amostral dos dados composicionais, denominado por simplex, de modo que este espaço consiga satisfazer essas mesmas particularidades.

Definição 2.1.3 (Simplex). *Considere-se o subconjunto \mathbb{R}_+^D e k uma constante real positiva. O simplex, para composições de D partes, é denotado por S^D e definido por:*

$$S^D = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D) \in \mathbb{R}^D \mid x_d > 0, \sum_{d=1}^D x_d = k \right\}.$$

Apesar de uma composição ser uma classe de equivalência, os representantes dessa classe no simplex também são designados por composições. As componentes de um vetor em S^D são denominadas de partes para destacar o seu carácter composicional. Na maior parte dos casos, no estudo dos dados composicionais, o interesse foca-se em apenas algumas partes da composição [6]. Portanto, importa destacar a definição de subcomposição apresentada de seguida.

Definição 2.1.4 (Subcomposição). *Dada uma composição \mathbf{x} e uma seleção de índices $S = \{d_1, d_2, \dots, d_s\}$, uma subcomposição \mathbf{x}_s , com s partes, é obtida pela aplicação da operação de fecho ao subvetor $(x_{d_1}, x_{d_2}, \dots, x_{d_s})$ de \mathbf{x} . O conjunto de índices S indica as partes seleccionadas para a subcomposição.*

Assim, uma subcomposição é considerada um subconjunto de componentes ou partes de uma composição. O estudo de uma subcomposição exige que os resultados não sejam contraditórios com os obtidos a partir da composição completa [5]. Veja-se o seguinte exemplo.

Exemplo 2.1.2. Considerando o conjunto de dados por grupo etário, cada município é uma composição de 5 partes, $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ em que $x_1 = \text{Idade } 0-14$, $x_2 = \text{Idade } 15-24$, $x_3 = \text{Idade } 25-39$, $x_4 = \text{Idade } 40-64$ e $x_5 = \text{Idade } 65+$. Por exemplo, a composição relativa do município de Albufeira é dada por $\mathbf{x} = (0.040, 0.167, 0.381, 0.311, 0.101)$. Suponha-se que se pretende apenas estudar a composição referente às partes x_1, x_3, x_5 (isto é, os grupos etários *Idade 0-14*, *Idade 25-39* e *Idade 65+*). Então, a subcomposição de \mathbf{x} é dada por

$$\mathbf{x}_3 = \mathcal{C}(x_1, x_3, x_5),$$

sendo que para o município de Albufeira ficaria

$$\begin{aligned} \mathbf{x}_3 &= \mathcal{C}\left(\frac{0.040}{0.040 + 0.381 + 0.101}, \frac{0.381}{0.040 + 0.381 + 0.101}, \frac{0.101}{0.040 + 0.381 + 0.101}\right) \\ &= (0.077, 0.730, 0.193). \end{aligned}$$

2.2 Geometria de Aitchison no simplex

Para uma análise estatística apropriada é essencial considerar propriedades geométricas inerentes ao espaço amostral das observações. A necessidade de um tratamento especial dos dados composicionais rege-se pelo facto destes dados não serem coerentes com a geometria usual Euclidiana. Os dados composicionais possuem uma geometria particular distinta no simplex, designada por geometria de Aitchison.

Na estrutura geométrica dos dados composicionais, o objetivo prende-se em conferir ao simplex de D partes uma estrutura de um espaço vetorial, de modo que seja possível definir bases, linhas retas e outros operadores no simplex. Para tal, Aitchison introduziu duas operações unicamente composicionais, conhecidas na literatura por perturbação (*perturbation*) e potenciação (*powering*). As suas definições encontram-se de seguida.

Definição 2.2.1 (Perturbação). *Considere-se duas composições $\mathbf{x}, \mathbf{y} \in S^D$. A perturbação de \mathbf{x} por \mathbf{y} é uma composição definida por*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \dots, x_D y_D),$$

onde $\mathcal{C}(\cdot)$ é a operação de fecho.

Observação 2.2.1. *A inversa de uma composição \mathbf{y} é $\mathbf{y}^{-1} = \mathcal{C}\left(\frac{1}{y_1}, \frac{1}{y_2}, \dots, \frac{1}{y_D}\right)$, pelo que a perturbação de \mathbf{x} pela inversa de \mathbf{y} , denotada por $\mathbf{x} \oplus \mathbf{y}^{-1}$ ou por $\mathbf{x} \ominus \mathbf{y}$, é dada por $\mathbf{x} \ominus \mathbf{y} = \mathcal{C}\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \dots, \frac{x_D}{y_D}\right)$.*

Exemplo 2.2.1. Do conjunto da situação profissional considere-se as composições relativas referentes aos municípios da Amadora e de Aveiro dadas por $\mathbf{x} = (0.102, 0.575, 0.323)$ e $\mathbf{y} = (0.083, 0.590, 0.327)$, respetivamente. Aplicando a definição de perturbação obtém-se

$$\begin{aligned} \mathbf{x} \oplus \mathbf{y} &= \mathcal{C}(0.102 \times 0.083, 0.575 \times 0.590, 0.323 \times 0.327) \\ &= \mathcal{C}(0.008, 0.339, 0.106) \\ &= \left(\frac{0.008}{0.008 + 0.339 + 0.106}, \frac{0.339}{0.008 + 0.339 + 0.106}, \frac{0.106}{0.008 + 0.339 + 0.106}\right) \\ &= (0.018, 0.748, 0.234). \end{aligned}$$

Pode-se, ainda, medir a perturbação de \mathbf{x} pela inversa de \mathbf{y} :

$$\begin{aligned}\mathbf{x} \ominus \mathbf{y} &= \mathcal{C} \left(\frac{0.102}{0.083}, \frac{0.575}{0.590}, \frac{0.323}{0.327} \right) \\ &= \mathcal{C}(1.229, 0.975, 0.988) \\ &= \left(\frac{1.229}{1.229 + 0.975 + 0.988}, \frac{0.975}{1.229 + 0.975 + 0.988}, \frac{0.988}{1.229 + 0.975 + 0.988} \right) \\ &= (0.398, 0.299, 0.303).\end{aligned}$$

Definição 2.2.2 (Potenciação). *Considere-se uma composição $\mathbf{x} \in S^D$ e um escalar $\alpha \in \mathbb{R}$. A potenciação de \mathbf{x} por α é uma composição definida por*

$$\alpha \otimes \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha),$$

sendo $\mathcal{C}(\cdot)$ a operação de fecho.

Exemplo 2.2.2. Suponha-se que se tem a composição do município da Amadora do Exemplo 2.2.1 e seja $\alpha = 3$ o escalar. Então, a potenciação de \mathbf{x} por 3 é dada por,

$$\begin{aligned}\alpha \otimes \mathbf{x} &= \mathcal{C}(0.102^3, 0.575^3, 0.323^3) \\ &= \mathcal{C}(0.001, 0.190, 0.034) \\ &= \left(\frac{0.001}{0.001 + 0.190 + 0.034}, \frac{0.190}{0.001 + 0.190 + 0.034}, \frac{0.034}{0.001 + 0.190 + 0.034} \right) \\ &= (0.004, 0.844, 0.151).\end{aligned}$$

O produto interno, a norma e a distância de Aitchison são três conceitos cruciais que constituem a geometria de Aitchison, permitindo-lhe conferir uma estrutura métrica no simplex. As duas operações composicionais permitem definir o simplex como uma estrutura de espaço vetorial [19].

Definição 2.2.3 (Produto interno de Aitchison). *O produto interno entre duas composições $\mathbf{x} = (x_1, x_2, \dots, x_D) \in S^D$ e $\mathbf{y} = (y_1, y_2, \dots, y_D) \in S^D$ é definido por*

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{d=1}^D \sum_{j=1}^D \ln \frac{x_d}{x_j} \ln \frac{y_d}{y_j} = \frac{1}{D} \sum_{d=1}^{D-1} \sum_{j=d+1}^D \ln \frac{x_d}{x_j} \ln \frac{y_d}{y_j}.$$

Definição 2.2.4 (Norma de Aitchison). *A norma de uma composição $\mathbf{x} = (x_1, x_2, \dots, x_D) \in S^D$ é a raiz quadrada do produto interno dela própria, definida do seguinte modo*

$$\|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A} = \sqrt{\frac{1}{D} \sum_{d=1}^{D-1} \sum_{j=d+1}^D \left(\ln \frac{x_d}{x_j} \right)^2}.$$

Definição 2.2.5 (Distância de Aitchison). *A distância de Aitchison entre duas composições, \mathbf{x} e $\mathbf{y} \in S^D$, é definida por*

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D} \sum_{d=1}^{D-1} \sum_{j=d+1}^D \left(\ln \frac{x_d}{x_j} - \ln \frac{y_d}{y_j} \right)^2}.$$

Exemplo 2.2.3. Do terceiro conjunto de dados considere-se duas composições $\mathbf{x}, \mathbf{y} \in S^3$, em que $\mathbf{x} = (0.107, 0.473, 0.421)$ e $\mathbf{y} = (0.068, 0.377, 0.556)$, correspondendo aos municípios de Estarreja e Murtosa, respetivamente. Cada uma das três definições anteriores são ilustradas a seguir.

i. Produto interno de Aitchison entre as composições \mathbf{x} e \mathbf{y} :

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle_A &= \frac{1}{2 \times 3} \left[2 \times \left(\ln \frac{0.107}{0.473} \times \ln \frac{0.068}{0.377} + \ln \frac{0.107}{0.421} \times \ln \frac{0.068}{0.556} + \ln \frac{0.473}{0.421} \times \ln \frac{0.377}{0.556} \right) \right] \\ &= 1.793 \end{aligned}$$

ii. Norma de Aitchison das composições \mathbf{x} e \mathbf{y} :

$$\begin{aligned} \|\mathbf{x}\|_A &= \sqrt{\frac{1}{3} \left[\left(\ln \frac{0.107}{0.473} \right)^2 + \left(\ln \frac{0.107}{0.421} \right)^2 + \left(\ln \frac{0.473}{0.421} \right)^2 \right]} = 1.169 \\ \|\mathbf{y}\|_A &= \sqrt{\frac{1}{3} \left[\left(\ln \frac{0.068}{0.377} \right)^2 + \left(\ln \frac{0.068}{0.556} \right)^2 + \left(\ln \frac{0.377}{0.556} \right)^2 \right]} = 1.581 \end{aligned}$$

iii. Distância de Aitchison entre as composições \mathbf{x} e \mathbf{y} :

$$\begin{aligned} d_A(\mathbf{x}, \mathbf{y}) &= \sqrt{\frac{1}{3} \left[\left(\ln \frac{0.107}{0.473} - \ln \frac{0.068}{0.377} \right)^2 + \left(\ln \frac{0.107}{0.421} - \ln \frac{0.068}{0.556} \right)^2 + \left(\ln \frac{0.473}{0.421} - \ln \frac{0.377}{0.556} \right)^2 \right]} \\ &= 0.530 \end{aligned}$$

Tal como ocorre na geometria Euclidiana, a norma e o produto de Aitchison determinam o ângulo α entre dois vetores composicionais \mathbf{x} e \mathbf{y} no espaço simplex, partindo da seguinte relação:

$$\cos \alpha = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_A}{\|\mathbf{x}\|_A \cdot \|\mathbf{y}\|_A}.$$

Exemplo 2.2.4. Do Exemplo 2.2.3, o ângulo no espaço simplex entre as composições dos municípios de Estarreja e Murtosa pode ser obtido do seguinte modo:

$$\cos \alpha = \frac{1.793}{1.169 \times 1.581} = 0.970; \quad \text{logo } \alpha = \cos^{-1}(0.970) = 14.070^\circ$$

2.3 Princípios de uma análise composicional

Para uma análise estatística de dados composicionais, John Aitchison (1986) definiu três princípios sobre os quais devem ser seguidos por qualquer técnica utilizada na análise destes dados onde apenas o rácio entre componentes contém informações relevantes. São eles:

1. Invariância de escala;
2. Invariância de permutação;
3. Coerência subcomposicional.

Embora os dois últimos princípios devam ser cumpridos por qualquer análise estatística, a invariância de escala é um princípio específico que resulta diretamente da definição de dados composicionais [19]. Para uma melhor compreensão destes princípios, uma breve ilustração sobre o significado de cada um deles será apresentado de seguida.

Princípio 2.3.1 (Invariância de escala). *A informação presente numa composição não depende das unidades particulares em que a composição é expressa.*

Este princípio vai ao encontro das Definições 2.1.1 e 2.1.2 que explicam que a multiplicação de um vetor composicional por uma constante arbitrária positiva não altera as proporções entre as partes da composição [19]. Portanto, qualquer característica de uma composição deve ser invariante sob uma mudança de escala.

Exemplo 2.3.1. Suponha-se a informação da composição \mathbf{x} do concelho de Estarreja dada no conjunto de dados relativo à situação profissional, $\mathbf{x} = (177, 785, 699)$. Esta composição é equivalente a ser expressa cada parte em termos de proporção: $\mathbf{y} = (0.107, 0.473, 0.421)$ já que $\mathbf{y} = \frac{1}{1661}\mathbf{x}$ (\mathbf{x} e \mathbf{y} são composições equivalentes). Ao efetuar-se uma análise composicional deve reconhecer-se que composições equivalentes representam a mesma composição. Um requisito indispensável na análise de dados composicionais é o facto de, dada uma função f , a relação $f(\mathbf{y}) = f(\mathbf{x})$ deve verificar-se sempre que \mathbf{x} e \mathbf{y} forem vetores equivalentes [8]. Uma função com esta propriedade é designada de função invariante quanto à escala (*scale invariant*) definida formalmente conforme se segue.

Definição 2.3.1 (Função invariante quanto à escala). *Seja f uma função definida em \mathbb{R}_+^D . A função f é invariante quanto à escala se, para qualquer número real positivo $\lambda \in \mathbb{R}_+$ e para qualquer composição $\mathbf{x} \in S^D$, satisfaz $f(\lambda\mathbf{x}) = f(\mathbf{x})$, isto é, a imagem de vetores composicionalmente equivalentes por meio de f mantém-se sempre a mesma.*

Princípio 2.3.2 (Invariância de permutação). *As conclusões de uma análise estatística de dados composicionais não devem depender da ordem das partes envolvidas.*

Segundo este princípio, a permutação das partes de uma composição não alteram a informação transmitida pelo vetor composicional [19]. Assim, ao aplicar uma análise estatística aos dados composicionais, a ordem das diferentes partes não deverá ter qualquer influência sobre os resultados.

Exemplo 2.3.2. A análise estatística do conjunto dos 308 municípios relativa à situação profissional onde se descrevem 3 partes, x_1, x_2, x_3 , produzirá iguais conclusões se os dados forem tomados na ordem, por exemplo, x_3, x_2, x_1 .

Princípio 2.3.3 (Coerência subcomposicional). *As análises sobre um conjunto de partes de uma composição não devem depender de outras partes não envolvidas, pelo que o estudo de uma subcomposição não pode conduzir a resultados contraditórios com os obtidos a partir da composição total.*

O princípio de coerência subcomposicional pode ser reformulado em dois critérios [20]:

- i. A distância entre duas composições completas deve ser maior ou igual à distância entre as respectivas subcomposições. Isto é, se $\Delta_p(\mathbf{x}, \mathbf{y})$ denota uma qualquer distância entre duas composições de p partes, então

$$\Delta_D(\mathbf{x}, \mathbf{y}) \geq \Delta_d(\mathbf{x}_d, \mathbf{y}_d)$$

onde, \mathbf{x}, \mathbf{y} são composições de D partes e $\mathbf{x}_d, \mathbf{y}_d$ são subcomposições das anteriores com d partes, $d < D$. A este critério atribui-se o nome de dominância subcomposicional (*Subcompositional dominance*);

- ii. Se uma parte “não informativa” for removida, os resultados não devem ser alterados.

Assim sendo, as técnicas para a análise de dados composicionais devem garantir que a sua aplicação a qualquer subcomposição não altere as proporções entre as partes. Por outras palavras, visto que as proporções são a única informação considerada, os resultados da análise devem permanecer invariantes ao usar as mesmas partes, tanto da composição como da subcomposição.

Exemplo 2.3.3. Considere-se o conjunto da habilitação académica cujas composições incluem 10 partes (*Bacharelato, Doutoramento, EB 1º Ciclo, EB 2º Ciclo, EB 3º Ciclo, Ens. Pos Secundário, Ens. Secundário, Licenciatura, Mestrado e Nenhum*). Seja \mathbf{x} a composição completa referente ao município de Monção, e \mathbf{y} a composição constituída pelas partes *Bacharelato, Doutoramento, Licenciatura e Mestrado*. Tendo em conta o significado das partes das duas composições enunciadas é evidente que a composição \mathbf{y} é uma subcomposição de \mathbf{x} . Assim, face ao Princípio 2.3.3, técnicas adequadas de análise deverão conduzir à mesma conclusão relativamente às partes *Bacharelato, Doutoramento, Licenciatura e Mestrado* para as duas composições.

2.4 Transformações de dados composicionais

Tecnicamente, a análise de dados composicionais é frequentemente associada, em primeiro lugar, à aplicação de uma transformação apropriada e, em seguida, ao recurso de metodologias estatísticas usuais sobre os dados transformados [19]. Embora, do ponto de vista técnico, seja verdade em muitos casos, a dificuldade deste procedimento é a interpretação dos resultados. Depois de aplicar uma transformação, não se trabalha mais com as composições originais, mas sim com as suas transformações, e a interpretação dos resultados deve ser adaptada em conformidade.

A transformação a ser aplicada aos dados composicionais baseia-se na análise estatística de log-razões das partes de uma composição conhecida na literatura especializada como Análise de Log-razões (*Logratio Analysis*). Esta abordagem surgiu face ao reconhecimento da importância do princípio de invariância de escala (Princípio 2.3.1), cuja aplicação prática exigia que

se trabalhasse com razões entre as componentes, que anula a constante de escala. A transformação log-razão é uma correspondência biunívoca em \mathbb{R} e o tratamento matemático de um quociente é mais simples em termos do seu logaritmo, por isso Aitchison propôs a adoção de uma técnica de transformação envolvendo logaritmos de razões das componentes [4].

Apesar de se ter a noção de que em qualquer análise de dados composicionais deve-se, sempre, recorrer a uma transformação dos dados, a questão centra-se em saber qual a transformação log-razão a escolher. Deste modo, surge a necessidade de introduzir o conceito de log-contraste (*logcontrast*), que pode ser considerado como uma combinação linear no simplex.

Definição 2.4.1 (Log-contraste). *Considere-se uma composição $\mathbf{x} = (x_1, x_2, \dots, x_D) \in S^D$ e $\mathbf{a} = (\alpha_1, \alpha_2, \dots, \alpha_D) \in \mathbb{R}^D$, com $\sum_{d=1}^D \alpha_d = 0$, o vetor dos coeficientes $\alpha_d \in \mathbb{R}, \forall d = 1, 2, \dots, D$. Um log-contraste de \mathbf{x} é uma combinação linear definida do seguinte modo:*

$$\mathbf{a}' \log \mathbf{x} = \sum_{d=1}^D \alpha_d \ln(x_d).$$

Uma propriedade a destacar deste conceito é a seguinte.

Propriedade 2.4.1. *Log-contrastes são invariantes quanto à escala.*

Demonstração. De facto, para qualquer $k > 0$, e para $\mathbf{a} = (\alpha_1, \alpha_2, \dots, \alpha_D)$ se tem:

$$\begin{aligned} \mathbf{a}' \log(k\mathbf{x}) &= \sum_{d=1}^D \alpha_d \ln(kx_d) \\ &= \sum_{d=1}^D \alpha_d \ln(x_d) + \sum_{d=1}^D \alpha_d \ln(k) \\ &= \sum_{d=1}^D \alpha_d \ln(x_d) + \ln(k) \times \sum_{d=1}^D \alpha_d \\ &= \sum_{d=1}^D \alpha_d \ln(x_d) + \ln(k) \times 0 \\ &= \sum_{d=1}^D \alpha_d \ln(x_d) = \mathbf{a}' \ln(\mathbf{x}) \end{aligned}$$

□

Conclui-se, assim, que as transformações log-razões que sejam log-contrastes são invariantes quanto à escala. De forma a remover a ambiguidade dos log-contrastes, por exemplo podendo manifestar-se na comparação de log-contrastes, surge o conceito de log-contrastes ortogonais apresentado de seguida.

Definição 2.4.2 (Log-contrastes ortogonais). *Dois log-contrastes, $\mathbf{a}' \log(\mathbf{x})$ e $\mathbf{b}' \log(\mathbf{x})$, são ortogonais se $\mathbf{a}'\mathbf{b} = 0$.*

Dificuldades relacionadas com a análise de dados composicionais podem ser ultrapassadas com a escolha de um log-contraste apropriado. Não existe um método específico que auxilie o

analista a adotar uma transformação que seja melhor que todas as outras. É, sim, fundamental que este análise de forma adequada os seus dados, de modo que essa mesma escolha dependerá do problema, da interpretação da composição e dos objetivos que pretende alcançar.

Tal como se constatou na Secção 2.2, os dados composicionais possuem um espaço amostral próprio (simplex) usando a geometria de Aitchison. Esta geometria é substancialmente diferente da geometria Euclidiana, pelo que os métodos estatísticos projetados para a geometria Euclidiana não podem ser aplicados diretamente nos dados composicionais [21]. Contudo, existe uma família de transformações log-razões que permite representar os dados composicionais do simplex no espaço real Euclidiano.

Aitchison (1986) introduziu a primeira transformação de log-razões denominada por transformação de log-razões aditiva (*alr*: *additive logratio*) [2]. No entanto, percebeu que esta transformação não era isométrica¹. Posto isto, introduziu uma nova proposta. Baseada na média geométrica das partes das composições, Aitchison (1986) introduziu a transformação de log-razões centrada (*clr*: *centered logratio*) [2]. Em 2003, Egozcue *et al* ([10]) propuseram a transformação de log-razões isométrica (*ilr*: *isometric logratio*) definida a partir de uma base ortonormal do simplex.

Nas subsecções a seguir, analisar-se-á as três transformações log-razões referidas acima.

2.4.1 Transformação *alr*

Definição 2.4.1.1 (Transformação *alr*). *Seja \mathbf{x} uma composição de D partes no simplex S^D . Designa-se por transformação de log-razões aditiva de \mathbf{x} e, denota-se por $\text{alr}(\mathbf{x})$, relativamente à j -ésima componente, a transformação $\text{alr} : \mathbf{x} \in S^D \rightarrow \mathbf{x}^{(j)} \in \mathbb{R}^{D-1}$, definida por:*

$$\begin{aligned} \mathbf{x}^{(j)} &= \text{alr}_j(\mathbf{x}) \\ &= (x_1^{(j)}, x_2^{(j)}, \dots, x_{D-1}^{(j)}) \\ &= \left(\ln \frac{x_1}{x_j}, \ln \frac{x_2}{x_j}, \dots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \dots, \ln \frac{x_D}{x_j} \right). \end{aligned}$$

Observação 2.4.1.1. *Se for dada uma matriz $\mathbf{X} = [x_{ld}]$ de dados composicionais de dimensão $n \times D$ com composições definidas por $\mathbf{x}_l = (x_{l1}, x_{l2}, \dots, x_{lD})$ nas linhas de \mathbf{X} , para $l = 1, 2, \dots, n$, então a matriz de coordenadas *alr*, relativamente à componente j dos dados, é formada pelas seguintes linhas:*

$$\begin{aligned} \mathbf{x}_l^{(j)} &= \text{alr}_j(\mathbf{x}_l) \\ &= \left(\ln \frac{x_{l1}}{x_{lj}}, \ln \frac{x_{l2}}{x_{lj}}, \dots, \ln \frac{x_{l,j-1}}{x_{lj}}, \ln \frac{x_{l,j+1}}{x_{lj}}, \dots, \ln \frac{x_{lD}}{x_{lj}} \right), \quad l = 1, 2, \dots, n. \end{aligned}$$

O índice $j \in \{1, 2, \dots, D\}$ refere-se à variável que é escolhida como sendo a variável razão nas coordenadas. Qualquer outra parte da composição poderia ser escolhida como referência para figurar no denominador, levando a diferentes transformações *alr* [19]. Esta escolha, usualmente, depende do contexto, mas também da adequação dos resultados para visualização e exploração de dados.

¹As distâncias entre composições em coordenadas transformadas são iguais às distâncias entre composições em coordenadas originais.

Uma vez obtidas as coordenadas *alr* é possível, posteriormente, voltar aos dados composicionais originais. Para isso, recorre-se à inversa da transformação *alr*, denotada por $alr_j^{-1} : \mathbf{x}^{(j)} \in \mathbb{R}^{D-1} \rightarrow \mathbf{x} \in S^D$, para qualquer $j \in \{1, 2, \dots, D\}$, e definida por $\mathbf{x} = (x_1, x_2, \dots, x_D)$ com

$$x_i = \frac{\exp(x_i^{(j)})}{\exp(x_1^{(j)}) + \dots + \exp(x_{j-1}^{(j)}) + \exp(x_{j+1}^{(j)}) + \dots + \exp(x_D^{(j)}) + 1}, \text{ para } i = 1, 2, \dots, D, i \neq j,$$

$$x_j = \frac{1}{\exp(x_1^{(j)}) + \dots + \exp(x_{j-1}^{(j)}) + \exp(x_{j+1}^{(j)}) + \dots + \exp(x_D^{(j)}) + 1}.$$

A transformação *alr* permite reduzir a perturbação e a potenciação a operações comuns de adição e multiplicação no espaço \mathbb{R}^{D-1} . Considerando duas composições \mathbf{x}_1 e $\mathbf{x}_2 \in S^D$, uma constante $\alpha \in \mathbb{R}$ e $j \in \{1, 2, \dots, D\}$, a partir da definição da soma de composições e produto escalar facilmente se prova que (veja-se, por exemplo, [4], pág. 27):

$$alr_j(\mathbf{x}_1 \oplus \mathbf{x}_2) = alr_j(\mathbf{x}_1) + alr_j(\mathbf{x}_2)$$

$$alr_j(\alpha \otimes alr_j(\mathbf{x}_1)) = \alpha alr_j(\mathbf{x}_1)$$

Todavia, como é salientado em [19], a transformação *alr* no espaço euclidiano não satisfaz o produto interno, a norma nem a distância de Aitchison apresentados na Secção 2.2; a título de exemplo repare-se que $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A \neq \langle alr_j(\mathbf{x}_1), alr_j(\mathbf{x}_2) \rangle$. Mais ainda, as coordenadas *alr* conduzem a interpretações que podem ser enganosas em termos das suas partes originais já que a interpretação de uma determinada coordenada não pode ser feita apenas em termos de uma parte. Por exemplo, veja-se que a primeira componente de $\mathbf{x}^{(j)}$ é $\ln\left(\frac{x_1}{x_j}\right)$ e contém informação de x_1 relativa apenas à j -ésima componente e, não a todas as outras partes.

As principais desvantagens práticas das coordenadas *alr* são a subjetividade da escolha da variável razão e o facto desta transformação originar um sistema de coordenadas não ortogonal [19]. Além disso, na transformação *alr* há a possibilidade de escolha da variável razão o que resulta em diferentes transformações *alr* para uma mesma composição \mathbf{x} . Tal ocorrência revela que esta transformação não satisfaz o princípio de invariância de permutação (Princípio 2.3.2) e, por isso, a análise de dados composicionais, através deste tipo de transformação, pode causar conclusões pouco fidedignas [4].

2.4.2 Transformação clr

De forma a evitar problemas relativamente ao uso da transformação *alr*, Aitchison estabeleceu a transformação de log-razões centrada onde se representa uma composição de D partes através de D coordenadas *clr*.

Definição 2.4.2.1 (Transformação *clr*). *Seja \mathbf{x} uma composição de D partes no simplex S^D . Designa-se por transformação de log-razões centrada de \mathbf{x} , e denota-se por $\text{clr}(\mathbf{x})$, a transformação $\text{clr} : \mathbf{x} \in S^D \rightarrow \mathbf{y} \in \mathbb{R}^D$ definida por:*

$$\begin{aligned} \mathbf{y} &= \text{clr}(\mathbf{x}) \\ &= (y_1, y_2, \dots, y_D) \\ &= \left(\ln \frac{x_1}{\sqrt[D]{\prod_{d=1}^D x_d}}, \ln \frac{x_2}{\sqrt[D]{\prod_{d=1}^D x_d}}, \dots, \ln \frac{x_D}{\sqrt[D]{\prod_{d=1}^D x_d}} \right). \end{aligned} \quad (2.1)$$

Observação 2.4.2.1. Se for dada uma matriz $\mathbf{X} = [x_{ld}]$ de dados composicionais de dimensão $n \times D$ com composições definidas por $\mathbf{x}_l = (x_{l1}, x_{l2}, \dots, x_{lD})$ nas linhas de \mathbf{X} , para $l = 1, 2, \dots, n$, então a matriz de coordenadas *clr* é formada pelas seguintes linhas:

$$\mathbf{y}_l = \text{clr}(\mathbf{x}_l) = \left(\ln \frac{x_{l1}}{\sqrt[D]{\prod_{d=1}^D x_{ld}}}, \ln \frac{x_{l2}}{\sqrt[D]{\prod_{d=1}^D x_{ld}}}, \dots, \ln \frac{x_{lD}}{\sqrt[D]{\prod_{d=1}^D x_{ld}}} \right), \quad l = 1, 2, \dots, n. \quad (2.2)$$

O denominador em (2.1) é a média geométrica dos elementos de \mathbf{x} , ou seja

$$g_m(\mathbf{x}) = \sqrt[D]{\prod_{d=1}^D x_d} = \exp \left(\frac{1}{D} \sum_{d=1}^D \ln x_d \right).$$

Note-se que em (2.2), a média geométrica é calculada para cada observação individual.

Numa primeira análise, há uma clara diferença entre a transformação *alr* e *clr*. Esta última transformação evita que haja subjetividade na escolha da variável razão uma vez que utiliza a média geométrica no denominador. Além disso, a transformação *clr* usa D componentes em vez de utilizar $D - 1$ componentes, tal como acontece com a transformação *alr*.

O facto da transformação *clr* conter no denominador a média geométrica determina que cada componente é comparada com uma quantidade global tratando, assim, todas as componentes simetricamente [11]. Esta característica possibilita que os nomes das variáveis originais possam ser usados para a interpretação dos resultados estatísticos com base nos dados transformados por *clr*. Este tipo de transformação, ao ter a vantagem de simplificar a interpretação das variáveis transformadas, é muitas vezes utilizada na construção de biplots para dados composicionais.

Uma desvantagem desta transformação *clr* é que os dados composicionais assim transformados são colineares porque a soma das coordenadas é igual a zero. Efetivamente, segundo (2.1) e (2.2),

$$\begin{aligned} \sum_{j=1}^D y_j &= \sum_{j=1}^D \ln \frac{x_j}{\exp \left(\frac{1}{D} \sum_{d=1}^D \ln x_d \right)} \\ &= \sum_{j=1}^D \left(\ln x_j - \frac{1}{D} \sum_{d=1}^D \ln x_d \right) \\ &= \sum_{j=1}^D \ln x_j - \sum_{j=1}^D \left(\frac{1}{D} \sum_{d=1}^D \ln x_d \right) \\ &= \sum_{j=1}^D \ln x_j - \frac{1}{D} D \sum_{d=1}^D \ln x_d = 0. \end{aligned} \quad (2.3)$$

Esta desvantagem tem um forte impacto na transformação *clr* quando o objetivo se prende na deteção de *outliers*, uma vez que dados colineares inviabilizam a aplicação de técnicas estatísticas robustas [11].

Visto que, na transformação *clr*, o denominador é a média geométrica das partes, a análise de dados composicionais em coordenadas *clr*-transformadas satisfaz o princípio de invariância

de permutação (Princípio 2.3.2). Contudo, a média geométrica de uma composição completa não é necessariamente igual à média geométrica de uma das suas subcomposições. Pelo que, não há garantias que esta transformação satisfaça o princípio de coerência subcomposicional (Princípio 2.3.3).

Contudo, analogamente à transformação *alr*, com a transformação *clr* é possível voltar aos dados composicionais originais. De facto, recorre-se à inversa da transformação *clr*, denotada por $clr^{-1} : \mathbf{y} \in \mathbb{R}^D \rightarrow \mathbf{x} \in S^D$ e definida por $\mathbf{x} = (x_1, x_2, \dots, x_D)$ com

$$x_i = \frac{\exp(y_i)}{\exp(y_1) + \dots + \exp(y_D)}, \quad \text{para } i = 1, 2, \dots, D. \quad (2.4)$$

Observação 2.4.2.1. A inversa da transformação *clr* em (2.4) pode ser escrita a partir da operação de fecho de uma composição (Definição 2.1.2) da seguinte forma:

$$\mathbf{x} = clr^{-1}(\mathbf{y}) = \mathcal{C}(e^{y_1}, e^{y_2}, \dots, e^{y_D}).$$

Observação 2.4.2.2. O vetor \mathbf{y} em (2.1) pode ser obtido em forma matricial do seguinte modo:

$$\mathbf{y} = clr(\mathbf{x}) = \mathbf{R} \cdot \ln(\mathbf{x})$$

com

$$\mathbf{R} = \mathbf{I}_D - \frac{1}{D} \mathbf{1}_D \mathbf{1}_D' = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \frac{1}{D} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}.$$

onde \mathbf{I}_D é a matriz identidade de dimensão $D \times D$, e $\mathbf{1}_D = [1 \dots 1]'$ é um vetor de comprimento D com entradas de 1.

De forma semelhante, também a transformação *clr* permite reduzir a perturbação e a potenciação à soma e ao produto no espaço \mathbb{R}^D . Além disso, outro aspeto importante centra-se no facto da representação de composições em coordenadas *clr*-transformadas poder ser usada para definir uma estrutura métrica no simplex. Considerando duas composições \mathbf{x}_1 e $\mathbf{x}_2 \in S^D$ e uma constante $\alpha \in \mathbb{R}$, tem-se [10]:

$$\begin{aligned} clr(\mathbf{x}_1 \oplus \mathbf{x}_2) &= clr(\mathbf{x}_1) + clr(\mathbf{x}_2) \\ clr(\alpha \otimes \mathbf{x}_1) &= \alpha \cdot clr(\mathbf{x}_1) \\ \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A &= \langle clr(\mathbf{x}_1), clr(\mathbf{x}_2) \rangle \end{aligned} \quad (2.5)$$

$$\|\mathbf{x}_1\|_A = \|clr(\mathbf{x}_1)\| \quad (2.6)$$

$$d_A(\mathbf{x}_1, \mathbf{x}_2) = d(clr(\mathbf{x}_1), clr(\mathbf{x}_2)). \quad (2.7)$$

Face aos resultados expostos, conclui-se que em (2.5) o produto interno de Aitchison é o mesmo que o produto interno Euclidiano entre as composições com coordenadas *clr*; em (2.6) a norma de Aitchison corresponde à norma Euclidiana de uma composição com coordenadas *clr*; e, de igual modo, em (2.7) a distância de Aitchison é a mesma que a distância Euclidiana entre composições com coordenadas *clr*. As três últimas propriedades são importantes uma vez que mostram que todos os conceitos de métrica no simplex são mantidos após a obtenção das coordenadas *clr*, evidenciando que estas coordenadas representam uma isometria.

2.4.3 Transformação ilr e Coordenadas Pivô

Um aspeto crucial para se trabalhar com a geometria de Aitchison consiste na criação de uma base ortonormal e nas suas correspondentes coordenadas. O procedimento para estabelecer uma base ortonormal no simplex foi proposto pela primeira vez por um trabalho de Egozcue e colaboradores, em 2003 [10]. A definição de uma base ortonormal no simplex é enunciada de seguida.

Definição 2.4.3.1 (Base ortonormal no simplex). *Seja S^D um simplex de D partes. O conjunto de vetores $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$, com $\mathbf{e}_i \in S^D$, $i = 1, 2, \dots, D-1$, é uma base ortonormal de S^D se:*

- i. $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_A = 0$, para $i \neq j$;
- ii. $\|\mathbf{e}_i\|_A = 1$, $i = 1, 2, \dots, D-1$.

Exemplo 2.4.3.1. O conjunto $\{\mathbf{e}_1, \mathbf{e}_2\}$ de vetores pertencentes ao simplex S^3 e dados por $\mathbf{e}_1 = \mathcal{C}(e^{\frac{\sqrt{6}}{3}}, e^{-\frac{\sqrt{6}}{6}}, e^{-\frac{\sqrt{6}}{6}}) = \left(\frac{e^{\frac{\sqrt{6}}{3}}}{a}, \frac{e^{-\frac{\sqrt{6}}{6}}}{a}, \frac{e^{-\frac{\sqrt{6}}{6}}}{a}\right)$ e $\mathbf{e}_2 = \mathcal{C}(1, e^{\frac{\sqrt{2}}{2}}, e^{-\frac{\sqrt{2}}{2}}) = \left(\frac{1}{b}, \frac{e^{\frac{\sqrt{2}}{2}}}{b}, \frac{e^{-\frac{\sqrt{2}}{2}}}{b}\right)$, para $a = e^{\frac{\sqrt{6}}{3}} + e^{-\frac{\sqrt{6}}{6}} + e^{-\frac{\sqrt{6}}{6}}$ e $b = 1 + e^{\frac{\sqrt{2}}{2}} + e^{-\frac{\sqrt{2}}{2}}$, constituem uma base ortonormal de S^3 pois satisfazem a Definição 2.4.3.1. De facto, notando primeiramente que:

$$\begin{aligned} clr(\mathbf{e}_1) &= \left(\ln \frac{\frac{e^{\frac{\sqrt{6}}{3}}}{a}}{\sqrt[3]{\frac{e^{\frac{\sqrt{6}}{3}}}{a} \cdot \frac{e^{-\frac{\sqrt{6}}{6}}}{a} \cdot \frac{e^{-\frac{\sqrt{6}}{6}}}{a}}}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a}}{\sqrt[3]{\frac{e^{\frac{\sqrt{6}}{3}}}{a} \cdot \frac{e^{-\frac{\sqrt{6}}{6}}}{a} \cdot \frac{e^{-\frac{\sqrt{6}}{6}}}{a}}}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a}}{\sqrt[3]{\frac{e^{\frac{\sqrt{6}}{3}}}{a} \cdot \frac{e^{-\frac{\sqrt{6}}{6}}}{a} \cdot \frac{e^{-\frac{\sqrt{6}}{6}}}{a}}} \right) \\ &= \left(\ln \frac{\frac{e^{\frac{\sqrt{6}}{3}}}{a}}{\frac{1}{a} \sqrt[3]{e^0}}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a}}{\frac{1}{a} \sqrt[3]{e^0}}, \ln \frac{\frac{e^{-\frac{\sqrt{6}}{6}}}{a}}{\frac{1}{a} \sqrt[3]{e^0}} \right) \\ &= \left(\ln e^{\frac{\sqrt{6}}{3}}, \ln e^{-\frac{\sqrt{6}}{6}}, \ln e^{-\frac{\sqrt{6}}{6}} \right) \\ &= \left(\frac{\sqrt{6}}{3}, -\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6} \right) \end{aligned}$$

e

$$\begin{aligned} clr(\mathbf{e}_2) &= \left(\ln \frac{\frac{1}{b}}{\sqrt[3]{\frac{1}{b} \cdot \frac{e^{\frac{\sqrt{2}}{2}}}{b} \cdot \frac{e^{-\frac{\sqrt{2}}{2}}}{b}}}, \ln \frac{\frac{e^{\frac{\sqrt{2}}{2}}}{b}}{\sqrt[3]{\frac{1}{b} \cdot \frac{e^{\frac{\sqrt{2}}{2}}}{b} \cdot \frac{e^{-\frac{\sqrt{2}}{2}}}{b}}}, \ln \frac{\frac{e^{-\frac{\sqrt{2}}{2}}}{b}}{\sqrt[3]{\frac{1}{b} \cdot \frac{e^{\frac{\sqrt{2}}{2}}}{b} \cdot \frac{e^{-\frac{\sqrt{2}}{2}}}{b}}} \right) \\ &= \left(\ln \frac{\frac{1}{b}}{\frac{1}{b} \sqrt[3]{e^0}}, \ln \frac{\frac{e^{\frac{\sqrt{2}}{2}}}{b}}{\frac{1}{b} \sqrt[3]{e^0}}, \ln \frac{\frac{e^{-\frac{\sqrt{2}}{2}}}{b}}{\frac{1}{b} \sqrt[3]{e^0}} \right) \\ &= \left(\ln e^0, \ln e^{\frac{\sqrt{2}}{2}}, \ln e^{-\frac{\sqrt{2}}{2}} \right) \\ &= \left(0, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right) \end{aligned}$$

então,

$$\text{i. } \langle \mathbf{e}_1, \mathbf{e}_2 \rangle_A = \left\langle \left(\frac{\sqrt{6}}{3}, -\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6} \right), \left(0, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right) \right\rangle = \frac{\sqrt{6}}{3} \times 0 + \frac{-\sqrt{6}}{6} \times \frac{\sqrt{2}}{2} + \frac{-\sqrt{6}}{6} \times \frac{-\sqrt{2}}{2} = 0$$

$$\text{ii. } \|\mathbf{e}_1\|_A = \sqrt{\left(\frac{\sqrt{6}}{3} \right)^2 + \left(-\frac{\sqrt{6}}{6} \right)^2 + \left(-\frac{\sqrt{6}}{6} \right)^2} = 1,$$

$$\|\mathbf{e}_2\|_A = \sqrt{0^2 + \left(\frac{\sqrt{2}}{2} \right)^2 + \left(-\frac{\sqrt{2}}{2} \right)^2} = 1$$

Definição 2.4.3.2 (Transformação *ilr*). *Seja $\mathbf{x} \in S^D$ uma composição de D partes e $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$, $\mathbf{e}_i \in S^D$, uma base ortonormal de S^D . Designa-se por transformação de log-razões isométricas de \mathbf{x} em relação à base $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$, e denota-se por $\text{ilr}(\mathbf{x})$, a transformação $\text{ilr} : \mathbf{x} \in S^D \rightarrow \mathbf{z} \in \mathbb{R}^{D-1}$ definida por:*

$$\begin{aligned} \mathbf{z} &= \text{ilr}(\mathbf{x}) \\ &= (z_1, z_2, \dots, z_{D-1}) \end{aligned}$$

com

$$z_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_A$$

ou, de acordo com (2.5),

$$z_i = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{e}_i) \rangle \text{ para } i = 1, 2, \dots, D-1.$$

Considerando os vetores linha

$$\mathbf{v}_i = [v_{i1} v_{i2} \dots v_{iD}] = \text{clr}(\mathbf{e}_i), \quad i = 1, 2, \dots, D-1, \quad (2.8)$$

tem-se, como verificado em (2.3), que

$$\sum_{j=1}^D v_{ij} = 0 \quad (2.9)$$

pelo que os vetores $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{D-1}$ pertencem a um hiperplano $\mathcal{H} : \{y = (y_1, y_2, \dots, y_D) \in \mathbb{R}^D : y_1 + y_2 + \dots + y_D = 0\}$. Assim, qualquer transformação do tipo *ilr* é baseada na escolha de uma base ortonormal do hiperplano \mathcal{H} que é formado pela transformação *clr*. Nessa base a composição $\mathbf{x} \in S^D$ passa a ser descrita por vetores não colineares transformados por *ilr* [11]. Por conseguinte, a transformação *ilr* resolve o problema da colinearidade dos dados resultante da transformação *clr*. No caso especial dos vetores $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{D-1}$ em (2.8), associados a uma base ortonormal de \mathcal{H} , serem definidos por:

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}} \left(0, \dots, 0, 1, -\frac{1}{D-i}, \dots, -\frac{1}{D-i} \right), \quad i = 1, 2, \dots, D-1 \quad (2.10)$$

com $i-1$ zeros nas primeiras componentes de cada vetor \mathbf{v}_i resulta que as coordenadas *ilr* da composição \mathbf{x} associada à base $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ com

$$\mathbf{e}_i = \text{clr}^{-1}(\mathbf{v}_i)$$

são definidas por

$$\mathbf{z} = \text{ilr}(\mathbf{x}) = (z_1, z_2, \dots, z_{D-1})$$

com

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{d=i+1}^D x_d}}, \quad \text{para } i = 1, 2, \dots, D-1. \quad (2.11)$$

De facto, neste caso, para $i = 1, 2, \dots, D-1$, tem-se:

$$\begin{aligned} z_i &= \langle \text{clr}(\mathbf{x}), \mathbf{v}_i \rangle \\ &= \left\langle \left(\ln \frac{x_1}{\sqrt[D]{\prod_{d=1}^D x_d}}, \dots, \ln \frac{x_D}{\sqrt[D]{\prod_{d=1}^D x_d}} \right), \sqrt{\frac{D-i}{D-i+1}} \left(0, \dots, 0, 1, -\frac{1}{D-i}, \dots, -\frac{1}{D-i} \right) \right\rangle \\ &= \underbrace{0 + \dots + 0}_{(i-1) \text{ termos}} + \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D]{\prod_{d=1}^D x_d}} - \sqrt{\frac{D-i}{D-i+1}} \cdot \frac{1}{D-i} \ln \frac{x_{i+1}}{\sqrt[D]{\prod_{d=1}^D x_d}} - \dots \\ &\quad \dots - \sqrt{\frac{D-i}{D-i+1}} \cdot \frac{1}{D-i} \ln \frac{x_D}{\sqrt[D]{\prod_{d=1}^D x_d}} \\ &= \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D]{\prod_{d=1}^D x_d}} - \sqrt{\frac{D-i}{D-i+1}} \cdot \frac{1}{D-i} \left(\ln \frac{x_{i+1}}{\sqrt[D]{\prod_{d=1}^D x_d}} + \dots + \ln \frac{x_D}{\sqrt[D]{\prod_{d=1}^D x_d}} \right) \\ &= \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D]{\prod_{d=1}^D x_d}} - \sqrt{\frac{D-i}{D-i+1}} \cdot \frac{1}{D-i} \left(\ln \frac{x_{i+1} \dots x_D}{\left(\sqrt[D]{\prod_{d=1}^D x_d} \right)^{D-i}} \right) \\ &= \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D]{\prod_{d=1}^D x_d}} - \sqrt{\frac{D-i}{D-i+1}} \cdot \frac{1}{D-i} \left(\ln \frac{x_{i+1} \dots x_D}{\left(\prod_{d=1}^D x_d \right)^{\frac{D-i}{D}}} \right) \\ &= \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D]{\prod_{d=1}^D x_d}} - \sqrt{\frac{D-i}{D-i+1}} \cdot \frac{1}{D-i} \left(\ln(x_{i+1} \dots x_D) - \frac{D-i}{D} \ln \left(\prod_{d=1}^D x_d \right) \right) \\ &= \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D]{\prod_{d=1}^D x_d}} - \sqrt{\frac{D-i}{D-i+1}} \left(\frac{1}{D-i} \ln(x_{i+1} \dots x_D) - \frac{1}{D} \ln \left(\prod_{d=1}^D x_d \right) \right) \\ &= \sqrt{\frac{D-i}{D-i+1}} \left(\ln \frac{x_i}{\sqrt[D]{\prod_{d=1}^D x_d}} - \frac{1}{D-i} \ln(x_{i+1} \dots x_D) + \frac{1}{D} \ln \left(\prod_{d=1}^D x_d \right) \right) \\ &= \sqrt{\frac{D-i}{D-i+1}} \left(\ln x_i - \frac{1}{D} \ln \left(\prod_{d=1}^D x_d \right) - \frac{1}{D-i} \ln(x_{i+1} \dots x_D) + \frac{1}{D} \ln \left(\prod_{d=1}^D x_d \right) \right) \\ &= \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{(x_{i+1} \dots x_D)^{\frac{1}{D-i}}} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{d=i+1}^D x_d}} = z_i \end{aligned}$$

As coordenadas *ilr* obtidas por (2.11) são, habitualmente, referidas como coordenadas pivô (*pivot coordinates*). A razão para tal designação surge de forma intuitiva: apenas uma parte será configurada para ser um pivô no sentido de ser comparada com todas as restantes partes. De acordo com (2.11), a primeira coordenada, x_1 , é a única que é comparada com as restantes que estão em denominador. Efetivamente, detalhando as $D-1$ componentes em (2.11) resulta que

$$\begin{aligned} z_1 &= \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt[D-1]{\prod_{d=2}^D x_d}} \quad (\text{isto é, } x_1 \text{ é comparada com } x_2, \dots, x_D) \\ z_2 &= \sqrt{\frac{D-2}{D-1}} \ln \frac{x_2}{\sqrt[D-2]{\prod_{d=3}^D x_d}} \quad (\text{isto é, } x_2 \text{ é comparada com } x_3, \dots, x_D) \\ &\vdots \\ z_{D-2} &= \sqrt{\frac{2}{3}} \ln \frac{x_{D-2}}{\sqrt{x_{D-1}x_D}} \quad (\text{isto é, } x_{D-2} \text{ é comparada com } x_{D-1}, x_D) \\ z_{D-1} &= \sqrt{\frac{1}{2}} \ln \frac{x_{D-1}}{x_D} \quad (\text{isto é, } x_{D-1} \text{ é comparada com } x_D) \end{aligned}$$

Exemplo 2.4.3.2. Seja $\mathbf{x} = (0.087, 0.351, 0.562)$ a composição por situação profissional do município da Sertã relativa ao terceiro conjunto referido no Capítulo 1. Ora, as coordenadas pivô para esta composição segundo (2.11) são dadas por:

$$\begin{aligned} z_1 &= \sqrt{\frac{3-1}{3-1+1}} \ln \frac{x_1}{\sqrt[3-1]{\prod_{d=2}^3 x_d}} = \sqrt{\frac{2}{3}} \ln \frac{0.087}{\sqrt{0.351 \times 0.562}} = -1.331 \\ z_2 &= \sqrt{\frac{3-2}{3-2+1}} \ln \frac{x_2}{\sqrt[3-2]{\prod_{d=3}^3 x_d}} = \frac{1}{\sqrt{2}} \ln \frac{0.351}{0.562} = -0.333 \end{aligned}$$

Note-se que as coordenadas *ilr* apenas são obtidas até à componente $D-1$.

Observação 2.4.3.1. Se $\mathbf{X} = [x_{ld}]$ é uma matriz de dados composicionais de dimensão $n \times D$ com composições definidas por $\mathbf{x}_l = (x_{l1}, x_{l2}, \dots, x_{lD})$ nas linhas de \mathbf{X} , para $l = 1, 2, \dots, n$, então a matriz de coordenadas pivô é uma matriz \mathbf{Z} de dimensão $n \times (D-1)$ e de elementos

$$z_{li} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_{li}}{\sqrt[D-i]{\prod_{d=i+1}^D x_{ld}}}, \quad l = 1, 2, \dots, n, \quad e \quad i = 1, 2, \dots, D-1$$

para a linha l e coluna i de \mathbf{Z} .

Como mencionado anteriormente, as coordenadas pivô têm a característica de que x_1 surge unicamente na coordenada z_1 . Esta situação não se verifica para as outras partes. Por exemplo, x_2 aparece em z_1 e em z_2 . Isolar uma parte numa única componente *ilr*-transformada é vantajosa no sentido de ser apenas em z_1 que se resume todas as informações relativas sobre

x_1 . Mais ainda, em z_1 , a parte x_1 é comparada individualmente com as restantes já que:

$$z_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt[D-1]{\prod_{d=2}^D x_d}} = \sqrt{\frac{1}{D(D-1)}} \left(\ln \frac{x_1}{x_2} + \ln \frac{x_1}{x_3} + \dots + \ln \frac{x_1}{x_D} \right).$$

Por tal facto, z_1 pode, assim, ser interpretado como contendo a dominância relativa de x_1 dentro da composição dada. Nenhuma outra parte pode ser interpretada desta forma, pelo que a definição de coordenadas pivô em (2.11) é especificamente projetada em favor de uma interpretação para a primeira parte x_1 da composição original [19].

Por outro lado, observe-se que a primeira coordenada pivô z_1 da transformação ilr , em (2.11), e o primeiro coeficiente y_1 da transformação clr , em (2.1), são proporcionais a um fator de escala dependendo apenas da dimensão D , isto é,

$$z_1 = \sqrt{\frac{D}{D-1}} y_1. \quad (2.12)$$

De facto, ao desenvolver (2.11) na primeira coordenada, (2.12) surge conforme se segue:

$$\begin{aligned} z_1 &= \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt[D-1]{\prod_{d=2}^D x_d}} \\ &= \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\left(\prod_{d=2}^D x_d\right)^{\frac{1}{D-1}}} \times \frac{(x_1)^{\frac{1}{D-1}}}{(x_1)^{\frac{1}{D-1}}} \\ &= \sqrt{\frac{D-1}{D}} \ln \frac{(x_1)^{\frac{D}{D-1}}}{\left(\prod_{d=1}^D x_d\right)^{\frac{1}{D-1}}} \\ &= \sqrt{\frac{D-1}{D}} \ln \frac{(x_1)^{\frac{D}{D-1}}}{\left[\left(\prod_{d=1}^D x_d\right)^{\frac{1}{D}}\right]^{\frac{D}{D-1}}} \\ &= \sqrt{\frac{D-1}{D}} \ln \left(\frac{x_1}{\sqrt[D]{\prod_{d=1}^D x_d}} \right)^{\frac{D}{D-1}} \\ &= \sqrt{\frac{D-1}{D}} \cdot \frac{D}{D-1} \ln \frac{x_1}{\sqrt[D]{\prod_{d=1}^D x_d}} \\ &= \sqrt{\frac{D}{D-1}} \ln \frac{x_1}{\sqrt[D]{\prod_{d=1}^D x_d}} = y_1 \end{aligned}$$

Por (2.1) também y_1 pode ser interpretado como z_1 em termos da dominância relativa de x_1 na composição. Essa relação é dada por,

$$y_1 = \sqrt{\frac{D-1}{D}} z_1. \quad (2.13)$$

Exemplo 2.4.3.3. Atendendo à composição \mathbf{x} do Exemplo 2.4.3.2, a coordenada pivô obtida neste exemplo pode também ser obtida segundo (2.12), de tal modo que se tem

$$z_1 = \sqrt{\frac{3}{3-1}} \ln \frac{x_1}{\sqrt[3]{\prod_{d=1}^3 x_d}} = \sqrt{\frac{3}{2}} \ln \frac{0.087}{\sqrt[3]{0.087 \times 0.351 \times 0.562}} = -1.331$$

E, por (2.13), a primeira coordenada *clr* será:

$$y_1 = \sqrt{\frac{3-1}{3}} z_1 = \sqrt{\frac{2}{3}} \times (-1.331) = -1.087$$

Da mesma forma, na transformação *ilr* é possível voltar aos dados composicionais originais. Recorrendo à inversa da transformação *ilr*, denotada por $ilr^{-1} : \mathbf{z} \in \mathbb{R}^{D-1} \rightarrow \mathbf{x} \in S^D$, esta é definida por $\mathbf{x} = (x_1, x_2, \dots, x_D)$ do seguinte modo [19]:

$$\begin{aligned} x_1 &= \exp \left(\sqrt{\frac{D-1}{D}} z_1 \right) \\ x_j &= \exp \left(- \sum_{d=1}^{j-1} \frac{1}{\sqrt{(D-d+1)(D-d)}} z_d + \sqrt{\frac{D-j}{D-j+1}} z_j \right), \quad j = 2, \dots, D-1, \\ x_D &= \exp \left(- \sum_{d=1}^{D-1} \frac{1}{\sqrt{(D-d+1)(D-d)}} z_d \right). \end{aligned}$$

Note-se que em S^D podem ser definidas inúmeras bases ortonormais. A transformação *ilr* depende da escolha de uma base ortonormal, pelo que diferentes bases ortonormais estão associadas a diferentes coordenadas *ilr*-transformadas. Importa realçar que em (2.11) as coordenadas pivô estão definidas daquele modo uma vez que, associado a essas coordenadas, encontra-se a base ortonormal definida em (2.10). Na Subsecção 2.5.2 será abordado em mais detalhe esta questão.

As equações (2.1) e (2.11) podem ser reescritas em notação matricial para expressar a relação linear existente entre a transformação *clr* e *ilr*. De facto, prova-se que:

$$\mathbf{y} = \mathbf{V}' \mathbf{z} \quad (2.14)$$

onde $\mathbf{V}' = [\mathbf{v}_1 \dots \mathbf{v}_{D-1}]$ é uma matriz de dimensão $D \times (D-1)$, onde em cada coluna estão os vetores de uma base ortonormal de \mathcal{H} dados por (2.10) [10]. Multiplicando à esquerda ambos os membros da equação (2.14) por \mathbf{V} e sabendo que $\mathbf{V}\mathbf{V}' = \mathbf{I}_{D-1}$, obtém-se:

$$\mathbf{z} = \mathbf{V} \mathbf{y} \quad (2.15)$$

Posteriormente, esta relação linear será útil no contexto da Análise de Componentes Principais (ACP) quando for introduzido o tópico da construção de biplots composicionais robustos.

Observação 2.4.3.2. Tendo em conta a transformação realizada, as equações (2.14) e (2.15) podem ser reescritas em termos da composição \mathbf{x} original do seguinte modo: $\text{clr}(\mathbf{x}) = \mathbf{V}' \text{ilr}(\mathbf{x})$ e $\text{ilr}(\mathbf{x}) = \mathbf{V} \text{clr}(\mathbf{x})$, respetivamente.

As coordenadas *ilr*-transformadas, obtidas a partir de uma base ortonormal, garantem que a correspondência entre o simplex S^D e o espaço Euclidiano \mathbb{R}^{D-1} seja isométrica. Assim, esta transformação preserva a propriedade de isometria entre o simplex e o espaço real Euclidiano.

Identicamente à transformação *clr*, também a representação de composições em coordenadas *ilr*-transformadas pode ser usada para definir uma estrutura métrica no simplex. No entanto, tem a particularidade de que o produto, a norma e a distância de Aitchison correspondem ao espaço real \mathbb{R}^{D-1} que é isomorfo a S^D . Considerando duas composições \mathbf{x}_1 e $\mathbf{x}_2 \in S^D$ e uma constante $\alpha \in \mathbb{R}$, obtém-se [10]:

$$\begin{aligned} ilr(\mathbf{x}_1 \oplus \mathbf{x}_2) &= ilr(\mathbf{x}_1) + ilr(\mathbf{x}_2) \\ ilr(\alpha \otimes \mathbf{x}_1) &= \alpha \cdot ilr(\mathbf{x}_1) \\ \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A &= \langle ilr(\mathbf{x}_1), ilr(\mathbf{x}_2) \rangle \\ \|\mathbf{x}_1\|_A &= \|ilr(\mathbf{x}_1)\| \\ d_A(\mathbf{x}_1, \mathbf{x}_2) &= d(ilr(\mathbf{x}_1), ilr(\mathbf{x}_2)). \end{aligned}$$

Assim, a analogia da métrica no espaço S^D à custa da métrica no espaço Euclidiano permite aplicar as técnicas estatísticas usuais aos dados composicionais em termos das suas coordenadas *ilr*-transformadas.

Duas grandes vantagens da transformação *ilr* resulta do facto desta transformação preservar propriedades vantajosas da transformação *clr* e satisfazer todos os princípios de uma análise de dados composicionais. Contudo, apesar das suas propriedades geométricas vantajosas, as correlações calculadas com base na transformação *ilr* não podem ser interpretadas de acordo com as variáveis originais. Isto ocorre pois as variáveis *ilr*-transformadas estão relacionadas com as variáveis originais através de funções não lineares (Definição 2.4.3.2).

Atualmente, para a maioria dos métodos estatísticos em dados composicionais, a transformação *ilr* é uma das mais utilizadas para avaliação dos resultados. No entanto, existem algumas exceções, como é o caso do biplot composicional. Neste caso, a transformação *clr* torna-se favorável, pois permite uma interpretação mais fácil. Por razões que serão posteriormente enunciadas na Secção 2.7, para a deteção de *outliers* a transformação *ilr* é a preferida na família das transformações log-razões. A junção destas duas transformações torna-se um aspeto fulcral na identificação de *outliers* em dados composicionais usando biplots.

2.5 Base ortonormal no simplex

No espaço do simplex, S^D , existem infinitas bases ortonormais. Uma base ortonormal permite representar qualquer elemento do simplex, ou seja, existindo uma composição, os seus elementos podem ser representados pelas suas coordenadas em eixos ortogonais tal como, normalmente, se faz no espaço real [10].

Representar os elementos do simplex S^D , através de uma base ortonormal com $D - 1$ vetores, pode ser visto como uma transformação que atribui a cada elemento x do simplex as suas correspondentes $D - 1$ coordenadas. Essa transformação é uma bijeção e é isométrica, ou seja, as distâncias e ângulos de Aitchison no simplex são transformados em distâncias e ângulos euclidianos comuns no espaço real $(D - 1)$ -dimensional das coordenadas [10]. Quando a base ortonormal é selecionada, os dados composicionais podem ser transformados por *ilr*, e

consequentemente, representados pelas suas coordenadas e analisados como um conjunto de dados multivariados no espaço real. Assim, a aquisição de coordenadas *ilr*-transformadas dos elementos do simplex permite um tratamento eficaz e prático dos dados composicionais. A única dificuldade em usar este tipo de transformação reside na seleção da base ortonormal de referência.

Algumas bases ortogonais especiais podem ser obtidas através de uma técnica conhecida por Partição Binária Sequencial (PBS). A primeira proposta foi sugerida, em 2003, por Egozcue e colaboradores ([10]). Esta técnica é definida em termos de uma partição predefinida das partes da composição, o que implica que uma base construída por PBS dependerá da escolha dessa mesma partição. Na literatura são referidas duas formas de escolher partições com interesse prático. Uma primeira escolha baseia-se no conceito de equilíbrio (*balances*) entre grupos de partes e foi sugerida por Egozcue e colaboradores, em 2005 ([22]). Uma segunda escolha foi sugerida por Filzmoser e colaboradores, em 2009 ([23]), com o interesse de garantir a interpretabilidade das coordenadas transformadas.

O método baseado no conceito de equilíbrio tem a desvantagem de ser necessário um conhecimento à priori do problema em estudo com vista a predefinir a separação das partes da composição de modo a esta ser interpretável para o problema em causa. A PBS tende a ficar confusa para composições que envolvem muitas partes e/ou quando nenhuma informação a priori sobre o problema está disponível, o que pode condicionar, na prática, a eficácia de análise com dados em coordenadas *ilr*-transformadas [24]. Com o objetivo de ultrapassar esse constrangimento, um segundo método de construção de partições foi desenvolvido que garante uma escolha adequada de bases de modo que cada uma das coordenadas ortogonais explique todas as log-razões de uma variável original [23].

Importa mencionar que os dados demográficos que serão analisados na presente dissertação, não contêm qualquer informação previa que se possa usar para construir uma partição das partes composicionais. Deste modo, para garantir uma melhor eficácia de análise dos três conjuntos de dados (Grupo etário, Habilitação académica e Situação profissional) escolher-se-á a PBS proposta por Filzmoser e colaboradores para obtenção de uma base dos dados em coordenadas *ilr*. Assim sendo, nas subsecções seguintes estudar-se-á esta técnica de partição.

2.5.1 Generalização das coordenadas pivô

As coordenadas pivô introduzidas na Subsecção 2.4.3 suportam uma interpretação especialmente da primeira parte composicional x_1 , uma vez que a primeira coordenada descreve exclusivamente toda a informação relativa sobre x_1 . No entanto, pode ser interessante obter uma interpretação específica para uma única parte dentro da composição sem que essa seja a primeira parte. Nesse caso, pode-se simplesmente permutar as partes composicionais de forma a que a parte que interessa fique colocada na primeira posição e, assim, as coordenadas pivô são construídas para a composição permutada [19].

Dada uma composição $\mathbf{x} \in S^D$, denota-se por $\mathbf{x}^{(\ell)} = (x_1^{(\ell)}, x_2^{(\ell)}, \dots, x_D^{(\ell)})$ uma permutação de \mathbf{x} . Ao executar D permutações tal que na ℓ -ésima permutação a parte x_ℓ , $\ell \in \{1, 2, \dots, D\}$, de \mathbf{x} ocupe a primeira posição, obtém-se D vetores composicionais que contêm a mesma informação relativa de \mathbf{x} definidos por $\mathbf{x}^{(\ell)} = (x_\ell, x_1, x_2, \dots, x_{\ell-1}, x_{\ell+1}, \dots, x_D)$, $\ell = 1, 2, \dots, D$.

Assim, a composição permutada atrás referida, $\mathbf{x}^{(\ell)}$, corresponde a ter

$$\mathbf{x}^{(\ell)} = (x_\ell, x_1, x_2, \dots, x_{\ell-1}, x_{\ell+1}, \dots, x_D). \quad (2.16)$$

Repare-se que apenas a parte $x_1^{(\ell)}$ é colocada na primeira posição e a ordem das restantes partes permanece inalterada. Por exemplo, para $D = 5$, permutando as partes de uma composição $\mathbf{x} \in S^5$, os cinco vetores seguintes correspondem às cinco permutações desejadas e contêm a mesma informação relativa de \mathbf{x} :

$$\begin{aligned} \mathbf{x}^{(1)} &= \mathbf{x} = (x_1, x_2, x_3, x_4, x_5); \\ \mathbf{x}^{(2)} &= (x_2, x_1, x_3, x_4, x_5); \\ \mathbf{x}^{(3)} &= (x_3, x_1, x_2, x_4, x_5); \\ \mathbf{x}^{(4)} &= (x_4, x_1, x_2, x_3, x_5); \\ \mathbf{x}^{(5)} &= (x_5, x_1, x_2, x_3, x_4). \end{aligned}$$

A transformação *ilr* para $\mathbf{x}^{(\ell)}$ resulta em $\mathbf{z}^{(\ell)} = (z_1^{(\ell)}, z_2^{(\ell)}, \dots, z_{D-1}^{(\ell)})$ onde, tendo em conta (2.11), as componentes de $\mathbf{z}^{(\ell)}$ são definidas por:

$$z_i^{(\ell)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(\ell)}}{\sqrt[D-i]{\prod_{d=i+1}^D x_d^{(\ell)}}} \quad \text{para } i = 1, 2, \dots, D-1. \quad (2.17)$$

Observe-se que, para cada permutação $\ell = 1, 2, \dots, D$, $x_1^{(\ell)}$ corresponde à única componente que pode ser explicada por $\mathbf{z}^{(\ell)}$ por apenas uma componente que é a componente $z_1^{(\ell)}$ [24]. Em $\mathbf{z}^{(\ell)}$ o foco centra-se em $z_1^{(\ell)}$, que explica toda a informação relativa sobre a parte x_ℓ . Usando as D permutações $z^{(1)}, z^{(2)}, \dots, z^{(D)}$, a interpretação de resultados da análise de dados composicionais em termos destas coordenadas ortogonais centra-se nas suas primeiras componentes conduzindo a conclusões em termos de coordenadas originais, x_1, x_2, \dots, x_D [5].

Observação 2.5.1.1. *Se for dada uma matriz $\mathbf{X} = [x_{ld}]$, de dados composicionais, de dimensão $n \times D$ com composições definidas por $\mathbf{x}_l = (x_{l1}, x_{l2}, \dots, x_{lD})$ nas linhas de \mathbf{X} , para $l = 1, 2, \dots, n$, então a matriz de coordenadas pivô $\mathbf{Z}^{(\ell)} = [z_{li}^{(\ell)}]$, de dimensão $n \times (D-1)$, com ênfase na parte x_ℓ , para algum $\ell = 1, 2, \dots, D$, é formada por elementos*

$$z_{li}^{(\ell)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_{li}^{(\ell)}}{\sqrt[D-i]{\prod_{d=i+1}^D x_{ld}^{(\ell)}}},$$

onde $x_{ld}^{(\ell)}$ é o d -ésimo elemento para a l -ésima linha da matriz de dados permutados dada por

$$(x_{l\ell}, x_{l1}, \dots, x_{l,\ell-1}, x_{l,\ell+1}, \dots, x_{lD}).$$

A permutação em (2.16) pode ser definida através de uma matriz de permutação $\mathbf{P}^{(\ell)}$, de dimensão $D \times D$ com entradas 0 e 1, onde o elemento 1 em cada linha encontra-se na coluna da posição permutada. A título ilustrativo, considere-se, por exemplo, a composição

$\mathbf{x} = (x_1, x_2, x_3)$ e a parte que se pretende interpretar é x_3 , então a composição permutada de interesse é $\mathbf{x}^{(3)} = (x_3, x_1, x_2)$, a qual pode ser obtida por

$$\mathbf{x}^{(3)} = \begin{pmatrix} x_3 \\ x_1 \\ x_2 \end{pmatrix} = \mathbf{P}^{(3)} \mathbf{x} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

onde $\mathbf{P}^{(3)}$ é a matriz de permutação. Com esta matriz de permutação é possível definir uma nova base para as coordenadas pivô, $\mathbf{z}^{(\ell)}$. Os vetores da base ortonormal para a composição não permutada foram definidos em (2.10), e destes obtida a matriz \mathbf{V} . A matriz contendo os vetores da base ortonormal para a composição permutada será

$$\mathbf{V}^{(\ell)} = \mathbf{P}^{(\ell)} \mathbf{V}'. \quad (2.18)$$

Assim, as linhas da matriz \mathbf{V}' são permutadas da mesma forma que as partes da composição. Considere-se a seguinte matriz

$$\mathbf{Q}^{(\ell)} = (\mathbf{V} \mathbf{V}^{(\ell)})'. \quad (2.19)$$

Ora, $\mathbf{Q}^{(\ell)}$ é uma matriz ortonormal pois,

$$(\mathbf{Q}^{(\ell)})' \mathbf{Q}^{(\ell)} = \mathbf{Q}^{(\ell)} (\mathbf{Q}^{(\ell)})' = \mathbf{I}_{D-1}.$$

As coordenadas pivô para uma composição permutada são obtidas em termos da composição não permutada usando a matriz $\mathbf{Q}^{(\ell)}$,

$$\begin{aligned} \mathbf{z}^{(\ell)} &= \mathbf{Q}^{(\ell)} \mathbf{z} \\ &= (\mathbf{V} \mathbf{V}^{(\ell)})' \mathbf{z} \text{ por (2.19)} \\ &= (\mathbf{V} \mathbf{P}^{(\ell)} \mathbf{V}')' \mathbf{z} \text{ por (2.18)} \\ &= \mathbf{V} (\mathbf{P}^{(\ell)})' \mathbf{V}' \mathbf{z}. \end{aligned} \quad (2.20)$$

A equação em (2.20) mostra, de uma maneira elegante, como é que a base pode ser alterada para expressar composições num sistema diferente da base ortonormal não permutada, permitindo uma interpretação concisa da ℓ -ésima parte composicional. Além disso, (2.20) demonstra que outra escolha do sistema de coordenadas pivô é apenas uma rotação do sistema original.

De forma semelhante, tal como se constatou em (2.12) para o caso especial de z_1 , também a relação entre $z_1^{(\ell)}$ e as coordenadas clr , y_ℓ , pode ser generalizada do seguinte modo,

$$z_1^{(\ell)} = \sqrt{\frac{D}{D-1}} y_\ell.$$

A vantagem de obter uma interpretação para cada parte composicional é resgatada pela necessidade de construir um sistemas de D coordenadas, onde apenas uma variável é de interesse primário (na primeira posição) [25]. É de destacar que para qualquer $\ell = 1, \dots, D$ a primeira coordenada $z_1^{(\ell)}$ corresponde sempre à coordenada clr -transformada y_ℓ , diferindo apenas pela constante $\sqrt{\frac{D}{D-1}}$.

2.5.2 Partição Binária Sequencial

A escolha de uma base ortonormal específica em S^D é crucial para a interpretação de coordenadas. As bases que resultam da PBS são preferíveis porque permitem a interpretação em termos de partes agrupadas da composição [26].

Seja $\mathbf{x} \in S^D$ uma composição e $\mathbf{x}^{(\ell)} = (x_1^{(\ell)}, x_2^{(\ell)}, \dots, x_D^{(\ell)})$ uma permutação de \mathbf{x} . Quando se considera uma partição do vetor $\mathbf{x}^{(\ell)}$ de modo que na primeira posição se tenha um grupo formado pela componente $x_1^{(\ell)}$, obtém-se um outro grupo formado pelas partes $x_2^{(\ell)}, x_3^{(\ell)}, \dots, x_D^{(\ell)}$. Por outro lado, estabelecendo uma segunda partição separa-se o grupo $\{x_2^{(\ell)}, x_3^{(\ell)}, \dots, x_D^{(\ell)}\}$ obtido na etapa anterior de modo que se tenha um grupo formado pela parte $x_2^{(\ell)}$ e outro grupo formado pelas partes $x_3^{(\ell)}, \dots, x_D^{(\ell)}$. Procedendo deste modo até à partição de ordem $D - 1$ irá obter-se um grupo formado pela parte $x_{D-1}^{(\ell)}$ e outro pela parte $x_D^{(\ell)}$.

A PBS para a partição descrita anteriormente é construída da seguinte forma. Num primeiro passo, as partes da composição são divididas em dois grupos: partes no primeiro grupo são codificadas por $+1$ e partes no segundo grupo são codificadas por -1 . Assim obtém-se a primeira coordenada que separa as partes $+1$ e -1 . Na segunda etapa e nas seguintes etapas, um grupo anterior de partes é dividido em dois novos grupos. Os grupos são codificados de forma semelhante por $+1$ e -1 , enquanto as componentes não envolvidas são codificadas com 0 . O número de etapas necessárias para que todos os grupos contenham uma única componente é exatamente $D - 1$, ou seja, a dimensão de S^D [26]. Na Tabela 2.1 encontra-se a PBS para esta partição onde a primeira partição confronta a parte $x_1^{(\ell)}$ com as restantes partes da composição.

Tabela 2.1: Codificação de uma PBS para construção de uma base ortonormal onde a primeira partição confronta a parte $x_1^{(\ell)}$ com as restantes partes da composição.

Etapa i da PBS	Resultados de cada etapa da PBS								r	s
	$x_1^{(\ell)}$	$x_2^{(\ell)}$	$x_3^{(\ell)}$	$x_4^{(\ell)}$	\dots	$x_{D-2}^{(\ell)}$	$x_{D-1}^{(\ell)}$	$x_D^{(\ell)}$		
1	$+1$	-1	-1	-1	\dots	-1	-1	-1	1	$D - 1$
2	0	$+1$	-1	-1	\dots	-1	-1	-1	1	$D - 2$
3	0	0	$+1$	-1	\dots	-1	-1	-1	1	$D - 3$
4	0	0	0	$+1$	\dots	-1	-1	-1	1	$D - 4$
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots	\vdots	\vdots
$D - 2$	0	0	0	0	\dots	$+1$	-1	-1	1	2
$D - 1$	0	0	0	0	\dots	0	$+1$	-1	1	1

Sem perda de generalidade, assume-se que na i -ésima etapa da PBS um grupo de $r + s$ partes é dividido em dois grupos. Um dos grupos é formado por r partes (etiquetadas com $+1$) e, tal como se verifica na Tabela 2.1, em cada etapa da PBS a etiqueta $+1$ surge apenas uma vez para identificar a parte que é separada do restante grupo de partes, pelo que o grupo formado por r partes contém unicamente 1 parte. Por outro lado, o outro grupo é formado por s partes (etiquetadas por -1) e irá conter as restantes $D - i$ partes, $i = 1, 2, \dots, D$. Nestas condições, o vetor \mathbf{e}_i da base ortonormal associado à i -ésima etapa da PBS é dado pela

seguinte expressão:

$$\begin{aligned}\mathbf{e}_i &= \text{clr}^{-1}(\mathbf{v}_i) = \mathcal{C}(\exp(\mathbf{v}_i)) \\ &= \mathcal{C}[\exp(v_{i1}v_{i2} \dots v_{iD})],\end{aligned}\tag{2.21}$$

em que v_{ij} corresponde à j -ésima coordenada clr -transformada do vetor \mathbf{e}_i , $i = 1, 2, \dots, D-1$, e é dada por

$$v_{ij} = \begin{cases} \sqrt{\frac{D-i}{D-i+1}}, & \text{se a etiqueta em } (i, j) = +1 \\ \frac{-1}{\sqrt{(D-i)(D-i+1)}}, & \text{se a etiqueta em } (i, j) = -1 \\ 0, & \text{se a etiqueta em } (i, j) = 0 \end{cases}$$

Pelo facto dos elementos v_{ij} de \mathbf{v}_i satisfazerem (2.9) surge o conceito de matriz de contrastes (*contrast matrix*) associado a uma base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ e, de um modo geral, é definida conforme se segue.

Definição 2.5.2.1 (Matriz de contrastes). *Seja $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ uma base ortonormal do simplex. Uma matriz $\mathbf{V}_{D-1 \times D} = [v_{ij}]$, tal que a i -ésima linha $\mathbf{v}_i = \text{clr}(\mathbf{e}_i)$, $i = 1, 2, \dots, D-1$, é designada por matriz de contrastes associada à base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$.*

A matriz de contrastes associada à PBS, apresentada na Tabela 2.1, contendo as coordenadas clr -transformadas dos vetores \mathbf{e}_i associados a cada etapa $i = 1, 2, \dots, D-1$ da partição, encontra-se na Tabela 2.2.

Tabela 2.2: Processo para obtenção da matriz de contrastes resultante da PBS obtida por partição.

Matriz de contrastes, $\mathbf{V} = [v_{ij}]$							
Etapa i da PBS	v_{i1}	v_{i2}	v_{i3}	\dots	v_{iD-1}	v_{iD}	$\mathbf{v}_i^{(\ell)}$
1	$\sqrt{\frac{D-1}{D-1+1}}$	$\frac{-1}{\sqrt{(D-1)(D-1+1)}}$	$\frac{-1}{\sqrt{(D-1)(D-1+1)}}$	\dots	$\frac{-1}{\sqrt{(D-1)(D-1+1)}}$	$\frac{-1}{\sqrt{(D-1)(D-1+1)}}$	$\mathbf{v}_1^{(\ell)}$
2	0	$\sqrt{\frac{D-2}{D-2+1}}$	$\frac{-1}{\sqrt{(D-2)(D-2+1)}}$	\dots	$\frac{-1}{\sqrt{(D-2)(D-2+1)}}$	$\frac{-1}{\sqrt{(D-2)(D-2+1)}}$	$\mathbf{v}_2^{(\ell)}$
3	0	0	$\sqrt{\frac{D-3}{D-3+1}}$	\dots	$\frac{-1}{\sqrt{(D-3)(D-3+1)}}$	$\frac{-1}{\sqrt{(D-3)(D-3+1)}}$	$\mathbf{v}_3^{(\ell)}$
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots
$D-1$	0	0	0	\dots	$\frac{1}{\sqrt{2}}$	$\frac{-1}{\sqrt{2}}$	$\mathbf{v}_{D-1}^{(\ell)}$

Exemplo 2.5.2.1. Seja $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ uma composição de 5 partes, sendo que cada parte corresponde ao conjunto do grupo etário referido no Capítulo 1. De acordo com o procedimento de construção da PBS, na Tabela 2.3 encontra-se a PBS para a composição \mathbf{x} .

Tabela 2.3: PBS de uma composição de 5 partes.

Etapa i da PBS	$x_1^{(\ell)}$	$x_2^{(\ell)}$	$x_3^{(\ell)}$	$x_4^{(\ell)}$	$x_5^{(\ell)}$	r	s
1	+1	-1	-1	-1	-1	1	4
2	0	+1	-1	-1	-1	1	3
3	0	0	+1	-1	-1	1	2
4	0	0	0	+1	-1	1	1

Os valores v_{ij} correspondentes às coordenadas *clr*-transformadas dos vetores da base ortonormal do simplex S^5 obtidas por esta partição encontram-se na Tabela 2.4, e foram obtidos com base na Tabela 2.2.

Tabela 2.4: Processo para obtenção da matriz de contrastes resultante da PBS de uma composição de 5 partes.

Etapa i da PBS	v_{i1}	v_{i2}	v_{i3}	v_{i4}	v_{i5}	$\mathbf{v}_i^{(\ell)}$
1	$\frac{2}{\sqrt{5}}$	$\frac{-1}{\sqrt{20}}$	$\frac{-1}{\sqrt{20}}$	$\frac{-1}{\sqrt{20}}$	$\frac{-1}{\sqrt{20}}$	$\mathbf{v}_1^{(\ell)}$
2	0	$\sqrt{\frac{3}{4}}$	$\frac{-1}{\sqrt{12}}$	$\frac{-1}{\sqrt{12}}$	$\frac{-1}{\sqrt{12}}$	$\mathbf{v}_2^{(\ell)}$
3	0	0	$\frac{2}{\sqrt{6}}$	$\frac{-1}{\sqrt{6}}$	$\frac{-1}{\sqrt{6}}$	$\mathbf{v}_3^{(\ell)}$
4	0	0	0	$\frac{1}{\sqrt{2}}$	$\frac{-1}{\sqrt{2}}$	$\mathbf{v}_4^{(\ell)}$

Assim, a base obtida por esta partição é dada à custa dos vetores $\mathbf{v}_i^{(\ell)}$ listados na Tabela 2.4, e usando (2.21):

$$\begin{aligned}\mathbf{e}_1 &= \mathcal{C} \left[\exp \left(\frac{2}{\sqrt{5}}, \frac{-1}{\sqrt{20}}, \frac{-1}{\sqrt{20}}, \frac{-1}{\sqrt{20}}, \frac{-1}{\sqrt{20}} \right) \right], \\ \mathbf{e}_2 &= \mathcal{C} \left[\exp \left(0, \sqrt{\frac{3}{4}}, \frac{-1}{\sqrt{12}}, \frac{-1}{\sqrt{12}}, \frac{-1}{\sqrt{12}} \right) \right], \\ \mathbf{e}_3 &= \mathcal{C} \left[\exp \left(0, 0, \frac{2}{\sqrt{6}}, \frac{-1}{\sqrt{6}}, \frac{-1}{\sqrt{6}} \right) \right], \\ \mathbf{e}_4 &= \mathcal{C} \left[\exp \left(0, 0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right) \right].\end{aligned}$$

Deste modo, a base ortonormal $\{\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_3, \mathbf{e}_4\}$ é a base obtida pela PBS associada a uma composição \mathbf{x} de 5 partes. Note-se que as composições com diferentes números de partes conduzem a diferentes bases. Assim, por exemplo, para o conjunto da habilitação académica referido no Capítulo 1, o qual contém composições de 10 partes, uma base ortonormal será constituída por 9 vetores ortogonais, $\{\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \mathbf{e}_6, \mathbf{e}_7, \mathbf{e}_8, \mathbf{e}_9\}$. Já para o conjunto da situação profissional constituído por composições formadas por 3 partes, a base terá apenas 2 vetores ortogonais, $\{\mathbf{e}_1, \mathbf{e}_2\}$.

Resumidamente, conclui-se que obter uma base ortonormal segundo a PBS importa obter uma matriz de contrastes. Os vetores da base ortonormal serão formados pelas coordenadas *clr*-transformadas que se obtêm das linhas da matriz de contrastes. Por sua vez, o número de vetores ortogonais será dado pelo número de etapas que a PBS contenha e corresponderá às $D - 1$ partes das composições.

Na matriz de contrastes, as linhas são obtidas segundo (2.10) que, colocando em evidência um fator de escala, equivale a ter

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}} \left(0, 0, \dots, 0, 1, -\frac{1}{D-i}, \dots, -\frac{1}{D-i} \right).$$

Pela Subsecção 2.5.1 averiguou-se que $z_i = z_i^{(\ell)}$, $i = 1, 2, \dots, D - 1$, pelo que às coordenadas pivô generalizadas também estão associadas a mesma base dada por $\mathbf{e}_i = clr^{-1}(\mathbf{v}_i)$ e obtida, também, para as coordenadas pivô não permutadas que originam as definições em (2.11) e (2.17).

Evidentemente que, se a finalidade fosse obter coordenadas pivô respeitantes a uma composição permutada, a base ortonormal teria de ser permutada e, por conseguinte, também a matriz \mathbf{V} conforme estabelecido em (2.18). Consequentemente, originar-se-ia novas coordenadas pivô definidas de acordo com (2.20).

2.5.3 Balances

Existem muitas possibilidades de escolher uma base ortonormal no simplex e construir as coordenadas ortonormais. Infelizmente, não há nenhuma base canónica no simplex (as D partes composicionais originais são representadas apenas por $D - 1$ novas coordenadas), de modo que são necessárias alternativas interpretáveis. Uma escolha possível representa o conceito de *balances* [22] (traduzido-se refere-se a equilíbrio), que permite uma interpretação das coordenadas ortonormais em termos do equilíbrio entre grupos de partes composicionais [24].

Em termos práticos, não é de todo apenas interessante interpretar a predominância relativa de partes, mas também o comportamento de grupos de partes composicionais dentro da composição. Suponha-se que os principais efeitos dentro de uma composição sejam causados por dois grupos de partes. Então, é de todo o interesse construir coordenadas que permitem uma interpretação dos dois grupos em termos de informação relativa. Estas coordenadas designadas por *balances*, referem-se ao equilíbrio entre os grupos. O procedimento para a construção destas coordenadas baseia-se na PBS. Tal como o nome indica, o interesse não reside somente no equilíbrio entre dois grupos, mas também a dominância relativa dos grupos é considerada de forma sequencial [19].

O conceito de *balances* surge no processo da PBS de uma dada composição. Tendo em conta o procedimento da PBS sumariado na Tabela 2.1, o equilíbrio entre os grupos de partes formados na i -ésima etapa da PBS de uma dada composição é dado por [19]:

$$\bar{z}_i = \sqrt{\frac{rs}{r+s}} \ln \frac{\left(\prod_{i=1}^r x_i^{(+)}\right)^{\frac{1}{r}}}{\left(\prod_{i=1}^s x_i^{(-)}\right)^{\frac{1}{s}}}, \quad (2.22)$$

onde $x_i^{(+)} \left(x_i^{(-)}\right)$ representa as composições com etiqueta $+$ ($-$), respetivamente. O número total de etapas necessárias é $D - 1$ e as coordenadas resultantes, $\bar{z}_1, \bar{z}_2, \dots, \bar{z}_{D-1}$, estão associadas a uma determinada base ortonormal em S^D [19].

A interpretação do equilíbrio entre grupos de partes é baseada na média geométrica como representante dos grupos presentes no numerador e denominador em (2.22). As médias geométricas são valores centrais das partes de cada grupo pelo que a razão entre médias (geométricas) indica o peso relativo de cada grupo. Por exemplo, um equilíbrio positivo indica que, em média (geométrica), o grupo de partes no numerador é dominante comparativamente ao grupo de partes no denominador, dado que este tem maior peso relativo na composição (o contrário se verifica para equilíbrios negativos) [6].

Exemplo 2.5.3.1. De maneira a exemplificar a interpretação de equilíbrio entre grupos, seja $\mathbf{x} = (0.031, 0.139, 0.337, 0.358, 0.135)$ a composição de 5 partes respeitante ao município de Anadia do conjunto do grupo etário. De acordo com (2.22) é necessário obter-se a PBS da composição \mathbf{x} e uma vez que \mathbf{x} contém 5 partes, no exemplo anterior, na Tabela 2.3, a PBS já se encontra obtida. Assim, na Tabela 2.5 tem-se a PBS e a expressão do equilíbrio entre grupos referente a cada etapa.

Tabela 2.5: PBS de uma composição de 5 partes e equilíbrio entre grupos de partes do município de Anadia.

Etapa i da PBS	$x_1^{(\ell)}$	$x_2^{(\ell)}$	$x_3^{(\ell)}$	$x_4^{(\ell)}$	$x_5^{(\ell)}$	r	s	Equilíbrio entre grupos
1	+1	-1	-1	-1	-1	1	4	$\bar{z}_1 = \sqrt{\frac{4}{5}} \ln \frac{x_1}{\sqrt[4]{x_2 x_3 x_4 x_5}}$
2	0	+1	-1	-1	-1	1	3	$\bar{z}_2 = \sqrt{\frac{3}{4}} \ln \frac{x_2}{\sqrt[3]{x_3 x_4 x_5}}$
3	0	0	+1	-1	-1	1	2	$\bar{z}_3 = \sqrt{\frac{2}{3}} \ln \frac{x_3}{\sqrt{x_4 x_5}}$
4	0	0	0	+1	-1	1	1	$\bar{z}_4 = \sqrt{\frac{1}{2}} \ln \frac{x_4}{x_5}$

Perante as expressões obtidas na Tabela 2.5, a composição de \mathbf{x} é transformada no vetor $\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4)$. Para a composição \mathbf{x} o vetor $\bar{\mathbf{z}}$ resulta em $\bar{\mathbf{z}} = (-1.745, -0.520, 0.349, 0.690)$ onde, de acordo com a construção, \bar{z}_1 descreve o equilíbrio entre os dois grupos de partes x_1 e x_2, x_3, x_4, x_5 ; \bar{z}_2 é o equilíbrio entre x_2 e x_3, x_4, x_5 ; \bar{z}_3 é o equilíbrio entre x_3 e x_4, x_5 ; e, por fim \bar{z}_4 descreve toda a informação relativa entre as partes x_4 e x_5 envolvendo apenas a razão logarítmica das partes.

Relativamente à interpretação de $\bar{\mathbf{z}}$ conclui-se que a variável *Idade 0-14* tem o menor peso relativo na composição; a variável *Idade 15-24* possui o menor peso relativo em relação às variáveis *Idade 25-39*, *Idade 40-64* e *Idade 65+*; a variável *Idade 25-39* tem maior peso relativo do que as variáveis *Idade 40-64* e *Idade 65+*; e a variável *Idade 40-64* tem maior peso relativo na composição do que a variável *Idade 65+*.

Sabe-se que as coordenadas pivô, definidas em (2.17), são construídas de tal forma que z_1 descreve toda a informação relativa sobre a parte de x_1 . Assim, \bar{z}_1 também pode ser visto como um equilíbrio entre os “grupos” x_1 e x_2, \dots, x_D . Por conseguinte, pode-se optar por referir, alternativamente, o termo de coordenadas pivô por *balances* dado por \bar{z}_1 , assim como para as restantes coordenadas, isto é, $\bar{z}_i \equiv z_i$.

2.6 Detecção de *outliers* multivariados

No início de uma análise estatística de um determinado conjunto de dados composicionais, geralmente foca-se o interesse em padrões da estrutura principal dos dados, por exemplo: grupos formados pelas observações ou relações entre as variáveis, bem como em desvios representados pelas observações atípicas [27].

Numa análise estatística, os *outliers* estão amplamente presentes nos conjuntos de dados reais [28]. Estas observações atípicas podem, muitas das vezes, danificar os estimadores clássicos, por isso é usual considerar que essas mesmas observações devem ser retiradas dos dados

e feita a análise sem qualquer influência destas observações. Todavia, os *outliers* são frequentemente as observações mais interessantes porque algum fenómeno atípico é responsável pela sua presença, o que ocorre é que eles são unicamente desvalorizados para obter um ajuste de modelo que acomode a maioria dos dados [19].

Os dados composicionais são observações multivariadas, consequentemente numa deteção de *outliers* em dados composicionais assume-se que os dados multivariados são composições e, por isso, foca-se o interesse em *outliers* multivariados. Na identificação das observações atípicas, em vez de se identificar os *outliers* diretamente no espaço original, é comum expressar as composições em coordenadas log-razões e, em seguida, aplicar os métodos usuais de deteção de *outliers* multivariados.

Os *outliers* multivariados não são necessariamente extremos, mas podem estar localizados em qualquer parte do espaço multivariado. Para detectar estas observações atípicas, é importante dispor de métodos capazes de revelar observações divergentes, independentemente da escolha do tipo de transformação [19].

O foco da subsecção seguinte é expor ferramentas úteis na deteção de *outliers* multivariados. Abordar-se-á dois procedimentos para identificar as observações atípicas, mas o objetivo principal centra-se em analisar ao pormenor um desses procedimentos.

2.6.1 Métodos para deteção de *outliers* multivariados

No estudo exploratório da análise de dados composicionais, a deteção de *outliers* deve ser um passo importante da análise. Os *outliers* identificados são candidatos a dados atípicos que contribuem para erros na especificação do modelo, estimativa dos parâmetros e resultados incorretos [27]. Por isso, é importante identificar essas observações atípicas antes da análise dos dados.

Existem dois procedimentos diferentes para identificar *outliers* multivariados: (1) métodos baseados na projeção e, (2) métodos baseados na estimativa da estrutura de covariância da matriz de dados. A ideia dos métodos em (1) consiste em projetar repetidamente os dados multivariados no espaço univariado uma vez que a deteção univariada de *outliers* torna-se muito mais simples (para mais detalhe recomendo as seguintes referências: Gnanadesikan *et al* 1972 ([29]), Peña *et al* 2001 ([30]) e Maronna *et al* 2002 ([31])). Apesar destes métodos, geralmente, serem computacionalmente intensivos são particularmente úteis para dados de elevada dimensionalidade e reduzido tamanho da amostra [32].

Para o procedimento (2), a deteção de *outliers* multivariados baseia-se na estimativa da estrutura de covariância. O que se pretende é identificar observações que se desviam de uma distribuição do modelo subjacente. Neste caso, apenas a distribuição normal multivariada é considerada, o que serve como uma importante distribuição de modelo para muitos métodos estatísticos [19].

Neste procedimento, a estimativa da estrutura de covariância é usada para atribuir uma distância a cada observação, indicando até que ponto a observação se encontra do centro da nuvem de dados (também designado por centróide) em relação à estrutura de covariância. Esta distância é calculada com base na métrica de Mahalanobis. Sem perda de generalidade, assumindo que uma amostra $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ de observações $(D - 1)$ -dimensionais contém dados

composicionais em coordenadas *ilr*, a distância de Mahalanobis é definida por:

$$\text{MD}(\mathbf{z}_l) = \sqrt{(\mathbf{z}_l - \mathbf{T})' \mathbf{C}^{-1} (\mathbf{z}_l - \mathbf{T})}, \quad l = 1, 2, \dots, n \quad (2.23)$$

onde o vetor \mathbf{T} de dimensão $(D-1)$ e a matriz \mathbf{C} de dimensão $(D-1) \times (D-1)$ representam as estimativas de localização e de covariância, respetivamente [14, 33, 21].

A propriedade de equivariância afim para as estimativas \mathbf{T} e \mathbf{C} , enunciada a seguir, é de extrema importância na distância de Mahalanobis.

Propriedade 2.6.1.1. *As estimativas \mathbf{T} e \mathbf{C} de localização e de covariância, respetivamente, são designadas por equivariantes afim, se para qualquer matriz não singular \mathbf{A} de dimensão $d \times d$ e para qualquer vetor $\mathbf{b} \in \mathbb{R}^d$ as seguintes condições são satisfeitas:*

$$\begin{aligned} \mathbf{T}(\mathbf{A}\mathbf{z}_1 + \mathbf{b}, \dots, \mathbf{A}\mathbf{z}_n + \mathbf{b}) &= \mathbf{A}\mathbf{T}(\mathbf{z}_1, \dots, \mathbf{z}_n) + \mathbf{b} \\ \mathbf{C}(\mathbf{A}\mathbf{z}_1 + \mathbf{b}, \dots, \mathbf{A}\mathbf{z}_n + \mathbf{b}) &= \mathbf{A}\mathbf{C}(\mathbf{z}_1, \dots, \mathbf{z}_n)\mathbf{A}' \end{aligned}$$

onde $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ são n composições de D partes.

Assim, estimativas que cumpram estas condições e, posteriormente, utilizadas na distância de Mahalanobis fazem com que as distâncias permaneçam inalteradas sob transformações afins regulares, isto é,

$$\text{MD}(\mathbf{A}\mathbf{z}_l + \mathbf{b}) = \text{MD}(\mathbf{z}_l), \quad l = 1, 2, \dots, n. \quad (2.24)$$

De facto, partindo de (2.23) verifica-se a igualdade (2.24):

$$\begin{aligned} \text{MD}(\mathbf{A}\mathbf{z}_l + \mathbf{b}) &= \sqrt{[\mathbf{A}\mathbf{z}_l + \mathbf{b} - \mathbf{T}(\mathbf{A}\mathbf{z}_l + \mathbf{b})]' \mathbf{C}(\mathbf{A}\mathbf{z}_l + \mathbf{b})^{-1} [\mathbf{A}\mathbf{z}_l + \mathbf{b} - \mathbf{T}(\mathbf{A}\mathbf{z}_l + \mathbf{b})]} \\ &= \sqrt{[\mathbf{A}\mathbf{z}_l + \mathbf{b} - (\mathbf{A}\mathbf{T}(\mathbf{z}_1, \dots, \mathbf{z}_n) + \mathbf{b})]' (\mathbf{A}\mathbf{C}(\mathbf{z}_1, \dots, \mathbf{z}_n)\mathbf{A}')^{-1} [\mathbf{A}\mathbf{z}_l + \mathbf{b} - (\mathbf{A}\mathbf{T}(\mathbf{z}_1, \dots, \mathbf{z}_n) + \mathbf{b})]} \\ &= \sqrt{[\mathbf{z}_l - \mathbf{T}(\mathbf{z}_1, \dots, \mathbf{z}_n)]' \mathbf{A}' [(\mathbf{A}')^{-1} \mathbf{C}(\mathbf{z}_1, \dots, \mathbf{z}_n)^{-1} \mathbf{A}^{-1}] \mathbf{A} [\mathbf{z}_l - \mathbf{T}(\mathbf{z}_1, \dots, \mathbf{z}_n)]} \\ &= \sqrt{(\mathbf{z}_l - \mathbf{T}(\mathbf{z}_l))' \mathbf{C}(\mathbf{z}_l)^{-1} (\mathbf{z}_l - \mathbf{T}(\mathbf{z}_l))} \\ &= \text{MD}(\mathbf{z}_l) \end{aligned}$$

A Propriedade 2.6.1.1 garante que os *outliers* identificados sejam invariantes com a escolha da transformação log-razão, independentemente da escolha de \mathbf{A} e \mathbf{b} [33] (na Secção 2.7 será discutida em mais detalhe esta questão).

A escolha das estimativas \mathbf{T} e \mathbf{C} é um passo crucial para garantir a qualidade da deteção de observações atípicas multivariadas [33]. No caso da distribuição normal multivariada, escolhe-se para as estimativas de localização e de covariância (populacional), a média aritmética e a matriz de covariância da amostra, respetivamente, tornando esta escolha a mais adequada para assegurar uma melhor eficiência estatística [32]. Para uma amostra $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ de n composições em coordenadas *ilr*-transformadas, a média aritmética e a matriz de covariância da amostra são dadas por, respetivamente:

$$\bar{\mathbf{z}} = \frac{1}{n} \sum_{l=1}^n \mathbf{z}_l \quad \text{e} \quad \mathbf{S}_Z = \frac{1}{n-1} \sum_{l=1}^n (\mathbf{z}_l - \bar{\mathbf{z}})(\mathbf{z}_l - \bar{\mathbf{z}})' \quad (2.25)$$

As próprias definições da média aritmética e da matriz de covariância revelam que estas estimativas clássicas partilham a propriedade da equivariância afim. De facto, considere-se uma amostra $Z = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\}$ e seja $(\hat{\boldsymbol{\mu}}(Z), \hat{\boldsymbol{\Sigma}}(Z)) \equiv (\bar{\mathbf{z}}, \mathbf{S}_Z)$ as estimativas clássicas dadas pela média aritmética e matriz de covariância amostral (fórmula em (2.25)). Estas estimativas são equivariantes afim uma vez que satisfazem a relação:

$$\hat{\boldsymbol{\mu}}(\mathbf{A}Z + \mathbf{b}) = \mathbf{A}\hat{\boldsymbol{\mu}}(Z) + \mathbf{b} \quad (2.26)$$

$$\hat{\boldsymbol{\Sigma}}(\mathbf{A}Z + \mathbf{b}) = \mathbf{A}\hat{\boldsymbol{\Sigma}}(Z)\mathbf{A}' \quad (2.27)$$

No entanto, quando se considera estas duas estimativas clássicas muitas vezes esta escolha leva a resultados inúteis uma vez que estas são influenciadas pelas observações atípicas [33]. Deste modo, para contornar esta questão foram desenvolvidas metodologias estatísticas robustas que reduzem a influência dos *outliers*, concentrando-se na estrutura dos dados principais. Portanto, a escolha das estimativas deve incidir em estimativas robustas. Assim, a distância de Mahalanobis em (2.23) é reformulada do seguinte modo:

$$\text{MD}(\mathbf{z}_l) = \sqrt{(\mathbf{z}_l - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{z}_l - \hat{\boldsymbol{\mu}})}, \quad l = 1, 2, \dots, n$$

onde $\hat{\boldsymbol{\mu}}$ e $\hat{\boldsymbol{\Sigma}}$ correspondem, respetivamente, às estimativas robustas da média e da matriz de covariância, garantindo que as distâncias utilizadas sejam robustas. Na Subsecção 2.6.2, explicar-se-á um método de obter estas estimativas robustas.

Dada uma matriz de dados composicionais $\mathbf{X}_{n \times D}$ considere-se, sem perda de generalidade, que a matriz $\mathbf{Z}_{n \times (D-1)}$ resulta da aplicação da transformação *ilr* à matriz $\mathbf{X}_{n \times D}$.

Definição 2.6.1.1 (Distribuição Normal de dados composicionais). *Seja $\mathbf{z} = (z_1, \dots, z_{D-1})$ as coordenadas ilr-transformadas de uma composição $\mathbf{x} \in S^D$. Diz-se que um conjunto de dados composicionais $\mathbf{X}_{n \times D}$ tem distribuição normal no simplex se $\mathbf{Z}_{n \times (D-1)} = \text{ilr}(\mathbf{X})$ tem distribuição normal multivariada em \mathbb{R}^{D-1} , denotando-se por $\mathbf{Z} \sim N(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Z}})$ e $\mathbf{X} \sim N_{SD}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Os procedimentos da deteção de *outliers* multivariados assumem que a maioria das observações da matriz \mathbf{Z} são geradas por uma distribuição normal multivariada. Pela Definição 2.6.1.1, se $\mathbf{X}_{n \times D}$ segue uma distribuição normal no simplex, $\mathbf{Z}_{n \times (D-1)}$ tem distribuição normal multivariada em \mathbb{R}^{D-1} , $\mathbf{Z} \sim N(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Z}})$ [6].

Os dados composicionais necessitam de ser transformados antes de se calcular as distâncias de Mahalanobis. Por razões que serão discutidas na Secção 2.7, a deteção de observações atípicas deve incidir no uso da transformação *ilr*. Por conseguinte, a distância de Mahalanobis baseia-se na matriz $\mathbf{Z}_{n \times (D-1)}$ contendo as coordenadas *ilr*. Perante uma distribuição normal multivariada, proveniente de $\mathbf{Z}_{n \times (D-1)}$, a distância de Mahalanobis torna-se robusta influenciada por estimativas robustas de $\boldsymbol{\mu}_{\mathbf{Z}}$ e $\boldsymbol{\Sigma}_{\mathbf{Z}}$. Sob o pressuposto da normalidade multivariada dos dados no simplex, a Propriedade 2.6.1.2 contribui para concluir que a distância de Mahalanobis, $\text{MD}(\mathbf{z}_l)$, $l = 1, 2, \dots, n$, segue uma distribuição de qui-quadrado com $(D-1)$ graus de liberdade, onde $(D-1)$ corresponde ao número de variáveis em \mathbf{Z} . Deste modo, se $\mathbf{Z} \sim N(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Z}})$ então $\text{MD}(\mathbf{Z}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \sim \chi_{D-1}^2$ tendo em conta o seguinte resultado:

Propriedade 2.6.1.2. *Se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $|\boldsymbol{\Sigma}| > 0$, então*

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_p^2$$

onde χ_p^2 denota a distribuição de qui-quadrado com p graus de liberdade.

O gráfico associado a quadrados de distâncias de Mahalanobis constantes, $MD(\mathbf{z}_l)^2 = \text{constante}$, corresponde a elipsóides. Essa constante permitirá a identificação dos *outliers*. Da Propriedade 2.6.1.2 essa constante corresponde a um quantil χ^2_{D-1} . Assim, em χ^2_{D-1} surge a questão de qual a ordem do quantil a considerar para detetar observações atípicas. O valor do quantil correspondente a essa ordem irá servir como valor de corte para separar as observações “regulares” daquelas que constituem potenciais *outliers* [33].

Existem várias opções para a ordem a tomar, na literatura, o quantil de ordem 0.975 é frequentemente usado como indicador do valor de corte para detetar *outliers*. Em síntese, considera-se que observações que possuem um valor do quadrado da distância de Mahalanobis superior ao valor de corte são consideradas observações atípicas [32], isto é,

$$MD(\mathbf{z}_l)^2 > \chi^2_{D-1;0.975}.$$

Na Figura 2.1 encontram-se as duas primeiras coordenadas *ilr*-transformadas de uma composição. Visualizam-se quatro elipses relativas aos quadrados de distâncias de Mahalanobis robusta constantes e iguais aos quantis de ordem 0.25, 0.50, 0.75 e 0.975. A última elipse, correspondente ao quantil de ordem 0.975, determina as observações *outliers* exibidas a vermelho e etiquetadas com símbolo “+” de maior tamanho [33].

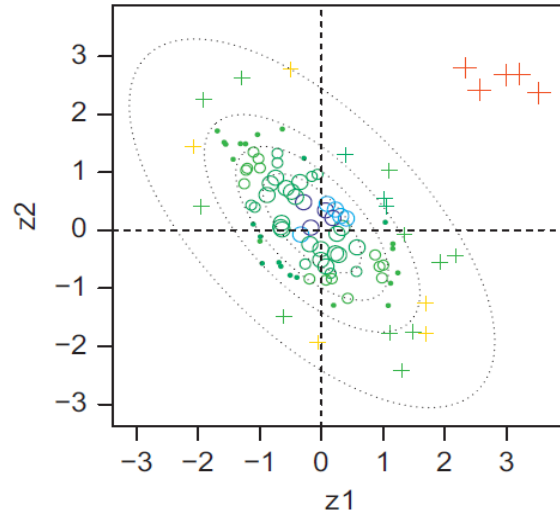


Figura 2.1: Representação das quatro elipses referentes aos quantis de ordem 0.25, 0.5 e 0.75 e ao valor de corte para detetar *outliers*, quantil de ordem 0.975 (Figura extraída de [33]).

É de realçar que na Figura 2.1 existem mais quatro grupos diferentes de símbolos. Os símbolos referenciados por “+” de menor tamanho encontram-se entre os quantis de ordem 0.75 e 0.975 não sendo considerados observações atípicas. Em oposição, no interior da elipse, encontram-se os símbolos em forma de círculo (“o”). Os círculos preenchidos (pontos) localizam-se entre os quantis de ordem 0.5 e 0.75 e os círculos que não estão preenchidos, tanto de menor como de maior tamanho, situam-se desde o centro da nuvem de dados até ao quantil de ordem 0.5 (ultrapassando o quantil de ordem 0.25) [33].

Observe-se que há também outras propostas na literatura para encontrar um valor de corte apropriado. Para uma abordagem mais pormenorizada aconselho a leitura da referência [34].

2.6.2 Estimador MCD

Para obter as estimativas robustas de quaisquer parâmetros populacionais, pode-se recorrer ao estimador MCD (*Minimum Covariance Determinant*) ou aos estimadores S [14]. Geralmente, opta-se por obter as estimativas robustas pelo estimador MCD uma vez que este apresenta a vantagem de ser um estimador eficiente e assintoticamente normal. Tornou-se bastante conhecido devido às suas boas propriedades de robustez e a um algoritmo rápido para o seu cálculo proposto por Rousseeuw e Van Driessen, em 1999, designado por *Fast-MCD* [35].

O estimador MCD caracteriza-se pela determinação de um subconjunto de pelo menos h observações cuja matriz de covariância amostral, \mathbf{S} , tenha o menor determinante. Assim, as estimativas robustas dos parâmetros populacionais $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são escolhidos, respetivamente, como a média aritmética e a matriz de covariância amostral deste subconjunto, multiplicados por um fator para garantir a consistência dos estimadores sob o pressuposto da normalidade dos dados [35].

A escolha de h determina tanto a robustez como a eficiência estatística dos estimadores, e tem que englobar a maioria dos dados podendo ser escolhido como um valor inteiro no intervalo $[\frac{n+D+1}{2}, n]$, onde D é o número de partes do conjunto. O valor de h deve ser pelo menos metade do tamanho total da amostra n , resultando numa melhor resistência às observações externas mas com uma eficiência estatística menor. Por outro lado, se h assumir um valor grande, por exemplo, próximo de n , a robustez das estimativas obtidas por MCD é fraca, mas a eficiência aumenta. A literatura sugere que a melhor escolha deverá ser, aproximadamente, $h = \frac{3}{4}n$. Esta escolha tolera uma fração de *outliers* de cerca de $\frac{n-h}{n} = \frac{1}{4}$ das observações. Assim, é necessário que $\frac{n-h}{n}$ seja maior que essa fração de observações atípicas, pois, caso contrário, as estimativas podem-se tornar pouco confiáveis [11].

A equivariância afim é uma propriedade desejável de qualquer estimativa em situações em que é fundamental que o resultado permaneça essencialmente inalterado sob quaisquer transformações lineares não singulares [14]. Na Secção 2.6.1 constatou-se que as estimativas clássicas, $\hat{\boldsymbol{\mu}}$ e $\hat{\boldsymbol{\Sigma}}$, satisfaziam essa propriedade (em (2.26) e (2.27)) e, consequentemente, o uso desta propriedade garante que as distâncias de Mahalanobis não sofram alterações sob qualquer transformação log-razão que se aplique (provado na Secção 2.7).

Porém, na identificação de observações atípicas a distância de Mahalanobis deve ser robusta e ao considerar estimativas robustas deve-se ter o cuidado de garantir que o estimador utilizado para estas estimativas consiga satisfazer a propriedade da equivariância afim. Ora, o estimador MCD é por si só um estimador que cumpre esta propriedade [33]. Portanto, para além das suas vantagens é de todo o interesse optar por este estimador.

Na secção seguinte serão apresentadas as propriedades das transformações log-razões onde, pormenorizadamente, se explica as relações entre as transformações, a distância de Mahalanobis e a propriedade da equivariância afim.

2.7 Propriedades das transformações na detecção de observações atípicas

A utilidade das distâncias robustas de Mahalanobis para a detecção de *outliers* multivariados tem sido demonstrada na literatura e tem muitas aplicações. Esta ferramenta não seria apropriada para dados fechados, mas apenas para dados após a transformação. O problema surge em identificar qual a transformação de log-razões do simplex mais adequada para o espaço real [32].

Respostas para cada uma das transformações apresentadas na Secção 2.4 e, ainda, questões cruciais abordadas na Secção 2.6.1 e 2.6.2 serão esclarecidas de seguida. Embora as demonstrações se encontram em [32], o estudo destas tornou-se útil para uma melhor compreensão das propriedades pelo que as demonstrações são aqui apresentadas com mais detalhe.

Teorema 2.7.1. *As distâncias de Mahalanobis para dados transformados por alr são invariantes em relação à escolha da variável razão se a estimativa de localização \mathbf{T} e a estimativa de dispersão \mathbf{C} forem equivariantes afim.*

Demonstração. Seja $\mathbf{X}_{n,D}$ uma matriz de dados composicionais. Portanto, tendo em conta que $\mathbf{x}_l = (x_{l1}, x_{l2}, \dots, x_{lD})$ tem-se $\sum_{d=1}^D x_{ld} = 1$ e $x_{ld} > 0$, para $l = 1, 2, \dots, n$, isto é, $\mathbf{x}_l \in S^D$. Seja $\mathbf{X}_{n,D-1}^{(j)}$ uma matriz que resulta da transformação *alr* de \mathbf{X} usando a coluna j . As linhas de $\mathbf{X}^{(j)}$ são da forma

$$\mathbf{x}_l^{(j)} = \left(\log \frac{x_{l1}}{x_{lj}}, \log \frac{x_{l2}}{x_{lj}}, \dots, \log \frac{x_{l,j-1}}{x_{lj}}, \log \frac{x_{l,j+1}}{x_{lj}}, \dots, \log \frac{x_{lD}}{x_{lj}} \right), \quad l = 1, 2, \dots, n. \quad (2.28)$$

Do mesmo modo, seja $\mathbf{X}^{(k)}$ uma matriz de dados que resulta da transformação *alr* de \mathbf{X} usando a coluna k , com $k \neq j$. Então, pelas propriedades dos logaritmos, é simples de mostrar que $\mathbf{X}^{(j)} = \mathbf{X}^{(k)} \mathbf{B}_{kj}$ ou $\mathbf{x}_l^{(j)} = \mathbf{B}_{kj}' \mathbf{x}_l^{(k)}$ com a matriz \mathbf{B}_{kj} de dimensão $(D-1) \times (D-1)$ definida do seguinte modo:

$$\mathbf{B}_{kj} = \begin{bmatrix} 1 & & & & & & 0 & & \\ & \ddots & & & & & \vdots & & \\ & & 1 & & & & 0 & & \\ -1 & \dots & -1 & -1 & -1 & \dots & -1 & \dots & -1 \\ & & & 1 & 0 & & 0 & & \\ & & & & \ddots & \ddots & \vdots & & \\ & & & & & 1 & 0 & & \\ & & & & & & 0 & 1 & \\ & & & & & & \vdots & & \ddots \\ & & & & & & 0 & & 1 \end{bmatrix}.$$

Esta matriz coincide com a matriz identidade exceto que na j -ésima linha da matriz inclui apenas entradas de -1 , a diagonal principal entre a linha $j+1$ e a linha $k-1$ é completada com entradas iguais a zero e, por final, a diagonal paralela a esta imediatamente anterior passa a ter entradas iguais a 1. Note-se que as entradas em branco na matriz são zeros. De forma a melhor ilustrar esta matriz \mathbf{B}_{kj} e a condição $\mathbf{X}^{(j)} = \mathbf{X}^{(k)} \mathbf{B}_{kj}$, suponha-se que $D = 5$. Neste

caso, a matriz $\mathbf{X}_{n,5}$ com n observações é

$$\mathbf{X}_{n,5} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} \end{bmatrix}.$$

A matriz que resulta da transformação *alr* aplicada à matriz $\mathbf{X}_{n,5}$ utilizando a coluna 2 ($j = 2$) é representada por:

$$\mathbf{X}_{n,4}^{(2)} = \begin{bmatrix} \log x_{11} - \log x_{12} & \log x_{13} - \log x_{12} & \log x_{14} - \log x_{12} & \log x_{15} - \log x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ \log x_{n1} - \log x_{n2} & \log x_{n3} - \log x_{n2} & \log x_{n4} - \log x_{n2} & \log x_{n5} - \log x_{n2} \end{bmatrix}.$$

Agora, considere-se novamente uma matriz que resulta da transformação *alr* aplicada à matriz $\mathbf{X}_{n,5}$ utilizando a coluna 5 ($k = 5$),

$$\mathbf{X}_{n,4}^{(5)} = \begin{bmatrix} \log x_{11} - \log x_{15} & \log x_{12} - \log x_{15} & \log x_{13} - \log x_{15} & \log x_{14} - \log x_{15} \\ \vdots & \vdots & \vdots & \vdots \\ \log x_{n1} - \log x_{n5} & \log x_{n3} - \log x_{n5} & \log x_{n3} - \log x_{n5} & \log x_{n4} - \log x_{n5} \end{bmatrix}.$$

Ora, para $D = 5$, a matriz $\mathbf{B}_{5,2}$ será

$$\mathbf{B}_{5,2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

De facto, a condição $\mathbf{X}^{(2)} = \mathbf{X}^{(5)}\mathbf{B}_{52}$ verifica-se.

A matriz quadrada \mathbf{B}_{kj} é não singular. Após cálculos algébricos consegue-se provar que o determinante é não nulo, concretamente que $\det(\mathbf{B}_{kj}) = (-1)^{k-j}$. Assim, a matriz admite inversa, \mathbf{B}_{kj}^{-1} e, por conseguinte, a matriz transposta de \mathbf{B}_{kj} também admite inversa, $(\mathbf{B}'_{kj})^{-1}$. Então, para \mathbf{T} e \mathbf{C} estimadores equivariantes afim (Propriedade 2.6.1.1), obtém-se:

$$\begin{aligned} \mathbf{T}(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)}) &= \mathbf{T}(\mathbf{B}'_{kj}\mathbf{x}_1^{(k)}, \mathbf{B}'_{kj}\mathbf{x}_2^{(k)}, \dots, \mathbf{B}'_{kl}\mathbf{x}_n^{(k)}) \\ &= \mathbf{B}'_{kj}\mathbf{T}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)}) \end{aligned}$$

e,

$$\begin{aligned} \mathbf{C}(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)}) &= \mathbf{C}(\mathbf{B}'_{kj}\mathbf{x}_1^{(k)}, \mathbf{B}'_{kj}\mathbf{x}_2^{(k)}, \dots, \mathbf{B}'_{kj}\mathbf{x}_n^{(k)}) \\ &= \mathbf{B}'_{kj}\mathbf{C}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)})\mathbf{B}_{kl} \end{aligned}$$

Consequentemente,

$$\begin{aligned}
 & \text{MD}^2(\mathbf{x}_l^{(j)}) \\
 &= \left[\mathbf{x}_l^{(j)} - \mathbf{T}(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)}) \right]' \left[\mathbf{C}(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)}) \right]^{-1} \left[\mathbf{x}_l^{(j)} - \mathbf{T}(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)}) \right] \\
 &= \left[\mathbf{B}'_{kj} \mathbf{x}_l^{(k)} - \mathbf{B}'_{kj} \mathbf{T}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)}) \right]' \left[\mathbf{B}'_{kj} \mathbf{C}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)}) \mathbf{B}_{kj} \right]^{-1} \times \\
 &\quad \times \left[\mathbf{B}'_{kl} \mathbf{x}_l^{(k)} - \mathbf{B}'_{kl} \mathbf{T}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)}) \right] \\
 &= \left[\mathbf{x}_l^{(k)} - \mathbf{T}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)}) \right]' \mathbf{B}_{kj} \mathbf{B}_{kl}^{-1} \left[\mathbf{C}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)}) \right]^{-1} (\mathbf{B}'_{kj})^{-1} \mathbf{B}'_{kj} \times \\
 &\quad \times \left[\mathbf{x}_l^{(k)} - \mathbf{T}(\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \dots, \mathbf{x}_n^{(k)}) \right] \\
 &= \text{MD}^2(\mathbf{x}_l^{(k)})
 \end{aligned}$$

donde se conclui o pretendido. \square

O Teorema 2.7.1 garante que os *outliers* identificados não dependerão da variável razão escolhida para a transformação *alr*, desde que as estimativas \mathbf{T} e \mathbf{C} sejam consideradas como equivariantes afim.

Observação 2.7.1. Em Álgebra Linear, a pseudoinversa A^+ de uma matriz A é uma generalização da matriz inversa de A . Em 1956, Penrose estabeleceu um resultado que caracteriza a pseudoinversa de uma matriz, também conhecida como inversa Moore-Penrose. Esse resultado é enunciado de seguida sob a forma de Lema.

Lema 2.7.1. Seja $A \in \mathbb{R}_r^{m \times n}$. Então, $G = A^+$ se e só se

$$(P1) \quad AGA = A$$

$$(P2) \quad GAG = G$$

$$(P3) \quad (AG)' = AG$$

$$(P4) \quad (GA)' = GA$$

Além disso, A^+ existe e é única.

Teorema 2.7.2. As distâncias de Mahalanobis para dados transformados por *clr* e *alr* são as mesmas se a estimativa de localização \mathbf{T} for a média aritmética e a estimativa de covariância \mathbf{C} for a matriz de covariância da amostra.

Atendendo ao resultado explícito na Observação 2.7.1 tem-se todas as condições para demonstrar o Teorema 2.7.2.

Demonstração. Seja $\mathbf{x} = (x_1, x_2, \dots, x_D) \in S^D$, com $\sum_{d=1}^D x_d = 1$ e $x_d > 0$, uma composição. Primeiramente, obtém-se uma relação em termos matricial entre as transformações *alr* e *clr* de \mathbf{x} . Sem perda de generalidade, suponha-se que a última variável D é a parte de referência usada para a transformação *alr*. Dada a composição \mathbf{X} recorde-se que a transformação de *alr* de \mathbf{X} é dada por:

$$\mathbf{x}^{(D)} = (\log x_1 - \log x_D, \log x_2 - \log x_D, \dots, \log x_{D-1} - \log x_D)$$

e a transformação clr , $\mathbf{y} = (y_1, y_2, \dots, y_D)$, é dada por:

$$y_d = \log x_d - \frac{1}{D} \sum_{j=1}^D \log x_j = \frac{D-1}{D} \log x_d - \frac{1}{D} \sum_{j=1, j \neq d}^D \log x_j, \quad d = 1, 2, \dots, D.$$

Através de manipulação algébrica verifica-se que uma relação entre as duas transformações pode ser descrita através das seguintes igualdades: $\mathbf{x}^{(D)} = \mathbf{F}\mathbf{y}$ e $\mathbf{y} = \mathbf{F}^*\mathbf{x}^{(D)}$, onde

$$\mathbf{F}_{D-1,D} = \begin{bmatrix} 1 & \dots & 0 & -1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1 & -1 \end{bmatrix} \quad \text{e} \quad \mathbf{F}^*_{D,D-1} = \begin{bmatrix} \frac{D-1}{D} & -\frac{1}{D} & \dots & -\frac{1}{D} \\ -\frac{1}{D} & \frac{D-1}{D} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{D} \\ \vdots & & \ddots & \frac{D-1}{D} \\ -\frac{1}{D} & \dots & \dots & -\frac{1}{D} \end{bmatrix}$$

De facto, relativamente à primeira igualdade observe-se que $\mathbf{F}\mathbf{y} = (y_1 - y_D, y_2 - y_D, \dots, y_{D-1} - y_D)$ e qualquer que seja $d = 1, 2, \dots, D-1$ se tem que

$$\begin{aligned} y_d - y_D &= \frac{D-1}{D} \log x_d - \frac{1}{D} \sum_{j=1, j \neq d}^D \log x_j - \frac{D-1}{D} \log x_D + \frac{1}{D} \sum_{j=d, j \neq D}^D \log x_j \\ &= \log x_d - \frac{1}{D} \log x_d - \frac{1}{D} \sum_{j=1, j \neq d}^D \log x_j - \log x_D + \frac{1}{D} \log x_D + \frac{1}{D} \sum_{j=1, j \neq D}^D \log x_j \\ &= \log x_d - \log x_D - \frac{1}{D} \sum_{j=1}^D \log x_j + \frac{1}{D} \sum_{j=1}^D \log x_j \\ &= \log x_d - \log x_D, \end{aligned}$$

Logo $\mathbf{F}\mathbf{y} = \mathbf{x}^{(D)}$.

Relativamente à segunda igualdade, observe-se que $\mathbf{F}^*\mathbf{x}^{(D)}$ representa um vetor de dimensão $D-1$, sendo a sua primeira componente dada por:

$$\begin{aligned} &\frac{D-1}{D}(\log x_1 - \log x_D) - \frac{1}{D}(\log x_2 - \log x_D) - \dots - \frac{1}{D}(\log x_{D-1} - \log x_D) \\ &= \frac{D-1}{D} \log x_1 - \frac{D-1}{D} \log x_D - \frac{1}{D} \log x_2 + \frac{1}{D} \log x_D - \dots - \frac{1}{D} \log x_{D-1} + \frac{1}{D} \log x_D \\ &= \frac{D-1}{D} \log x_1 - \frac{1}{D} \sum_{d=2}^{D-1} \log x_d - \frac{D-1}{D} \log x_D + \frac{1}{D} \log x_D \times (D-2) \\ &= \frac{D-1}{D} \log x_1 - \frac{1}{D} \sum_{d=2}^{D-1} \log x_d - \frac{1}{D} \log x_D \\ &= \frac{D-1}{D} \log x_1 - \frac{1}{D} \sum_{d=2}^D \log x_d \\ &= y_1 \end{aligned}$$

Do mesmo modo, se obtém os restantes resultados para as demais componentes de \mathbf{y} , o que se conclui que $\mathbf{F}^*\mathbf{x}^{(D)} = \mathbf{y}$.

Tendo em conta as definições de \mathbf{F} e \mathbf{F}^* , ocorre que relativamente à matriz \mathbf{F} , \mathbf{F}^* satisfaz as propriedades de (P1) a (P4) do Lema 2.7.1. Na verdade, $\mathbf{F}\mathbf{F}^* = \mathbf{I}_{D-1} = (\mathbf{F}\mathbf{F}^*)'$ (prop. P3) (onde \mathbf{I}_{D-1} é a matriz identidade de ordem $D-1$), $\mathbf{F}^*\mathbf{F}$ é simétrica, e portanto, $(\mathbf{F}^*\mathbf{F})' = \mathbf{F}^*\mathbf{F}$ (prop. P4), $\mathbf{F}\mathbf{F}^*\mathbf{F} = \mathbf{F}$, e $\mathbf{F}^*\mathbf{F}\mathbf{F}^* = \mathbf{F}^*$. Logo, \mathbf{F}^* é a pseudoinversa de \mathbf{F} . Assim, \mathbf{F}^* será a partir de agora denotada por \mathbf{F}^+ . Note-se que conclusões análogas podem ser retiradas se, porventura, se considerar para a transformação *alr* outra variável de referência para a razão, mas as estruturas das matrizes serão diferentes pois também \mathbf{F} será diferente.

Considere-se, agora, as matrizes transformadas por *alr* e *clr* denotadas por $\mathbf{X}_{n,D-1}^{(D)}$ e $\mathbf{Y}_{n,D}$ com linhas $\mathbf{x}_l^{(D)}$ e \mathbf{y}_l , para $l = 1, 2, \dots, n$, respetivamente. Usar-se-á a notação $\bar{\mathbf{x}}^{(D)}$ e $\bar{\mathbf{y}}$ para denotar os vetores de média (em coluna) dessas observações, respetivamente, e, $\mathbf{S}_{\mathbf{x}^{(D)}}$ e $\mathbf{S}_{\mathbf{y}}$ para as matrizes de covariância da amostra. Relativamente às matrizes de covariância verifica-se a seguinte relação entre elas:

$$\begin{aligned} \mathbf{S}_{\mathbf{y}} &= \frac{1}{n} \sum_{l=1}^n (\mathbf{y}_l - \bar{\mathbf{y}})(\mathbf{y}_l - \bar{\mathbf{y}})' \\ &= \frac{1}{n} \sum_{l=1}^n \left(\mathbf{F}^+ \mathbf{x}_l^{(D)} - \mathbf{F}^+ \bar{\mathbf{x}}^{(D)} \right) \left(\mathbf{F}^+ \mathbf{x}_l^{(D)} - \mathbf{F}^+ \bar{\mathbf{x}}^{(D)} \right)' \\ &= \mathbf{F}^+ \frac{1}{n} \sum_{l=1}^n \left(\mathbf{x}_l^{(D)} - \bar{\mathbf{x}}^{(D)} \right) \left(\mathbf{x}_l^{(D)} - \bar{\mathbf{x}}^{(D)} \right)' (\mathbf{F}^+)' \\ &= \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' . \end{aligned}$$

Seja $\mathbf{S}_{\mathbf{y}}^* = \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F}$. Recordando ainda que $\mathbf{F}\mathbf{F}^+ = \mathbf{I}$ e as propriedades da inversa generalizada de Moore-Ponrose do Lema 2.7.1, retiram-se as seguintes relações:

$$\begin{aligned} \mathbf{S}_{\mathbf{y}} \mathbf{S}_{\mathbf{y}}^* \mathbf{S}_{\mathbf{y}} &= \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \\ &= \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}\mathbf{F}^+)' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \\ &= \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \\ &= \mathbf{S}_{\mathbf{y}}, \end{aligned}$$

$$\begin{aligned} \mathbf{S}_{\mathbf{y}}^* \mathbf{S}_{\mathbf{y}} \mathbf{S}_{\mathbf{y}}^* &= \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \\ &= \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}\mathbf{F}^+)' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \\ &= \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \\ &= \mathbf{S}_{\mathbf{y}}^*, \end{aligned}$$

$$\begin{aligned}
 (\mathbf{S}_y \mathbf{S}_y^*)' &= \left[\mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \right]' \\
 &= \left[\mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F} \mathbf{F}^+)' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \right]' \\
 &= \left[\mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \right]' \\
 &= (\mathbf{F}^+ \mathbf{F})' \\
 &= \mathbf{F}^+ \mathbf{F} \\
 &= \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \\
 &= \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F} \mathbf{F}^+)' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \\
 &= \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \\
 &= \mathbf{S}_y \mathbf{S}_y^*
 \end{aligned}$$

$$\begin{aligned}
 (\mathbf{S}_y^* \mathbf{S}_y)' &= \left[\mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \right]' \\
 &= \left[\mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \right]' \\
 &= \left[\mathbf{F}' (\mathbf{F}^+)' \right]' \\
 &= \left[(\mathbf{F}^+ \mathbf{F})' \right]' \\
 &= \mathbf{F}^+ \mathbf{F} \\
 &= \mathbf{F}^+ \mathbf{F} \mathbf{F}^+ \mathbf{F} \\
 &= (\mathbf{F}^+ \mathbf{F})' \mathbf{F} \mathbf{F}^+ \\
 &= \mathbf{F}' \mathbf{F} \mathbf{F}^+ (\mathbf{F}^+)' \\
 &= \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{S}_{\mathbf{x}^{(D)}} \mathbf{F} \mathbf{F}^+ (\mathbf{F}^+)' \\
 &= \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} \mathbf{F}^+ \mathbf{S}_{\mathbf{x}^{(D)}} (\mathbf{F}^+)' \\
 &= \mathbf{S}_y^* \mathbf{S}_y
 \end{aligned}$$

Estas relações mostram que \mathbf{S}_y^* relativa à matriz \mathbf{S}_y , satisfaz (P1)-(P4) do Lema 2.7.1, pelo que \mathbf{S}_y^* é a inversa generalizada de Moore-Penrose de \mathbf{S}_y , ou seja, $\mathbf{S}_y^+ = \mathbf{S}_y^*$ e, consequentemente,

$$\begin{aligned}
 \text{MD}^2 \left(\mathbf{x}_l^{(D)} \right) &= \left(\mathbf{x}_l^{(D)} - \bar{\mathbf{x}}^{(D)} \right)' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \left(\mathbf{x}_l^{(D)} - \bar{\mathbf{x}}^{(D)} \right) \\
 &= (\mathbf{F} \mathbf{y}_l - \mathbf{F} \bar{\mathbf{y}})' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} (\mathbf{F} \mathbf{y}_l - \mathbf{F} \bar{\mathbf{y}}) \\
 &= (\mathbf{y}_l - \bar{\mathbf{y}})' \mathbf{F}' \mathbf{S}_{\mathbf{x}^{(D)}}^{-1} \mathbf{F} (\mathbf{y}_l - \bar{\mathbf{y}}) \\
 &= (\mathbf{y}_l - \bar{\mathbf{y}})' \mathbf{S}_y^* (\mathbf{y}_l - \bar{\mathbf{y}}) \\
 &= (\mathbf{y}_l - \bar{\mathbf{y}})' \mathbf{S}_y^+ (\mathbf{y}_l - \bar{\mathbf{y}}) \\
 &= \text{MD}^2(\mathbf{y}_l) \quad l = 1, 2, \dots, n.
 \end{aligned}$$

Agora, usando o Teorema 2.7.1 e a notação em (2.28), obtém-se

$$\text{MD}^2(\mathbf{x}_l^{(j)}) = \text{MD}^2(\mathbf{x}_l^{(D)}) = \text{MD}^2(\mathbf{y}_l), \quad \text{para } j = 1, 2, \dots, D-1,$$

completando, assim, a prova. □

O resultado do Teorema 2.7.2 é inválido do ponto de vista da robustez. A igualdade da distância de Mahalanobis para as transformações *clr* e *alr* é válida apenas para as estimativas não robustas definidas pela média aritmética e matriz de covariância da amostra, visto que para estimadores robustos, como é o caso do estimador MCD, o teorema não é válido. Para o caso de se considerar o estimador MCD na estimação dos parâmetros populacionais $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, outro resultado para a distância de Mahalanobis é conhecido. Antes do enunciar recorde-se uma propriedade das matrizes inversa enunciada aqui sob a forma de Lema.

Lema 2.7.2. *Se uma matriz quadrada A tem uma matriz inversa B à direita ou à esquerda, então A é não singular e B é uma inversa de A .*

Teorema 2.7.3. *As distâncias de Mahalanobis para dados transformados por *ilr* são as mesmas que no caso da transformação *alr* se a estimativa de localização \mathbf{T} e a estimativa de covariância \mathbf{C} forem equivariantes afim.*

Atendendo ao Lema 2.7.2 tem-se todas as condições para demonstrar o Teorema 2.7.3.

Demonstração. Seja $\mathbf{x}^{(D)}$, \mathbf{y} e \mathbf{z} matrizes que resultam das transformações *alr*, *clr* e *ilr*, respetivamente, para uma composição $\mathbf{x} \in S^D$. Observe-se que a variável D é escolhida como sendo a variável razão para a transformação *alr*. Da prova do Teorema 2.7.2, sabe-se $\mathbf{x}^{(D)} = \mathbf{F}\mathbf{y}$ e $\mathbf{y} = \mathbf{F}^+\mathbf{x}^{(D)}$. Considere-se as relações existentes entre a transformação *clr* e *ilr* enunciadas em (2.14) e (2.15). Consequentemente, das transformações *alr* e *ilr* é possível obter os resultados seguintes:

$$\begin{aligned}\mathbf{x}^{(D)} &= \mathbf{F}\mathbf{y} \\ &= \mathbf{F}\mathbf{V}'\mathbf{z}, \text{ por (2.14)}\end{aligned}$$

e,

$$\begin{aligned}\mathbf{z} &= \mathbf{V}\mathbf{y} \\ &= \mathbf{V}\mathbf{F}^+\mathbf{x}^{(D)} \text{ pois } \mathbf{y} = \mathbf{F}^+\mathbf{x}^{(D)}\end{aligned}$$

onde $\mathbf{F}\mathbf{V}'$ e $\mathbf{V}\mathbf{F}^+$ são ambas matrizes quadradas de dimensão $(D-1) \times (D-1)$. Além disso, partindo da expressão $\mathbf{y} = \mathbf{F}^+\mathbf{x}^{(D)}$, multiplicando por \mathbf{V} do lado esquerdo e usando as relações descritas anteriormente, de forma detalhada tem-se o seguinte resultado:

$$\begin{aligned}\mathbf{y} &= \mathbf{F}^+\mathbf{x}^{(D)} \\ \Leftrightarrow \mathbf{V}\mathbf{y} &= \mathbf{V}\mathbf{F}^+\mathbf{x}^{(D)} \\ \Leftrightarrow \mathbf{z} &= \mathbf{V}\mathbf{F}^+\mathbf{x}^{(D)} \\ \Leftrightarrow \mathbf{F}\mathbf{V}'\mathbf{z} &= \mathbf{F}\mathbf{V}'\mathbf{V}\mathbf{F}^+\mathbf{x}^{(D)} \\ \Leftrightarrow \mathbf{x}^{(D)} &= \mathbf{F}\mathbf{V}'\mathbf{V}\mathbf{F}^+\mathbf{x}^{(D)}\end{aligned}\tag{2.29}$$

Face ao resultado obtido em (2.29) e comparando ambos os membros resulta que $\mathbf{F}\mathbf{V}'\mathbf{V}\mathbf{F}^+ = \mathbf{I}$. Então, pelo Lema 2.7.2, $\mathbf{V}\mathbf{F}^+$ é a matriz inversa da matriz não singular $\mathbf{F}\mathbf{V}'$. Por fim, utilizando (2.24) e o Teorema 2.7.1 resulta em

$$\text{MD}^2(\mathbf{z}) = \text{MD}^2(\mathbf{V}\mathbf{F}^+\mathbf{x}^{(D)}) = \text{MD}^2(\mathbf{x}^{(j)})$$

□

O Teorema 2.7.3 completa as relações entre as três transformações mencionadas. Quando se utilizam estimativas clássicas, isto é, a média aritmética e a matriz de covariância da amostra (total), as três transformações conduzem às mesmas distâncias de Mahalanobis. Visto que a detecção de observações atípicas é mais confiável com estimativas robustas serão estas a considerar no cálculo da distância de Mahalanobis. Contudo, verifica-se que as distâncias de Mahalanobis robustas são as mesmas para os dados transformados por *alr* e *ilr*, desde que as estimativas utilizadas sejam equivariantes afim.

Portanto, optar pela transformação *alr* ou *ilr* será uma decisão que cabe ao analista decidir. Contudo, tendo em conta as características destas duas transformações enunciadas nas Subsecções 2.4.1 e 2.4.3, respetivamente, a transformação *ilr* prevalece sobre a segunda o que torna o processo de decisão e posterior análise dos dados composicionais mais fácil e vantajoso.

Relativamente à transformação *clr* constatou-se na Subsecção 2.4.2 que esta resulta em dados colineares. Tal facto torna-a inapropriada em técnicas estatísticas robustas baseadas na matriz de covariância [11]. No caso das distâncias de Mahalanobis robustas, as estimativas de μ e Σ são obtidas pelo estimador MCD que só pode ser determinado para conjuntos de dados não singulares cuja característica da matriz de dados seja igual ao número de variáveis [14]. Deste modo, a transformação *clr* torna-se desaconselhável em dados que contêm *outliers* na qual uma abordagem robusta deve ser tida em conta.

A transformação *ilr* não apresenta o problema da colinearidade, pois a matriz de covariância resultante desta transformação $\Sigma_{\mathbf{Z}}$ é não singular, ($|\Sigma_{\mathbf{Z}}| \neq 0$) [6]. Além disso, goza de propriedades estatísticas compatíveis com qualquer tipo de análise estatística no espaço Euclidiano. Assim sendo, esta transformação é a ideal na detecção de *outliers* em dados composicionais.

Apesar do que se pode concluir anteriormente, quando se pretende representar os dados composicionais num biplot a fim de se identificar os *outliers*, a transformação *ilr* tem a desvantagem de que as novas variáveis não sejam diretamente interpretáveis em termos das variáveis originais. E, portanto também se torna útil o uso da transformação *clr* na análise gráfica destes dados (ver Subsecção 3.2.4).

2.8 Componentes irregulares

Como referido em secções anteriores, a análise de dados composicionais baseia-se nas log-razões entre as partes da composição, pelo que as composições devem ser abordadas em termos de relações logarítmicas. Isto implica que as componentes que assumem valor zero não podem ser tratadas diretamente neste contexto uma vez que o logaritmo de zero não existe. Infelizmente, os zeros ocorrem com bastante frequência em conjuntos de dados. Zeros e valores ausentes (*missing values*) são, portanto, componentes irregulares segundo a qual deve-se adotar métodos apropriados a fim de se conseguir lidar com este tipo de componente [6].

Atendendo à natureza das componentes irregulares, existem várias classificações que podem ser atribuídas. Os zeros podem assumir três tipos de classificação: zeros estruturais (*structural zeros*), zeros arredondados (*rounded zeros*) ou zeros de contagem (*zero counts*). Dada a natureza dos dados de qualquer um dos três conjuntos a analisar na presente dissertação (dados de contagem) a existência de componentes irregulares corresponderão a zeros de contagem.

Numa análise estatística, quando os conjuntos de dados contêm componentes irregulares, nem sempre a maioria dos métodos pode ser aplicada diretamente a esses conjuntos. Uma forma intuitiva de lidar com esta questão seria retirar as observações. Contudo, no caso multivariado, pode resultar numa grave perda de informação [36]. Em vez de se excluir essas componentes é preferível optar por um pré-processamento dos dados e preencher/substituir os espaços com valores apropriados, designando-se este procedimento por imputação. Após os dados em falta terem sido imputados, o conjunto de dados pode ser analisado pelas usuais técnicas estatísticas multivariadas.

Ao longo das últimas décadas, para a imputação de zeros e valores ausentes, muitos algoritmos baseados em modelos foram desenvolvidos. Os métodos multivariados mais aconselháveis são baseados nas semelhanças entre os objetos e/ou variáveis. De todos os métodos que existem apenas abordar-se-á o mais conhecido baseado na distância do k -vizinho mais próximo (k -Nearest Neighbor, k -NN).

De seguida, analisar-se-á em mais pormenor o tipo de componente irregular e o algoritmo k -NN escolhido para imputação. Para mais detalhe sobre a classificação de zeros e os algoritmos destas componentes irregulares recomendo a leitura das seguintes referências: Pawlowsky *et al* 2015 ([6]) e Filzmoser *et al* 2018 ([19]).

2.8.1 Zeros de contagem

Dos três conjuntos de dados para estudo na presente dissertação apenas um, o conjunto de dados relativo à habilitação académica, contém observações “incompletas”, ou seja, tem componentes irregulares. Recorde-se que estes dados dizem respeito a contagens de pessoas que afirmaram, nos Censos de 2011, que no período de 2005 a 2011 mudaram de residência passando a habitar no município em causa. De acordo com as contagens obtidas, há municípios onde não se observam quaisquer novos residentes em algumas categorias do conjunto de dados por habilitação académica. Consequentemente, este conjunto possui características que levam à presença de componentes irregulares. Observou-se ao todo 86 municípios repartidos por quatro variáveis (partes da composição), onde se observa zeros de contagem. Considere-se o seguinte exemplo, contendo apenas seis desses 86, relativo a esse conjunto.

Do conjunto de dados em questão foram extraídas as informações presentes na Tabela 2.6. A cada município faz corresponder a proporção de habitantes com doutoramento entre os que se deslocaram para esse município. Por exemplo, para Abrantes observa-se o valor 0.00283. Significa que 0.283% dos habitantes de Abrantes em 2011, que lá não residiam em 2005, tinham Doutoramento. Em contrapartida, verifica-se que Aguiar da Beira, Alfândega da Fé e Alijó têm valores ausentes, ou seja, entende-se que estes municípios não têm habitantes que foram residir para esses municípios tendo como habilitação académica um Doutoramento.

Tabela 2.6: Proporção de habitantes que se deslocaram em cada município com Doutoramento.

Municípios	Proporção com Doutoramento
Abrantes	0.00283
Águeda	0.00219
Aguiar da Beira	0
Alandroal	0.01533
Alfândega da Fé	0
Alijó	0

Observando então a existência de zeros no conjunto de dados, é importante atribuir uma classificação a estes zeros. De acordo com as três classificações mencionadas previamente, pode-se descartar a hipótese de serem zeros arredondados uma vez que os zeros não provêm de nenhum valor que necessite de arredondamento. Os zeros estruturais são valores intrinsecamente zero devido a uma dada limitação. Diz-se que uma determinada variável tem um zero estrutural numa dada observação quando essa parte não está adequadamente definida ou simplesmente não pode existir devido a alguma razão pré determinada [37]. Ora, não há nenhuma restrição/condição/circunstância física/legal/natural que impeça os municípios de terem residentes com Doutorado. Por isso, elimina-se o pressuposto dos zeros serem estruturais. Todavia, dado que a informação das observações corresponde a contagens, a sua natureza leva a concluir que os zeros presentes nos dados são zeros de contagem.

2.8.2 Algoritmo k -NN para imputação

O método do k -vizinho mais próximo é um dos métodos mais utilizados para lidar com componentes irregulares. A ideia deste método é usar uma medida de distância para encontrar as $k \geq 1$ observações mais similares de uma composição, e preencher/substituir as componentes irregulares usando a informação variável dos vizinhos [36].

No contexto dos dados composicionais, a medida de distância apropriada deste método é a distância de Aitchison (definida na Secção 2.2). Quando uma composição contém componentes irregulares em várias partes, a imputação é feita sequencialmente (uma célula após a outra), procurando os k vizinhos mais próximos entre as observações [36]. Esta opção é preferida entre várias alternativas possíveis (ver [36]), pois permite que o analista possa escolher as k observações durante a imputação sequencial.

O método do k -NN usado para imputação é numericamente estável uma vez que nenhum esquema iterativo é requerido. No entanto, contém algumas limitações. Particularmente, o número ideal k tem que ser determinado sendo, geralmente, obtido por simulação. De forma aleatória seleciona-se observações que serão omitidas e estima-se os espaços vazios com base em diferentes valores de k , medindo o erro entre os valores imputados e os valores originais das observações. O valor de k que gerar o menor erro pode ser considerado como o k ótimo. Uma outra limitação diz respeito ao seu uso sobre amostras de tamanho reduzido uma vez que poucos “bons” vizinhos podem estar disponíveis [21, 19].

Para calcular a parte que falta de uma composição, o algoritmo k -NN baseia-se na mediana das observações que correspondem aos k vizinhos mais próximos. Como as proporções entre as partes são as mesmas para composições proporcionalmente equivalentes, as observações precisam primeiro de ser ajustadas de acordo com o tamanho total das partes [19]. Veja-se a seguinte descrição do algoritmo [36].

Considere-se uma composição $\mathbf{x}_l = (x_{l1}, x_{l2}, \dots, x_{lD})$, com $l = 1, 2, \dots, n$, e seja $M_l \subset \{1, 2, \dots, D\}$ o conjunto de índices correspondente às partes omissas da composição \mathbf{x}_l . Então, $O_l = \{1, 2, \dots, D\} \setminus M_l$ refere-se às partes composicionais de \mathbf{x}_l “preenchidas”, isto é, que não contém as partes com valores omissos. Para imputar esses valores omissos x_{lj} , para $\forall j \in M_l$, são consideradas as partes preenchidas nas posições j e em O_l , e são identificados os k vizinhos mais próximos $\mathbf{x}_{l_1}, \mathbf{x}_{l_2}, \dots, \mathbf{x}_{l_k}$ para a composição \mathbf{x}_l usando a distância de Aitchison. A j -ésima parte da composição de todos os k vizinhos mais próximos é de interesse para a imputação. Primeiro, essas partes precisam de ser ajustadas por fatores que comparam o

tamanho das partes de O_l . Os fatores de ajuste são obtidos por:

$$f_{l\kappa} = \frac{\sum_{o \in O_l} x_{lo}}{\sum_{o \in O_l} x_{l\kappa o}}, \quad \text{para } \kappa = 1, 2, \dots, k. \quad (2.30)$$

Utilizar estes fatores como pesos para as observações torna os k vizinhos mais próximos comparáveis. O valor imputado substituindo a parte omissa x_{lj} é

$$x_{lj}^* = \text{mediana}\{f_{l1}x_{l1j}, f_{l2}x_{l2j}, \dots, f_{lk}x_{lkj}\}.$$

Usando a mediana obtém-se robustez para os *outliers* nas partes j dos k vizinhos mais próximos. Embora a escolha dos pesos de ajuste em (2.30) seja coerente com a definição de dados composicionais, uma versão mais robusta é preferível. A sugestão proposta na literatura é usar os fatores de ajuste dados por:

$$f_{lj}^* = \frac{\text{mediana}_{o \in O_l} x_{lo}}{\text{mediana}_{o \in O_l} x_{l\kappa o}}, \quad \text{para } \kappa = 1, 2, \dots, k, \quad (2.31)$$

que tem mostrado levar a resultados mais estáveis para dados que contêm observações atípicas.

Considere-se o exemplo esquematizado pela Tabela 2.7. Pela sua visualização, ao todo existem 10 municípios que podem ser agrupados em 4 grupos diferentes consoante as partes da composição com zeros de contagem, referenciados pelas siglas NA's, presentes nos vários tipos de habilitações académicas.

Tabela 2.7: Alguns municípios que possuem zeros de contagem (NA's) nas diferentes partes das suas composições.

Habilitações académicas				
Municípios	Bacharelato	Doutoramento	Ens. Pos Secundário	Mestrado
Aguiar da Beira	—	NA	NA	—
Barrancos	—	NA	NA	—
Boticas	—	NA	NA	—
Corvo	—	NA	NA	—
Tabuaço	—	NA	NA	—
Calheta (R.A.A.)	NA	—	—	—
Penamacor	—	NA	NA	NA
Vila Nova de Paiva	—	NA	NA	NA
Terras de Bouro	—	NA	—	NA
Vila Flor	—	NA	—	NA

Suponha-se que se pretende utilizar o algoritmo k -NN para efetuar a imputação dos zeros de contagem referente ao município da Calheta (R.A.A.). Ora, esta composição possui apenas um zero de contagem no *Bacharelato* para ser imputado. Escolhendo $k = 3$ como sendo o número de vizinhos mais próximos, o algoritmo irá determinar pela distância de Aitchison as três composições mais próximas da composição que se pretende imputar. Visto que o conjunto de dados por habilitação académica possui *outliers*, o algoritmo ao selecionar as três composições determina os fatores de ajuste baseados na mediana das observações, seguindo a fórmula (2.31).

A imputação dos zeros de contagem foi realizada, separadamente, para valores diferentes de k , nomeadamente $k = 2, 3, 4$ e 10 . Constatou-se que os resultados da imputação não variavam com a escolha de k . Assim, para prosseguir com a análise dos dados demográficos, por mera opção, seleccionou-se $k = 3$.

Aplicando o algoritmo às composições anteriores, o resultado é apresentado na Tabela 2.8.

Tabela 2.8: Resultado do algoritmo k -NN para imputação dos zeros de contagem para os diferentes municípios assinalados pela sigla NA na Tabela 2.7.

Municípios	Habilitações académicas			
	<i>Bacharelato</i>	<i>Doutoramento</i>	<i>Ens. Pos Secundário</i>	<i>Mestrado</i>
Aguiar da Beira	—	0.00117	0.00703	—
Barrancos	—	0.00221	0.01993	—
Boticas	—	0.00303	0.01643	—
Corvo	—	0.00289	0.01300	—
Tabuaço	—	0.00334	0.01197	—
Calheta (R.A.A.)	0.02072	—	—	—
Penamacor	—	0.00166	0.00777	0.00807
Vila Nova de Paiva	—	0.00540	0.00513	0.00725
Terras de Bouro	—	0.00287	—	0.01099
Vila Flor	—	0.00275	—	0.00852

Capítulo 3

Metodologias Gráficas

3.1 Estatística descritiva de dados composicionais

Na estatística descritiva, o foco principal é resumir informação acerca de um determinado conjunto de dados. Medidas estatísticas usuais como a média aritmética, a variância ou o desvio padrão para uma variável ou, ainda, a matriz de covariância no caso multivariado são ferramentas proeminentes para sumariar o conjunto de dados a analisar.

As medidas estatísticas da análise multivariada não são muito informativas para os dados composicionais devido às suas características particulares. Tentativas de aplicar estas medidas às composições originais que obedecem à geometria de Aitchison são bastante problemáticas. Tal facto deve-se a que as medidas estatísticas não lidam com o princípio de invariância de escala (Secção 2.3) [19]. Portanto, se tais medidas devem ser interpretáveis diretamente em termos de partes de uma composição, abordagens alternativas são necessárias. Nesta subsecção apresentar-se-á um conjunto de medidas estatísticas aplicadas a dados composicionais para que informação útil e vantajosa possa ser extraída.

Uma alternativa adequada à média aritmética quando o espaço de resultados é o simplex caracteriza-se pela média geométrica das componentes [38]. Neste contexto a média geométrica é designada por centro uma vez que caracteriza o centro da distribuição da amostra em questão. A sua definição é apresentada de seguida.

Definição 3.1.1 (Centro). *Seja $\mathbf{X} = [x_{ld}]$, $l = 1, 2, \dots, n$, $d = 1, 2, \dots, D$, uma amostra aleatória de l composições de D partes. O centro da amostra é o vetor de médias geométricas das partes definido por*

$$cen(\mathbf{X}) = \mathcal{C} \left[\left(\prod_{l=1}^n x_{l1} \right)^{\frac{1}{n}}, \left(\prod_{l=1}^n x_{l2} \right)^{\frac{1}{n}}, \dots, \left(\prod_{l=1}^n x_{lD} \right)^{\frac{1}{n}} \right],$$

onde $\mathcal{C}(\cdot)$ é a operação de fecho.

Exemplo 3.1.1. Designa-se por \mathbf{X}_{53} uma amostra de 5 composições de 3 partes, x_1 , x_2 e x_3 , do conjunto de dados por situação profissional que representam, respetivamente, as variáveis *Desempregado*, *Empregado* e *Inativo*. Considere-se essas 5 composições que correspondem aos municípios de Beja, Esposende, Ílhavo, Sines e Viseu.

A matriz \mathbf{X} é dada por

$$\mathbf{X} = \begin{bmatrix} 0.104 & 0.510 & 0.386 \\ 0.068 & 0.546 & 0.386 \\ 0.094 & 0.561 & 0.345 \\ 0.072 & 0.571 & 0.357 \\ 0.094 & 0.502 & 0.404 \end{bmatrix}$$

Segundo a Definição 3.1.1, o centro da amostra \mathbf{X} é obtido por

$$\begin{aligned} \text{cen}(\mathbf{X}) &= \mathcal{C} \left[\left(\prod_{l=1}^5 x_{l1} \right)^{\frac{1}{5}}, \left(\prod_{l=1}^5 x_{l2} \right)^{\frac{1}{5}}, \left(\prod_{l=1}^5 x_{l3} \right)^{\frac{1}{5}} \right] \\ &= \mathcal{C}(0.085, 0.537, 0.375) \\ &= (0.085, 0.539, 0.376). \end{aligned}$$

Portanto, a média aritmética de \mathbf{X} é o vetor $(0.085, 0.539, 0.376)$ na qual estão representadas as médias geométricas das 3 partes/componentes.

Como medida de variabilidade dos dados composicionais existe a variância de log-razão (*logratio variance*) entre duas partes composicionais, definida a seguir.

Definição 3.1.2 (Variância de log-razão). *Seja $\mathbf{x} \in S^D$ uma composição de D partes. A variância de log-razão entre duas partes x_i e x_j é dada por*

$$\tau_{ij} = \text{var} \left(\ln \frac{x_i}{x_j} \right).$$

De forma a que se obtenha uma ideia mais abrangente sobre a variabilidade dos dados composicionais é necessário calcular a variância de log-razão entre todos os pares de partes das composições da amostra. Deste modo, obtém-se uma matriz de variação (*variation matrix*), enunciada em seguida [2, 6].

Definição 3.1.3 (Matriz de variação). *Seja $\mathbf{X} = [x_{ld}]$, $l = 1, 2, \dots, n$, $d = 1, 2, \dots, D$, uma amostra aleatória de n composições de D partes. A matriz de variação de \mathbf{X} é uma matriz quadrada $D \times D$, denotada por \mathbf{T} , que determina a variância das log-razões entre partes da composição, sendo definida da seguinte forma:*

$$\mathbf{T} = [\tau_{ij}] = \left[\text{var} \left(\ln \frac{x_i}{x_j} \right) \right], \quad i, j = 1, 2, \dots, D.$$

Perante esta definição tem-se uma matriz simétrica e com a diagonal preenchida por elementos nulos, isto é,

$$\mathbf{T} = \begin{bmatrix} 0 & \tau_{12} & \tau_{13} & \tau_{14} & \dots & \tau_{1D} \\ & 0 & \tau_{23} & \tau_{24} & \dots & \tau_{2D} \\ & & 0 & \tau_{34} & \dots & \tau_{3D} \\ & & & 0 & \dots & \vdots \\ & & & & \ddots & \tau_{D-1,D} \\ & & & & & 0 \end{bmatrix}.$$

Exemplo 3.1.2. Tendo em conta o Exemplo 3.1.1, a matriz de variação de \mathbf{X} é

$$\mathbf{T} = \begin{bmatrix} 0 & 0.093 & 0.206 \\ & 0 & 0.134 \\ & & 0 \end{bmatrix}.$$

Com esta matriz realça-se o facto de que a maior variação entre duas componentes provém da relação entre *Empregado* e *Inativo*, registando-se $\tau_{13} = 0.206$. Por outro lado, a menor variação é entre *Empregado* e *Desempregado* com um valor de $\tau_{12} = 0.093$. Tal significa então que o rácio entre proporção de *Empregado* e *Desempregado* é pouco variável entre os 5 municípios, ao contrário do rácio entre proporção de *Empregado* e *Inativo* que apresenta maior variabilidade.

Alternativamente à matriz de variação, pode obter-se uma matriz de variação normalizada (*normalized variation matrix*) para dados composicionais tendo como suporte a Definição 3.1.3 [6]. Adicionando uma constante de normalização aos elementos da matriz de variação, obtém-se

$$\tau_{ij}^* = \text{var} \left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j} \right) = \frac{1}{2} \tau_{ij},$$

e, consequentemente, sendo $\mathbf{T}^* = [\tau_{ij}^*]$, tem-se que $\mathbf{T}^* = \frac{1}{2} \mathbf{T}$, e a matriz \mathbf{T}^* continua a ser simétrica com a diagonal de zeros.

Admitindo a normalidade das log-razões, em [6] deduziu-se o estimador de máxima verossimilhança para a variância de log-razão dado por

$$\hat{\tau}_{ij} = \frac{1}{n} \sum_{l=1}^n \left(\ln \frac{x_{li}}{x_{lj}} - \ln \frac{\hat{g}(\mathbf{x}_i)}{\hat{g}(\mathbf{x}_j)} \right)^2,$$

onde $\hat{g}(\mathbf{x}_i)$ e $\hat{g}(\mathbf{x}_j)$ correspondem, respetivamente, às médias geométricas dos vetores de partes \mathbf{x}_i e \mathbf{x}_j .

Por último, para fins teóricos, é importante certificar-se de que a variabilidade total de um conjunto de dados composicionais não depende de uma representação de coordenadas específica. Para se medir a dispersão global de uma amostra de dados composicionais $\mathbf{X}_{n \times D}$, utiliza-se a variação total (*Sample total variance*), sendo a sua definição exibida a seguir. Observe-se que, por vezes, a variância total é também designada por variância métrica (*metric variance*) [9].

Definição 3.1.4 (Variância total). *Seja $\mathbf{X} = [x_{ld}]$, $l = 1, 2, \dots, n$, $d = 1, 2, \dots, D$, uma amostra aleatória de n composições de D partes. A variância total da amostra \mathbf{X} é definida do seguinte modo:*

$$\text{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \text{var} \left(\ln \frac{x_i}{x_j} \right) = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \tau_{ij} = \frac{1}{D} \sum_{i=1}^D \sum_{j=1}^D \tau_{ij}^*.$$

Exemplo 3.1.3. Perante o último exemplo apresentado, a variância total é dada por

$$\text{totvar}(\mathbf{X}) = \frac{1}{3} (0.093 + 0.134 + 0.206) = 0.144.$$

3.2 Representação gráfica de dados composicionais

A representação gráfica de um conjunto de dados é um método muito vantajoso em qualquer análise exploratória de dados uma vez que permite ao analista visualizar tendências nos dados. No contexto da análise de dados composicionais duas ferramentas visuais podem ser utilizadas, nomeadamente, os diagramas ternários e os biplots.

Os diagramas ternários são gráficos de dispersão fechados de três componentes usados para representar os dados composicionais no simplex S^3 , na sua forma natural: composicional e relativa, sem haver a necessidade de aplicar uma transformação. Porém têm a particularidade de que apenas podem exibir somente três partes de composições.

Em 1971 Gabriel introduziu o biplot, um gráfico para dados multivariados que permite visualizar simultaneamente os indivíduos e as variáveis no mesmo esboço. Mais tarde, em 2002, Aitchison e Greenacre adaptaram a metodologia biplot a dados composicionais. Atendendo a que o objetivo desta dissertação centra-se na deteção de *outliers* multivariados, uma análise robusta deste gráfico deve ser tida em consideração.

Nas subsecções seguintes apresentar-se-á a teoria por detrás destes dois métodos e as ilustrações sobre cada um deles serão apresentadas no Capítulo 4, onde uma abordagem prática das técnicas de análise serão aplicadas ao conjunto dos dados demográficos.

3.2.1 Diagrama ternário

O diagrama ternário resulta da representação dos dados composicionais com uma restrição de soma constante, isto é, o gráfico exibe uma visualização gráfica do simplex de três partes,

$$S^3 = \{ \mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1 > 0, x_2 > 0, x_3 > 0, x_1 + x_2 + x_3 = k \},$$

para uma dada constante k , geralmente 1 ou 100 para os casos de proporções ou percentagens, respectivamente [19].

A maior parte da literatura aplicada a análise de dados composicionais, essencialmente em Geologia, restringe os gráficos a (sub) composições de três partes visto que a representação gráfica de composições com mais do que três partes torna-se difícil de visualizar [6]. Para $D = 3$, o simplex pode ser representado no plano \mathbb{R}^3 numa superfície triangular com vértices $X_1 = [k, 0, 0]$, $X_2 = [0, k, 0]$ e $X_3 = [0, 0, k]$, tal como mostra a Figura 3.1.

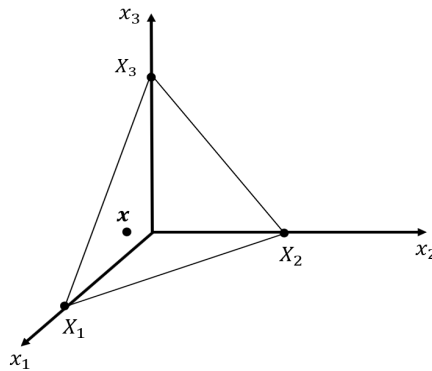


Figura 3.1: Representação do simplex em \mathbb{R}^3 .

Definição 3.2.1 (Diagrama Ternário). *Um diagrama ternário é um triângulo equilátero de vértices X_1, X_2, X_3 sendo o comprimento de cada segmento considerado como proporção de uma parte dada, isto é, um dos lados é igual a $\sqrt{2}k$ para $k = x_1 + x_2 + x_3$, onde uma composição $\mathbf{x} = (x_1, x_2, x_3)$ é representada a uma distância x_1 do lado oposto ao vértice X_1 , a uma distância x_2 do lado oposto ao vértice X_2 e a uma distância x_3 do lado oposto ao vértice X_3 .*

Na Figura 3.2 visualiza-se o diagrama ternário no plano \mathbb{R}^2 que é uma representação do triângulo observado na Figura 3.1. Geralmente, o tripleto (x_1, x_2, x_3) é designado por coordenadas baricêntricas de \mathbf{x} e o centro do diagrama ternário denominado por baricentro.

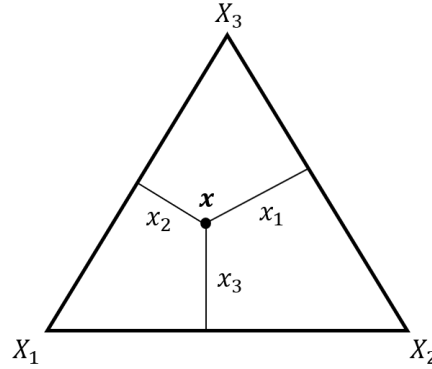


Figura 3.2: Diagrama ternário (adaptado de [2]).

As bordas do diagrama ternário correspondem a marginais nulas na representação em \mathbb{R}^3 , da Figura 3.1, e os próprios vértices representam composições com uma parte igual a k e as restantes duas partes iguais a zero, conforme referido anteriormente. A soma das distâncias permanece constante para qualquer escolha das partes de \mathbf{x} [19].

A construção de um diagrama ternário em \mathbb{R}^2 inicia-se com a representação dos vértices em coordenadas cartesianas, no sentido contrário dos ponteiros do relógio. Assume-se que $X_1 = (u_0, v_0)$ são as coordenadas do vértice X_1 (*Origem*) e os restantes vértices são $X_2 = (u_0 + 1, v_0)$ e $X_3 = (u_0 + \frac{1}{2}, v_0 + \frac{\sqrt{3}}{2})$, onde a segunda coordenada do vértice X_3 se obtém pelo teorema de Pitágoras. Deste modo, o diagrama ternário terá a forma exibida na Figura 3.3.

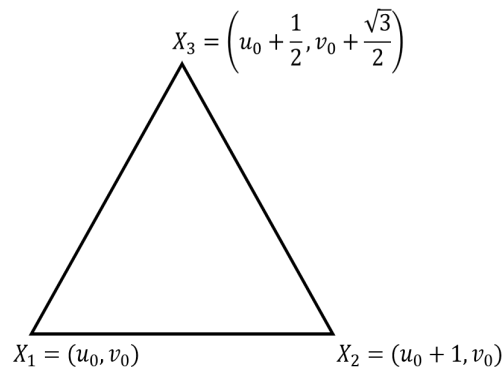


Figura 3.3: Representação de um diagrama ternário com coordenadas cartesianas.

Para se representar um ponto correspondente a uma composição \mathbf{x} de três partes, $\mathbf{x} = (x_1, x_2, x_3)$, fechado para uma constante k , torna-se necessário conhecer as suas coordenadas (u, v) , que são obtidas através da combinação linear convexa das coordenadas dos vértices, dada por

$$(u, v) = \frac{1}{k} (x_1 X_1 + x_2 X_2 + x_3 X_3).$$

Para interpretar o diagrama ternário pode-se recorrer à propriedade de que os segmentos ortogonais que ligam uma composição \mathbf{x} (ver Figura 3.2) com os três lados de um triângulo equilátero têm soma dos seus comprimentos constante [37]. Assim:

- 1) uma composição representada sobre (ou muito próxima de) uma aresta do triângulo indica a dominância das duas partes que formam essa aresta;
- 2) uma composição assinalada sobre um vértice indica a dominância da parte associada a esse vértice.

Portanto, ao analisar dados composicionais por meio de diagramas ternários deve-se ter em atenção os seguintes padrões:

- i. as (sub) composições concentram-se num vértice: indica a dominância da parte associada a esse vértice (Figura 3.4 (a));
- ii. as (sub) composições distribuem-se ao longo de uma aresta: indica a dominância das duas partes associadas a essa aresta (Figura 3.4 (b));
- iii. as (sub) composições agrupam-se em torno do baricentro do simplex: indica que as partes representadas têm proporções aproximadamente iguais (Figura 3.4 (c)).

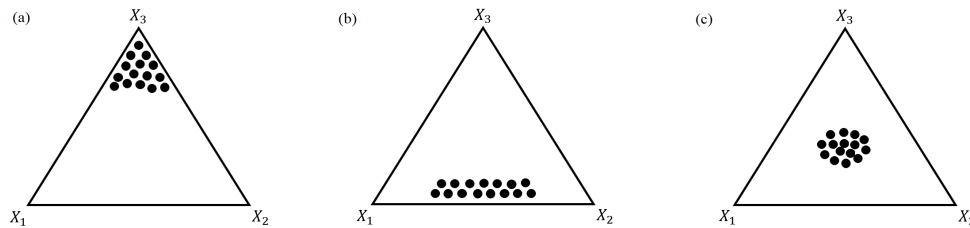


Figura 3.4: Representação em diagramas ternários dos padrões (i), (ii) e (iii), respetivamente.

Para além destes três padrões existem ainda mais três segundo o qual os dados composicionais podem ser representados num diagrama ternário, são eles:

- iv. as (sub) composições formam um padrão linear paralelo a um dos lados: indica que as proporções da parte associada ao vértice oposto nas (sub) composições são (aproximadamente) constantes (Figura 3.5 (a));
- v. as (sub) composições formam um padrão linear (aproximadamente) perpendicular a um dos lados: indica que as partes associadas a esse lado são (aproximadamente) proporcionais (reduzida variabilidade relativa) (Figura 3.5 (b));
- vi. as (sub) composições encontram-se dispersas no simplex: indica que as partes apresentam elevada variabilidade relativa entre si (Figura 3.5 (c)).

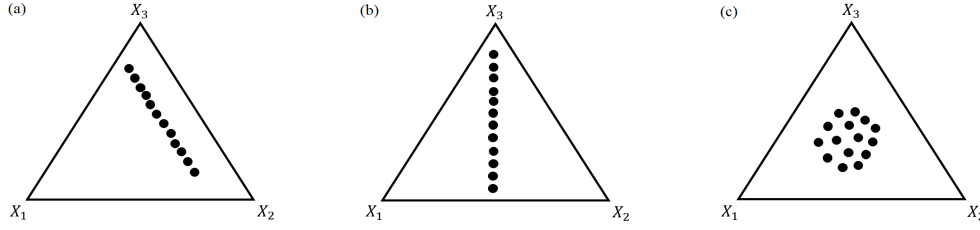


Figura 3.5: Representação em diagramas ternários dos padrões (iv), (v) e (vi), respetivamente.

Ao exibir dados composicionais num diagrama ternário, situações que frequentemente causam problemas são as características apresentadas em (i), (ii) e (iii). Uma forma de contorná-las é adotar o método de centralização que, usualmente, exibirá as características (iv), (v) e (vi).

A centralização é um caso especial da perturbação (Definição 2.2.1) e consiste em perturbar cada linha da matriz de dados, com composições completas, pela inversa do centro, de modo que o conjunto de dados fique distribuído em torno do baricentro do simplex. Realizar uma análise com base nos dados centrados permite uma melhor observação da real tendência no conjunto de dados [37, 3].

Exemplo 3.2.1. Suponha-se que se tem a matriz \mathbf{X} do Exemplo 3.1.1 cujo vetor associado ao centro de \mathbf{X} é $(0.085, 0.539, 0.376)$. Deste modo, a inversa do centro de \mathbf{X} é dada por

$$\ominus \text{cent}(\mathbf{X}) = \mathcal{C}\left(\frac{1}{0.085}, \frac{1}{0.539}, \frac{1}{0.376}\right) = \mathcal{C}(11.765, 1.855, 2.660) = (0.723, 0.114, 0.163).$$

Portanto, cada observação (linha) da matriz de dados pode ser perturbada pela inversa do centro²:

$$\begin{aligned} \mathbf{X}_c &= \begin{bmatrix} 0.104 & 0.510 & 0.386 \\ 0.068 & 0.546 & 0.386 \\ 0.094 & 0.561 & 0.345 \\ 0.072 & 0.571 & 0.357 \\ 0.094 & 0.502 & 0.404 \end{bmatrix} \oplus \begin{pmatrix} 0.723 \\ 0.114 \\ 0.163 \end{pmatrix} \\ &= \mathcal{C} \begin{bmatrix} 0.075192 & 0.058140 & 0.062918 \\ 0.049164 & 0.062244 & 0.062918 \\ 0.067962 & 0.063954 & 0.056235 \\ 0.052056 & 0.065094 & 0.058191 \\ 0.067962 & 0.057228 & 0.065852 \end{bmatrix} = \begin{bmatrix} 0.383 & 0.296 & 0.321 \\ 0.282 & 0.357 & 0.361 \\ 0.361 & 0.340 & 0.299 \\ 0.297 & 0.371 & 0.332 \\ 0.356 & 0.300 & 0.345 \end{bmatrix}. \end{aligned}$$

Na Figura 3.6 visualiza-se o diagrama ternário com os dados composicionais da matriz \mathbf{X} e após aplicar o método de centralização. Verifica-se que os dados sofrem uma deslocação posicionando-se em torno do baricentro do simplex.

²A operação entre matrizes de dados no simplex obedece à regra usual da Teoria de Álgebra Matricial, mas tendo em conta a teoria do simplex. O fecho de uma matriz corresponde ao fecho das linhas.

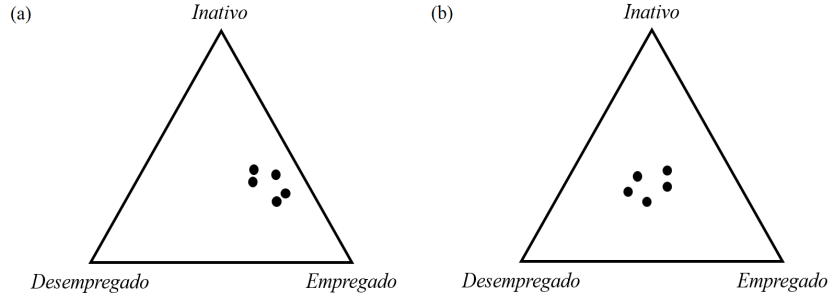


Figura 3.6: Diagrama ternário com os dados composicionais (a) iniciais (b) após o método de centralização.

Conclui-se que as partes representadas no diagrama (*Desempregado*, *Empregado* e *Inativo*) têm proporções aproximadamente iguais.

3.2.2 Biplot

Para verificar se existe alguma relação entre duas variáveis, o gráfico de dispersão é das ferramentas mais utilizadas. O biplot pode ser classificado como uma generalização deste gráfico: pretende-se representar um gráfico de dispersão contendo marcadores para as variáveis e para as observações.

Definição 3.2.2.1 (Biplot). *O biplot é uma representação gráfica, em duas dimensões, de uma matriz de dados $\mathbf{X}_{n \times p}$, onde as n linhas correspondentes às observações são representadas como projeção da nuvem de dados no espaço reduzido e, simultaneamente, sob o mesmo gráfico, são representadas as p colunas da matriz de dados através da projeção dos eixos das variáveis também no mesmo espaço reduzido.*

Observação 3.2.2.1. *O termo “bi” em biplot refere-se à representação conjunta das observações e das variáveis da matriz de dados no mesmo gráfico e não ao facto da exibição ser bidimensional.*

Sem perda de generalidade, seja $\mathbf{X}_{n \times D}$ uma matriz de dados tal que:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1D} \\ x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nD} \end{bmatrix} \quad (3.1)$$

onde o elemento x_{ld} representa o valor da l -ésima observação na d -ésima variável.

Habitualmente, a construção de biplots inicia-se com uma transformação da matriz \mathbf{X} , de acordo com a natureza dos dados, para que se obtenha uma matriz transformada sobre a qual se aplica o biplot. Vários tipos de transformação podem ser usados, por exemplo: a centralização em relação à média geral, a centralização em relação à média das variáveis, a normalização das variáveis, a raiz quadrada, as transformações log-razões, entre outras [7].

Relativamente aos dados multivariados sem restrições é frequente usar-se a centralização em relação à média das variáveis. Deste modo, suponha-se que \mathbf{X} em (3.1) é uma matriz centrada pela média das variáveis.

* Decomposição em Valores Singulares

De forma a que seja possível representar num biplot a matriz de dados é necessário recorrer a uma fatorização de \mathbf{X} . A Decomposição em Valores Singulares (*Singular Value Decomposition*, DVS) de uma matriz genérica é um dos resultados mais importantes na Teoria das Matrizes uma vez que permite fatorizar qualquer matriz. Baseando-se neste resultado, Gabriel desenvolveu a teoria dos biplots [39].

Qualquer matriz de dados $\mathbf{X}_{n \times D}$ (para $n > D$) definida por (3.1), com característica r ($r \leq \min(n, D)$), pode ser fatorizada na forma

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{W}', \quad (3.2)$$

designando-se por DVS da matriz \mathbf{X} , onde

- $\mathbf{U}_{n \times r}$ é a matriz ortogonal de vetores singulares à esquerda, isto é, vetores próprios de $\mathbf{X}\mathbf{X}'$;
- $\mathbf{D}_{r \times r}$ é a matriz diagonal formada pelos valores singulares positivos, isto é, raízes quadradas dos valores próprios de $\mathbf{X}'\mathbf{X}$ dispostos por ordem decrescente: $d_1 \geq d_2 \geq \dots \geq d_r > 0$;
- $\mathbf{W}_{p \times r}$ é a matriz ortogonal de vetores singulares à direita, isto é, de vetores próprios de $\mathbf{X}'\mathbf{X}$;

Observação 3.2.2.2. Se \mathbf{X} tem DVS dada por (3.2), então a transposta de \mathbf{X} tem DVS dada por $\mathbf{X}' = \mathbf{W}\mathbf{D}\mathbf{U}'$.

Pelo teorema de Eckart-Young ([40]) pode-se usar os primeiros maiores $r^* < r$ valores singulares e correspondentes vetores singulares para obter uma matriz $\hat{\mathbf{X}} = [\hat{x}_{ld}]$ de dimensão $n \times D$ e característica r^* , sendo esta a melhor aproximação no sentido dos mínimos quadrados de \mathbf{X} , ou seja, tal que

$$\|\mathbf{X} - \hat{\mathbf{X}}\| = \sqrt{\sum_{l=1}^n \sum_{d=1}^D (x_{ld} - \hat{x}_{ld})^2} = \min_{\mathbf{Y}} \|\mathbf{X} - \mathbf{Y}\|, \quad (3.3)$$

para todas as possíveis matrizes \mathbf{Y} de característica r^* , onde $\|\cdot\|$ denota a norma matricial de Frobenius. A solução do problema (3.3) é

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{P},$$

onde $\mathbf{P}_{D \times D} = \mathbf{W}_* \mathbf{W}_*'$ e \mathbf{W}_* é uma matriz ortonormal, $D \times r^*$, cujas colunas correspondem aos vetores próprios associados aos primeiros r^* maiores valores próprios da matriz $\mathbf{X}'\mathbf{X}$ [41].

Para $r^* = 2$, a matriz $\hat{\mathbf{X}}$ seria dada por

$$\hat{\mathbf{X}} = \begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{bmatrix} \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \\ \vdots & \vdots \\ v_{1D} & v_{2D} \end{bmatrix}'. \quad (3.4)$$

O biplot da matriz de dados \mathbf{X} é construído com base na matriz aproximada $\hat{\mathbf{X}}$, no espaço reduzido de dimensão $r^* = 2$. A precisão desse biplot diz respeito à precisão na aproximação de \mathbf{X} por $\hat{\mathbf{X}}$, e a qualidade da aproximação em (3.4) corresponde à proporção de variabilidade explicada, usualmente expressa em percentagem, dada por

$$\pi_r = \frac{d_1^2 + d_2^2}{\sum_{l=1}^r d_l^2} \times 100\%.$$

* Biplots

Apresentada a Decomposição em Valores Singulares segue-se de forma simplificada o conteúdo subjacente à construção e interpretação dos biplots.

Seja $\mathbf{X}_{n \times D} = [X_1 X_2 \dots X_D]$ a matriz de dados com característica r fatorizada por (3.2), isto é, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{W}'$. Para $0 \leq \alpha \leq 1$, a DVS pode ser escrita da seguinte forma:

$$\mathbf{X} = \mathbf{U}\mathbf{D}^{1-\alpha}(\mathbf{W}\mathbf{D}^\alpha)'$$

Definindo

$$\mathbf{G} = \mathbf{U}\mathbf{D}^{1-\alpha} \quad (3.5)$$

$$\mathbf{H} = \mathbf{W}\mathbf{D}^\alpha \quad (3.6)$$

obtém-se

$$\mathbf{X} = \mathbf{G}\mathbf{H}'. \quad (3.7)$$

A matriz \mathbf{G} tem dimensão $n \times r$ existindo uma correspondência entre as linhas de \mathbf{G} e as observações da matriz \mathbf{X} . Por sua vez, \mathbf{H} tem dimensão $D \times r$ havendo uma correspondência entre as linhas de \mathbf{H} e as variáveis da matriz \mathbf{X} [16]. Deste modo, o biplot no espaço \mathbb{R}^r tem a propriedade de que o produto escalar entre a l -ésima linha de \mathbf{G} e a d -ésima coluna de \mathbf{H}' é igual ao elemento x_{ld} da matriz \mathbf{X} [7].

Quando aplicado na prática, representado em \mathbb{R}^2 , nem sempre é possível obter uma representação gráfica exata, exceto se a matriz \mathbf{X} tiver característica $r = 2$ e, assim, o biplot é representado num espaço de dimensão reduzida essencialmente de duas dimensões, \mathbb{R}^2 , tornando-se mais fácil a sua interpretação.

Para $r = 2$, a DVS fornece uma decomposição adequada para a fatorização da matriz $\hat{\mathbf{X}}$ obtida em (3.4), conforme apresentada em (3.7). Consequentemente, escolhendo $\mathbf{G} = (d_1^{1-\alpha}\mathbf{u}_1 \ d_2^{1-\alpha}\mathbf{u}_2)$ e $\mathbf{H} = (d_1^\alpha\mathbf{v}_1 \ d_2^\alpha\mathbf{v}_2)$ resulta

$$\begin{aligned} \hat{\mathbf{X}} = \mathbf{G}\mathbf{H}' &= \begin{bmatrix} d_1^{1-\alpha}u_{11} & d_2^{1-\alpha}u_{21} \\ d_1^{1-\alpha}u_{12} & d_2^{1-\alpha}u_{22} \\ \vdots & \vdots \\ d_1^{1-\alpha}u_{1n} & d_2^{1-\alpha}u_{2n} \end{bmatrix} \begin{bmatrix} d_1^\alpha v_{11} & d_1^\alpha v_{12} & \dots & d_1^\alpha v_{1D} \\ d_2^\alpha v_{21} & d_2^\alpha v_{22} & \dots & d_2^\alpha v_{2D} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_D \end{bmatrix}, \end{aligned} \quad (3.8)$$

onde os n vetores $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n$, representarão os n marcadores das n observações e os D vetores $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_D$ representarão os D marcadores das D variáveis descritas na matriz $\hat{\mathbf{X}}$ que serão representados graficamente no biplot.

A constante $\alpha \in [0, 1]$ é designada por parâmetro de forma e consoante o valor que se escolha em (3.5) e (3.6) obtém-se diferentes biplots. Os diferentes valores de α fornecem exatamente a mesma matriz de aproximação e destacam diferentes aspetos da matriz de dados. Para a interpretação do biplot, os dois valores de α mais usados são $\alpha = 0$ e $\alpha = 1$ [7]. Cada escolha origina um biplot com características e interpretações diferentes.

Observação 3.2.2.3. *No caso do parâmetro de forma, $\alpha \in [0, 1]$, tomar o valor $\alpha = 0.5$, o biplot correspondente designa-se na literatura por *SQRT biplot*. Com este tipo de representação atinge-se a mesma qualidade de representação para os indivíduos e variáveis. No entanto, essa qualidade não é máxima nem para os indivíduos nem para as variáveis como quando se obtém fazendo $\alpha = 0$ e $\alpha = 1$, respetivamente.*

Independentemente da escolha de n a construção do biplot é feita sobre as duas primeiras componentes principais resultantes da aplicação de Análise de Componentes Principais (ACP) sobre a matriz de dados \mathbf{X} .

Dada uma matriz de dados \mathbf{X} definida por (3.1), para realizar uma ACP é necessário obter os vetores próprios e os valores próprios da matriz de covariância amostral $\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\mathbf{X}$ de \mathbf{X} . Os vetores próprios de \mathbf{S} correspondem a cada uma das colunas de \mathbf{W} que definem os coeficientes das componentes principais e são proporcionais aos marcadores $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_D$ a representar no biplot.

Em (3.2), ao multiplicar à direita ambos os membros por \mathbf{W} obtém-se \mathbf{XW} que representa os *scores*, ou seja, os indivíduos nas componentes principais, dado por

$$\mathbf{XW} = \mathbf{UDW}'\mathbf{W} = \mathbf{UD}. \quad (3.9)$$

Tal significa que

$$(\mathbf{XW})'(\mathbf{XW}) = (\mathbf{UD})'(\mathbf{UD}) = \mathbf{DU}'\mathbf{UD} = \mathbf{D}^2, \quad (3.10)$$

onde \mathbf{D}^2 é uma matriz diagonal $D \times D$ contendo os valores próprios de \mathbf{S} por ordem decrescente, $d_1^2 \geq d_2^2 \geq \dots \geq d_D^2 > 0$, que representam os quadrados dos valores singulares de \mathbf{X} contidos na diagonal da matriz \mathbf{D} . Por (3.10), as variâncias das componentes principais coincidem, a menos de uma constante igual a $(n-1)$, aos valores próprios de $\mathbf{X}'\mathbf{X}$. Assim, quando se tomam as duas primeiras componentes principais (isto é, a representação dos indivíduos pelos seus *scores*) ter-se-á a melhor qualidade de representação das observações, ou seja, a máxima variabilidade das observações no espaço de dimensão dois. Por (3.9) e (3.10), tal ocorrerá quando os marcadores $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n$ dos indivíduos forem representados pela matriz \mathbf{XW} , ou seja, \mathbf{UD} .

* Biplot com $\alpha = 1$

Para $\alpha = 1$, as matrizes \mathbf{G} e \mathbf{H} definidas em (3.5) e (3.6), respetivamente, ficam do seguinte modo:

$$\mathbf{G} = \mathbf{U} \text{ e } \mathbf{H} = \mathbf{WD}, \quad (3.11)$$

com $\mathbf{X} = \mathbf{GH}'$. Prova-se que esta fatorização preserva a métrica das colunas (*Column Metric Preserving*), favorecendo com máxima qualidade a representação das variáveis [42]. O biplot associado a esta fatorização é frequentemente denominado na literatura por biplot clássico de Gabriel, ou ainda, de modo menos comum, por **biplot de covariância** ou GH biplot.

A Análise de Componentes Principais é um método fundamental no processamento de dados exploratórios para obter informação sobre a estrutura dos dados multivariados [19]. Este método foca-se em explicar a estrutura de covariância dos dados através de novas variáveis, designadas por componentes principais, que são definidas como combinações lineares das variáveis originais, reduzindo a dimensionalidade dos dados.

Tendo em conta (3.7), a matriz de covariâncias amostral de \mathbf{X} será dada por

$$\begin{aligned}\mathbf{S} &= \frac{1}{n-1} \mathbf{X}'\mathbf{X} = \frac{1}{n-1} (\mathbf{GH}')'\mathbf{GH}' \\ &= \left(\frac{1}{\sqrt{n-1}} \mathbf{HG}' \right) \left(\frac{1}{\sqrt{n-1}} \mathbf{GH}' \right).\end{aligned}$$

Dado que $\mathbf{G} = \mathbf{U}$, tem-se que $\mathbf{G}'\mathbf{G} = \mathbf{U}'\mathbf{U} = \mathbf{I}_D$, resultando que

$$\mathbf{S} = \frac{1}{n-1} \mathbf{HH}'. \quad (3.12)$$

Como \mathbf{H} é dado pelos marcadores $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_D$ de acordo com (3.8) tem-se que:

$$\begin{aligned}\mathbf{HH}' &= \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \vdots \\ \mathbf{h}_D \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_D \end{bmatrix} \\ &= \begin{bmatrix} \|\mathbf{h}_1\|^2 & \dots & \langle \mathbf{h}_1, \mathbf{h}_d \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{h}_l, \mathbf{h}_d \rangle & \dots & \|\mathbf{h}_D\|^2 \end{bmatrix}\end{aligned}$$

e, portanto, igual a (3.12) significa que

- $\|\mathbf{h}_l\|^2 = (n-1) \text{var}(X_l)$
- $\langle \mathbf{h}_l, \mathbf{h}_j \rangle = (n-1) \text{cov}(X_l, X_d)$

ou seja, geometricamente tem-se que

- o comprimento do marcador \mathbf{h}_l é proporcional ao desvio padrão da variável X_l
- o cosseno do ângulo entre marcadores \mathbf{h}_l e \mathbf{h}_d é proporcional à correlação entre X_l e X_d , isto é,

$$\begin{aligned}\cos(\mathbf{h}_l, \mathbf{h}_d) &= \frac{\langle \mathbf{h}_l, \mathbf{h}_d \rangle}{\|\mathbf{h}_l\| \|\mathbf{h}_d\|} \\ &= \frac{\text{cov}(X_l, X_d)}{\sqrt{\text{var}(X_l) \text{var}(X_d)}} \\ &= \text{correlação entre } X_l \text{ e } X_d\end{aligned}$$

Para que os comprimentos das setas associadas às colunas X_1, X_2, \dots, X_D da matriz \mathbf{X} correspondam aos valores dos desvios padrão descritos na diagonal de \mathbf{S} , deve-se conceber em (3.7) a seguinte fatorização [25]:

$$\mathbf{X} = \mathbf{G}\mathbf{H}' = (\sqrt{n-1}\mathbf{G}) \left(\frac{1}{\sqrt{n-1}}\mathbf{H}' \right).$$

Assim, deverá ser tomado os marcadores $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n$ e $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n$ das linhas de $\mathbf{G}^* = \sqrt{n-1}\mathbf{G}$ e de $\mathbf{H}^{*'} = \frac{1}{\sqrt{n-1}}\mathbf{H}'$, respetivamente.

✱ Biplot com $\alpha = 0$

Para $\alpha = 0$, as matrizes \mathbf{G} e \mathbf{H} definidas em (3.5) e (3.6), respetivamente, apresentam-se do seguinte modo:

$$\mathbf{G} = \mathbf{U}\mathbf{D} \text{ e } \mathbf{H} = \mathbf{W},$$

com $\mathbf{X} = \mathbf{G}\mathbf{H}'$. Tendo em conta o que se mencionou atrás, esta escolha de $\mathbf{G} = \mathbf{U}\mathbf{D}$ determina a melhor representação de linhas da matriz \mathbf{X} pelo que a fatorização $\mathbf{X} = \mathbf{G}\mathbf{H}' = (\mathbf{U}\mathbf{D})\mathbf{W}'$ preserva a métrica das linhas (*Row Metric Preserving*), favorecendo a representação das observações. Por isso, se conclui que este tipo de biplot beneficia a exibição das observações. O biplot associado a esta fatorização é habitualmente denominado por JK biplot ou, de modo menos comum, por **biplot de forma**.

3.2.3 Biplot composicional

Os biplots para dados composicionais baseiam-se na teoria apresentada na subsecção anterior e adaptados às características dos dados composicionais.

Sem perda de generalidade, seja $\mathbf{X} = [x_{ld}] = [X_1 \ X_2 \ \dots \ X_D]$ uma matriz $n \times D$ de dados composicionais onde X_1, X_2, \dots, X_D representam as D partes onde se descrevem os dados. A representação num biplot do conjunto de dados requer, primeiramente, a aplicação de uma transformação log-razão aos dados antes de centrá-los, de modo que os vetores singulares à esquerda e à direita reproduzam a escala relativa dos dados composicionais [7].

A transformação log-razão que se usa para construir biplots de dados composicionais é a transformação *clr*. Por conseguinte, o biplot de \mathbf{X} é construído com base numa matriz \mathbf{Z} transformada cujas entradas são as coordenadas *clr*-transformadas calculadas sobre a matriz de dados \mathbf{X} , tendo sido previamente centrada em relação à média das colunas [6].

Partindo de $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_D]$, a transformação *clr* desta matriz é denotada por $\mathbf{Z}^* = [z_{ld}^*]$, com

$$z_{ld}^* = \ln \frac{x_{ld}}{\sqrt[D]{\prod_{d=1}^D x_{ld}}} = \ln x_{ld} - \frac{1}{D} \sum_{d=1}^D \ln x_{ld} = s_{ld} - s_{l+}$$

onde $s_{ld} = \ln x_{ld}$ e $s_{l+} = \frac{1}{D} \sum_{d=1}^D s_{ld}$. Logo, $\frac{1}{D} \sum_{d=1}^D z_{ld}^* = 0$.

Da matriz \mathbf{Z}^* obtém-se a matriz $\mathbf{Z} = [z_{ld}] = [Z_1 \ Z_2 \ \dots \ Z_D]$ das observações *clr*-transformadas e centrada em relação à média das colunas, isto é,

$$z_{ld} = z_{ld}^* - \frac{1}{n} \sum_{d=1}^n z_{ld}^* = s_{ld} - s_{l+} - \frac{1}{n} \sum_{l=1}^n (s_{ld} - s_{l+}) = s_{ld} - s_{l+} - s_{+d} + s_{++}$$

onde $s_{+d} = \frac{1}{n} \sum_{l=1}^n s_{ld}$ e $s_{++} = \frac{1}{nD} \sum_{l=1}^n \sum_{d=1}^D s_{ld}$. Portanto, $\frac{1}{n} \sum_{l=1}^n z_{ld} = 0$ e $\sum_{d=1}^D \ln z_{ld} = 0$.

Assim, sendo $\hat{\mathbf{Z}}$ a melhor aproximação de \mathbf{Z} no sentido dos mínimos quadrados, na fatorização em (3.8) os vetores $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n$ são designados por marcadores das linhas de $\hat{\mathbf{Z}}$ e dizem respeito às projeções das n composições no plano, por sua vez os vetores $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_D$ são denominados por marcadores das colunas de $\hat{\mathbf{Z}}$ referindo-se às projeções das D coordenadas *clr* no plano [6].

Interpretação do biplot composicional

A Figura 3.7 exibe um biplot de uma matriz de dados composicionais $\mathbf{X}_{n \times D}$, para $D = 5$, onde se observam os seguintes elementos:

- (i) uma origem do referencial que representa o centro O de todas as variáveis (círculo verde);
- (ii) pontos denotados como marcadores das n observações;
- (iii) vetores para cada uma das partes denominados por raios com origem no centro O ;
- (iv) vértices para cada uma das D partes (variáveis) em coordenadas *clr*-transformadas (centradas);
- (v) segmentos de reta que ligam dois vértices designados por ligações (*links*).

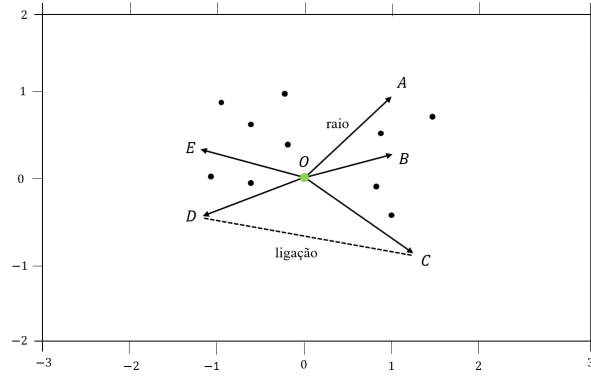


Figura 3.7: Ilustração de um biplot de uma matriz de dados \mathbf{X} genérica com a seguinte representação dos símbolos: ● linhas (observações); → colunas (variáveis); — ligação (razão entre variáveis); ● origem do conjunto de dados (centro O) (adaptado de [7]).

Observação 3.2.3.1. Os raios são denotados por \mathbf{h}_d , com $d = 1, 2, \dots, D$ e a ligação entre dois raios \mathbf{h}_d e \mathbf{h}_k nos seus vértices por $\mathbf{h}_d - \mathbf{h}_k$. Observe-se que, geometricamente, esta ligação corresponde à diferença entre dois vetores (os raios). Num biplot composicional os elementos básicos são as ligações, nomeadamente a sua direção e o seu comprimento, e não os raios tal como quando se interpreta um biplot de covariância.

Assim, para os biplots de dados composicionais, as ligações fornecem as diretrizes para exploração da variabilidade dos dados composicionais de acordo com as seguintes propriedades [37, 6]:

1. A ligação entre dois raios \mathbf{h}_d e \mathbf{h}_k nos seus vértices, $\mathbf{h}_d - \mathbf{h}_k$, indica a variabilidade da log-razão entre as partes envolvidas, isto é,

$$\|\mathbf{h}_d - \mathbf{h}_k\|^2 \propto \text{var} \left(\ln \frac{X_d}{X_k} \right).$$

De facto, pelas propriedades dos biplots [39], tem-se que

$$\begin{aligned} \|\mathbf{h}_d - \mathbf{h}_k\| &\propto \sqrt{\sum_{l=1}^n (z_{ld} - z_{lk})^2} \\ &= \sqrt{\sum_{l=1}^n [s_{ld} - s_{l+} - s_{+d} + s_{++} - (s_{lk} - s_{l+} - s_{+k} + s_{++})]^2} \\ &= \sqrt{\sum_{l=1}^n [s_{ld} - s_{lk} - (s_{+d} - s_{+k})]^2} \\ &= \sqrt{\sum_{l=1}^n [\ln x_{ld} - \ln x_{lk} - (\ln x_{+d} - \ln x_{+k})]^2} \\ &= \sqrt{\sum_{l=1}^n \left(\ln \frac{x_{ld}}{x_{lk}} - \frac{1}{n} \sum_{l=1}^n \ln \frac{x_{ld}}{x_{lk}} \right)^2} \\ &\propto \sqrt{\text{var} \left(\ln \frac{X_d}{X_k} \right)} \end{aligned} \quad (3.13)$$

Conclui-se, portanto, que a distância entre dois vértices é a menos de uma constante de proporcionalidade uma aproximação do desvio padrão das correspondentes log-razões. Assim, se a qualidade de representação dos dados no biplot for suficientemente elevada, deve-se ter em conta dois aspetos:

- (a) duas variáveis *clr*-transformadas com ligação muito curta entre si significa que as correspondentes variáveis originais são proporcionais e têm log-razão quase constante, coincidindo com valores baixos na matriz de variação (Definição 3.1.3);
- (b) duas variáveis *clr*-transformadas com ligação muito longa significa que as correspondentes variáveis originais apresentam variabilidade muito grande entre si, coincidindo com valores elevados na matriz de variação;

Assim, por exemplo, se dois vértices coincidirem (ou quase coincidirem) significa que a ligação $\mathbf{h}_d - \mathbf{h}_k$ tem comprimento aproximadamente igual a zero e, portanto, $\text{var} \left(\ln \frac{X_d}{X_k} \right)$ será 0 (ou aproximadamente 0), sendo o rácio $\frac{X_d}{X_k}$ aproximadamente constante,

$$\ln \left(\frac{X_d}{X_k} \right) \approx \text{constante}.$$

Consequentemente, neste caso, há uma redundância das duas partes envolvidas.

2. O ângulo formado pela ligação $\mathbf{h}_d - \mathbf{h}_k$ (entre as variáveis Z_d e Z_k) e pela ligação $\mathbf{h}_a - \mathbf{h}_b$ (entre as variáveis Z_a e Z_b) indica o valor do coeficiente de correlação entre as duas log-razões correspondentes,

$$\cos(\mathbf{h}_d - \mathbf{h}_k, \mathbf{h}_a - \mathbf{h}_b) \approx \text{corr} \left(\ln \frac{X_d}{X_k}, \ln \frac{X_a}{X_b} \right).$$

De facto, recorrendo a cálculos algébricos realizados em (3.13) tem-se que

$$\sqrt{\text{var}(Z_d - Z_k)} = \sqrt{\frac{1}{n-1} \sum_{l=1}^n (z_{ld} - z_{lk})^2} = \text{var} \left(\ln \frac{X_d}{X_k} \right),$$

e pelas propriedades dos biplots [39], obtém-se que

$$\begin{aligned} \cos(\mathbf{h}_d - \mathbf{h}_k, \mathbf{h}_a - \mathbf{h}_b) &\approx \text{corr}(Z_d - Z_k, Z_a - Z_b) \\ &\propto \frac{\sum_{l=1}^n (z_{ld} - z_{lk})(z_{la} - z_{lb})}{\sqrt{\text{var}(Z_d - Z_k) \text{var}(Z_a - Z_b)}} \\ &= \frac{\sum_{l=1}^n [s_{ld} - s_{lk} - (s_{+d} - s_{+k})][s_{la} - s_{lb} - (s_{+a} - s_{+b})]}{\sqrt{\text{var} \left(\ln \frac{X_d}{X_k} \right) \text{var} \left(\ln \frac{X_a}{X_b} \right)}} \\ &= \frac{\sum_{l=1}^n \left(\ln \frac{x_{ld}}{x_{lk}} - \frac{1}{n} \sum_{l=n} \ln \frac{x_{ld}}{x_{lk}} \right) \left(\ln \frac{x_{la}}{x_{lb}} - \frac{1}{n} \sum_{l=n} \ln \frac{x_{la}}{x_{lb}} \right)}{\sqrt{\text{var} \left(\ln \frac{X_d}{X_k} \right) \text{var} \left(\ln \frac{X_a}{X_b} \right)}} \\ &\propto \frac{\text{cov} \left(\ln \frac{X_d}{X_k}, \ln \frac{X_a}{X_b} \right)}{\sqrt{\text{var} \left(\ln \frac{X_d}{X_k} \right) \text{var} \left(\ln \frac{X_a}{X_b} \right)}} = \text{corr} \left(\ln \frac{X_d}{X_k}, \ln \frac{X_a}{X_b} \right) \end{aligned}$$

Conclui-se, portanto, que o cosseno do ângulo entre as ligações é uma aproximação da correlação entre as correspondentes log-razões das variáveis originais. Assim,

- (a) Se duas ligações, $\mathbf{h}_d - \mathbf{h}_k$ e $\mathbf{h}_a - \mathbf{h}_b$, formam um ângulo de 90° entre si então a correlação entre as log-razões das partes envolvidas será aproximadamente zero,

$$\text{corr} \left(\ln \frac{X_d}{X_k}, \ln \frac{X_a}{X_b} \right) \approx 0,$$

revelando que as log-razões das correspondentes partes estarão, provavelmente, não correlacionadas;

- (b) Se duas ligações formam um ângulo de 0° ou 180° , as log-razões das partes originais envolvidas estarão perfeitamente correlacionadas (direta ou indiretamente) com uma correlação igual a 1 (ou aproximadamente 1),

$$\text{corr} \left(\ln \frac{X_d}{X_k}, \ln \frac{X_a}{X_b} \right) \approx 1,$$

demonstrando que há uma relação linear entre o logaritmo do rácio das respetivas partes.

Para interpretação do biplot composicional, a qualidade deste gráfico depende da proporção de variância total retida pelo biplot. Todas as conclusões que advêm da análise do biplot devem ser confrontadas com outras ferramentas exploratórias dos dados composicionais, por exemplo a matriz de variação e o diagrama ternário [37].

3.2.4 Biplot composicional robusto

O procedimento para detecção de *outliers* será mais compreensivo se os resultados forem exibidos graficamente, por isso os *outliers* tendem a ser identificados por meios de representações gráficas dos dados. O biplot apresenta-se como uma ferramenta adequada visto que permite visualizar padrões na estrutura dos dados multivariados no espaço reduzido 2-dimensional.

Dado que os biplots composicionais são sensíveis a *outliers*, para muitas situações práticas é necessário adotar uma abordagem robusta. Tal abordagem não é direta porque a transformação *clr* resulta em dados colineares (ver Secção 2.4.2) e a estimativa da matriz de covariância não consegue lidar com a colinearidade dos dados. Por este motivo, a transformação *ilr* auxilia na resolução deste problema e ambas as transformações cooperam na construção de biplots robustos.

Em [11] propôs-se o uso destes biplots para lidar e identificar *outliers* em dados composicionais. Para a realização da ACP e construção de biplots robustos sugere-se a transformação dos dados em coordenadas *ilr*-transformadas (de acordo com uma determinada base) a fim de se obter *loadings* e *scores* robustos. Porém, para a interpretação da ACP realizada sobre a transformação *ilr*, recomenda-se que se volte a transformar os dados em coordenadas *clr*-transformadas onde a interpretação do biplot é conhecida.

Seguindo a abordagem proposta por [11] considere-se uma amostra de dados composicionais $\mathbf{X}_{n \times D}$ e a sua matriz em coordenadas *ilr*-transformadas $\mathbf{Z}_{n \times (D-1)}$, de valor médio $\boldsymbol{\mu}_{\mathbf{Z}}$ e matriz de covariância $\boldsymbol{\Sigma}_{\mathbf{Z}}$, com estimativas robustas $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ e $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$, respetivamente, obtidas pelo estimador MCD.

Tendo em conta a DVS de $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$, isto é, $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} = \mathbf{W}_{\mathbf{Z}} \mathbf{D}_{\mathbf{Z}}^2 \mathbf{W}_{\mathbf{Z}}'$, a matriz dos *scores* será a matriz $\mathbf{Z}_{n \times (D-1)}^*$ a qual descreve os dados \mathbf{Z} centrados no espaço das componentes principais robustas, ou seja,

$$\mathbf{Z}^* = [\mathbf{Z} - \mathbf{1} \hat{\boldsymbol{\mu}}_{\mathbf{Z}}'] \mathbf{W}_{\mathbf{Z}},$$

onde $\mathbf{1}$ é um vetor n dimensional com entradas iguais à unidade e $\mathbf{W}_{\mathbf{Z}}$ é a matriz dos *loadings*, cujas colunas contêm os vetores próprios de $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$.

Se a matriz de dados original $\mathbf{X}_{n \times D}$ tiver característica D , a matriz \mathbf{Z} terá característica $D - 1$, e o estimador MCD poderá ser usado para obter estimativas robustas $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ e $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ resultando em *loadings* e *scores* robustos contidos na matriz $\mathbf{W}_{\mathbf{Z}}$ e \mathbf{Z}^* , respetivamente [11].

No entanto, em coordenadas *ilr*-transformadas o biplot não é interpretável. Consequentemente, a relação linear entre as transformações *clr* e *ilr* obtida em (2.14) e (2.15) é de extrema importância, pois permite que os resultados de uma análise perante a transformação *ilr* sejam facilmente convertidos para serem interpretados no espaço *clr*, sem perda de informação.

Assim, transpondo (2.15) para notação matricial tem-se $\mathbf{Z} = \mathbf{YV}'$, e utilizando esta relação a matriz dos *scores* robustos em coordenadas *clr* é dada por

$$\mathbf{Y}^* = \mathbf{Z}^* \mathbf{V}$$

De modo semelhante, a notação matricial de (2.14) é $\mathbf{Y} = \mathbf{ZV}$ pelo que se obtém

$$\begin{aligned}\hat{\Sigma}_{\mathbf{Y}} &= \hat{\Sigma}_{\mathbf{ZV}} \\ &= \mathbf{V}' \hat{\Sigma}_{\mathbf{Z}} \mathbf{V} \\ &= \mathbf{V}' \mathbf{W}'_{\mathbf{Z}} \mathbf{D}_{\mathbf{Z}}^2 \mathbf{W}_{\mathbf{Z}} \mathbf{V} \\ &= \mathbf{W}'_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}}^2 \mathbf{W}_{\mathbf{Y}},\end{aligned}$$

onde $\mathbf{W}_{\mathbf{Y}}$ é a matriz dos *loadings* robustos em coordenadas *clr*-transformadas.

Observação 3.2.4.1. *Em virtude da relação de linearidade entre as duas transformações, os valores próprios não nulos de $\hat{\Sigma}_{\mathbf{Z}}$ são iguais aos de $\hat{\Sigma}_{\mathbf{Y}}$, pelo que a percentagem de variabilidade explicada contida na diagonal da matriz $\mathbf{D}_{\mathbf{Z}}^2$ permanece inalterável para a matriz $\mathbf{D}_{\mathbf{Y}}^2$.*

Assim sendo, para construir o biplot composicional robusto para uma matriz de coordenadas *clr*-transformadas \mathbf{Y} , é necessário aplicar a DVS. Fatorizando a matriz de acordo com (3.2) obtém-se

$$\mathbf{Y} = \mathbf{U}_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}} \mathbf{W}'_{\mathbf{Y}},$$

onde $\mathbf{D}_{\mathbf{Y}}$ é a matriz diagonal cujas entradas são as raízes quadradas dos elementos de $\mathbf{D}_{\mathbf{Y}}^2$. Portanto, mediante (3.9) tem-se a matriz dos *scores* robustos dada por:

$$\mathbf{Y}^* = \mathbf{U}_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}} \tag{3.14}$$

Em (3.14), ao multiplicar à direita ambos os membros por $\mathbf{D}_{\mathbf{Y}}^{-1}$, resulta que

$$\mathbf{Y}^* \mathbf{D}_{\mathbf{Y}}^{-1} = \mathbf{U}_{\mathbf{Y}}$$

Por intermédio de (3.7), $\mathbf{Y} = \mathbf{G}_{\mathbf{Y}} \mathbf{H}'_{\mathbf{Y}}$ e o biplot de covariâncias composicional robusto em coordenadas *clr*-transformadas é obtido, segundo (3.11), escolhendo-se

$$\mathbf{G}_{\mathbf{Y}} = \mathbf{Y}^* \mathbf{D}_{\mathbf{Y}}^{-1} \text{ e } \mathbf{H}'_{\mathbf{Y}} = \mathbf{D}_{\mathbf{Y}} \mathbf{W}'_{\mathbf{Y}},$$

onde \mathbf{Y}^* e $\mathbf{W}_{\mathbf{Y}}$ são, respetivamente, as matrizes de *scores* e *loadings* robustos da ACP.

Para interpretação deste biplot, as propriedades que advêm do biplot composicional são análogas ao biplot composicional robusto. Porém, para posterior identificação de *outliers*, optar pela robustez da representação deste gráfico torna-se uma mais valia para uma correta interpretação dos resultados.

Capítulo 4

Aplicação em dados demográficos

4.1 Matrizes de dados

No Capítulo 1 foi realizada uma breve introdução acerca dos dados demográficos para o presente estudo. Tal como se referiu, estes dados são provenientes dos Censos de 2011 e correspondem ao fluxo migratório interno em Portugal, isto é, à deslocação de pessoas da sua área de residência para outros municípios de Portugal. Um dos objetivos deste trabalho prende-se em analisar os dados sob o ponto de vista composicional. As três matrizes de dados iniciais dos três conjuntos continham informação absoluta pelo que se construiu novas matrizes contendo informação relativa (proporções), aplicando a operação de fecho (Definição 2.1.2) para $k = 1$.

Suponha-se a matriz inicial do conjunto de dados por grupo etário representada na Tabela 4.1. Note-se que o “Total 1” refere-se ao número de habitantes internos que entraram, até à data de 2011, em cada município.

Tabela 4.1: Tabela representativa da matriz inicial do conjunto de dados do grupo etário.

Municípios	Idade 0-14	Idade 15-24	Idade 25-39	Idade 40-64	Idade 65+	Total 1
Abrantes	81	299	719	723	299	2121
Águeda	136	374	1005	926	295	2736
Aguiar da Beira	5	28	195	143	43	414
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Vouzela	19	89	255	180	80	623
Total 2	33847	119861	324255	316261	99295	893519

De forma a que se obtivesse informação relativa, isto é, informação essa que é dada pela distribuição das idades dos novos habitantes que passaram a residir em cada município, foi necessário obter rácios. Utilizando a operação de fecho, construiu-se uma nova matriz de dados, $\mathbf{X}_{308 \times 5}$, e a informação relativa foi obtida pela frequência relativa, rácio entre a frequência absoluta e o “Total 1”. Por exemplo, a distribuição das idades dos habitantes em Abrantes, referente à idade dos 0 aos 14 anos, é $\frac{81}{2121} = 0.03819$. Para a nova matriz de dados eliminou-se as variáveis “Total 1” e “Total 2”, tal como mostra a Tabela 4.2.

Tabela 4.2: Tabela representativa da matriz do conjunto de dados composicional (aqui referenciada como matriz composicional) do grupo etário.

Municípios	Idade 0-14	Idade 15-24	Idade 25-39	Idade 40-64	Idade 65+
Abrantes	0.03819	0.14097	0.33899	0.34088	0.14097
Águeda	0.04971	0.13670	0.36732	0.33845	0.10782
Aguiar da Beira	0.01208	0.06763	0.47101	0.34541	0.10386
⋮	⋮	⋮	⋮	⋮	⋮
Vouzela	0.03050	0.14286	0.40931	0.28892	0.12841

Assim sendo, para o conjunto de dados por habilitação académica e por situação profissional efetuou-se o mesmo procedimento e a matriz composicional dos conjuntos é denotada por $\mathbf{X}_{308 \times 10}$ e $\mathbf{X}_{308 \times 3}$, respetivamente, sendo a análise dos dados demográficos baseada nas matrizes composicionais.

4.2 Metodologias de análise dos dados

De modo a que se pudesse examinar e extrair, adequadamente, toda a informação necessária dos três conjuntos efetuou-se a análise dos dados separadamente, onde em cada um deles o procedimento foi o mesmo mas adaptado ao conjunto.

O procedimento de análise foi realizado de acordo com metodologias numéricas e gráficas. As metodologias numéricas tiveram como finalidade a descoberta de observações atípicas numa perspetiva analítica, baseando-se na teoria do Capítulo 2 e com o auxílio de funções implementadas nas bibliotecas do *RStudio*. Por sua vez, tal como o nome indica, as metodologias gráficas serviram para uma visualização panorâmica dos resultados onde graficamente se identificaram os *outliers* com o uso dos biplots composicionais robustos. Para uma melhor interpretação da informação extraída foram utilizados outros métodos gráficos e, ainda, algumas das estatísticas descritivas do Capítulo 3.

Para proceder com a análise dos dados, as matrizes composicionais obtidas tiveram de ser transformadas de maneira a que as metodologias estatísticas usuais pudessem ser aplicadas. No Capítulo 2 conclui-se que das três transformações log-razões apresentadas, a transformação *ilr* é a que melhor se adequa para a deteção de *outliers* multivariados. Por conseguinte, deve-se determinar a base ortonormal a fim de se obter as coordenadas *ilr*-transformadas, tendo sido escolhida a técnica da PBS proposta por Filzmoser e seus colaboradores. No *RStudio* existem duas funções que permitem obter, isoladamente, a base ortonormal e as coordenadas *ilr*, provenientes da biblioteca *robCompositions* ([43]), sendo elas: *orthbasis()* e *pivotCoord()*, respetivamente. Note-se que com a função *orthbasis()* também é possível obter a matriz de contrastes e a tabela da PBS.

No Capítulo 2 constatou-se que pelos métodos baseados na estimativa da estrutura de covariância da matriz de dados, a deteção das observações atípicas decorria do uso da métrica de Mahalanobis robusta. Para esse fim obtinha-se as estimativas robustas $\hat{\boldsymbol{\mu}}$ e $\hat{\boldsymbol{\Sigma}}$ pelo estimador MCD, e estas eram implementadas na métrica. De seguida, todas as observações que continham um valor da distância superior ao valor de corte correspondente ao quantil de ordem 0.975 eram *outliers*.

Nas metodologias numéricas usou-se quatro funções disponíveis em diferentes bibliotecas

do *RStudio*, entre elas:

1. a função *covMcd()*, acessível na biblioteca *robustbase* ([44]), utiliza o algoritmo *Fast-MCD* de Rousseeuw e Van Driessen para as estimativas robustas $\hat{\mu}$ e $\hat{\Sigma}$, e possibilita a identificação de *outliers*;
2. a função *aq.plot()*, disponível na biblioteca *mvoutlier* ([45]), deteta os *outliers* tendo como *output* quatro gráficos sendo o mais relevante aquele que representa as observações atípicas detetadas pelo quantil χ_D^2 ;
3. a função *dd.plot()*, acessível na biblioteca *mvoutlier*, para além de identificar os *outliers* representa graficamente a distância de Mahalanobis clássica versus a distância de Mahalanobis robusta baseada no estimador MCD. O gráfico, designado por *D-D Plot*, contém uma linha a tracejado que referencia o conjunto de observações onde ambas as distâncias são iguais, e as linhas horizontais e verticais são desenhadas em função do valor de corte. Todas as observações que se situam para lá destas linhas (no 1º quadrante) são *outliers*, além disso diferentes símbolos e cores são usados dependendo da métrica usada: distância de Mahalanobis ou distância Euclidiana [34];
4. a função *outCoDa()*, disponível na biblioteca *robCompositions*, permite detetar observações atípicas usando métodos estatísticos clássicos e robustos. Esta função tem um detalhe interessante de que não necessita, como *input*, a matriz composicional sob coordenadas *ilr*-transformadas, ela própria tem a característica de converter os dados nas respetivas coordenadas.

Após a execução destas funções averiguou-se que nas três últimas havia instabilidade, ou seja, o número de *outliers* detetados variava para o mesmo conjunto de dados e, apenas, para a primeira função é que esse número era fixo. Por exemplo, para o conjunto da habilitação académica a função *aq.plot()* identificava observações atípicas entre 58 a 62 observações, enquanto o *dd.plot()* detetava 64 a 69 observações. Tal facto pode ser explicado uma vez que associado a cada grupo de *outliers* que se identifique encontra-se uma determinada estimativa $\hat{\mu}$ e $\hat{\Sigma}$, por conseguinte o estimador MCD para cada grupo é também diferente. O que se pretende sempre obter é a estimativa da matriz de covariância associada ao menor valor do seu determinante, o tal valor de MCD. Deste modo, de forma a contornar a questão da instabilidade, as três funções foram repetidas dez vezes e escolheu-se o grupo de *outliers* para os quais o valor do determinante da matriz de covariância era o mínimo. Note-se que no *output* das três funções não surge o valor de MCD pelo que houve necessidade de alterar o código acessível no *RStudio* de cada uma de tal forma que esse valor fosse impresso. O valor de MCD obtido era bastante baixo, por isso escolheu-se por imprimir o seu logaritmo. Assim, no código usado a estimativa da matriz de covariância está associada ao menor valor do logaritmo do seu determinante.

Nas metodologias gráficas, o biplot composicional robusto é construído com recurso à função *mvoutlier.CoDa()*, acessível na biblioteca *mvoutlier*. Além de construir biplots composicionais robustos, esta função permite a identificação dos *outliers* representando-os por cores progressivamente mais vivas, destacando-se a seguinte ordem: azul, verde, amarelo e vermelho. As cores são definidas de acordo com a média das medianas das distâncias de Mahalanobis das observações entre quantis de referência (a título ilustrativo, estes quantis de referência encontram-se assinalados na Figura 2.1 pelas elipses). Assim, observações com valores altos da média das medianas são representadas pela cor vermelha, enquanto as que possuem valores mais baixos são representadas pela cor azul [33]. De igual modo, esta função também possuía

instabilidade. Consequentemente, também o grupo das observações atípicas foi escolhido com base na repetição associada ao menor valor de MCD, tendo sido o respetivo código alterado. Por sua vez, a representação do biplot composicional robusto teve de ser feita de acordo com a solução obtida, pelo que o código foi novamente alterado inserindo-se a solução. Assim garantia-se a representação correta do biplot para o grupo de *outliers* encontrado. Note-se que entende-se por solução o menor valor de MCD e as estimativas robustas $\hat{\mu}$ e $\hat{\Sigma}$.

A função *mvoutlier.CoDa()* não só permite obter biplots composicionais robustos como também outros tipos de gráficos que visam uma melhor interpretação dos dados composicionais com observações atípicas. Por exemplo, os gráficos de dispersão univariados para as coordenadas *ilr*-transformadas. Nestes gráficos, os *outliers* são representados para cada uma das coordenadas pivô podendo assumir valores altos ou baixos consoante a distância de Mahalanobis robusta com base no estimador MCD. Um facto interessante destes gráficos de dispersão é que, individualmente, cada um deles representa cada uma das variáveis originais nas coordenadas *ilr*. Ou seja, o primeiro gráfico da variável z_1 contém toda a informação da parte x_1 , o segundo gráfico da variável z_2 contém toda a informação da parte x_2 e, assim, sucessivamente.

Observação 4.2.1. *Os biplots composicionais robustos e os gráficos de dispersão univariados exibem os outliers por números. O Apêndice A contém as tabelas para cada um dos conjuntos referentes à função mvoutlier.CoDa() com os municípios outliers e respetiva numeração.*

Para as cinco funções implementadas foi selecionado o grupo de *outliers* com o menor valor de MCD e constatou-se que o número de observações atípicas era diferente em todas elas. De forma a que se pudesse analisar o comportamento desses *outliers* decidiu-se numa primeira fase escolher o grupo que continha o maior número de observações. Assim, para esse grupo aplicou-se algumas ferramentas estatísticas para dados multivariados, destacando-se a Análise de *Clusters* e os *Heatmaps*. Os resultados obtidos não foram favoráveis e a interpretação dos mesmos não propiciaram a que se concluísse nada em concreto. Portanto optou-se por escolher numa segunda fase apenas os *outliers* comuns a todas as funções e analisou-se o comportamento destes através de gráficos de coordenadas paralelas e gráficos de dispersão para algumas variáveis originais.

Para os três conjuntos de dados, a análise efetuada com estes dois últimos gráficos não revelou quaisquer conclusões adicionais para além das que foram obtidas com os gráficos de dispersão univariados. Destaca-se que as conclusões retiradas com os gráficos de dispersão univariados apenas incidiam em *outliers* provenientes da função *mvoutlier.CoDa()*, pelo que as observações atípicas que não fizessem parte deste grupo não eram analisadas. Todavia foi realizado um diagnóstico e constatou-se que esses *outliers* que não apareciam no grupo das observações atípicas da função seguiam o comportamento de outros *outliers* detetados pelas restantes funções e que integravam o grupo dos *outliers* comuns. Por conseguinte, as conclusões que se sucediam para a análise das observações atípicas em falta eram análogas às que já tinham sido realizadas. Deste modo, para não sobrecarregar a discussão dos resultados, decidiu-se apenas apresentar o estudo dos gráficos de dispersão univariados.

Com os grupos das observações atípicas comuns decidiu-se representar os *outliers* em cartogramas a fim de ser facilmente visível os municípios *outliers* de Portugal Continental, Madeira e Açores (Grupo Ocidental, Central e Oriental), e identificá-los. Na secção a seguir serão apresentados os resultados obtidos para cada um dos três conjuntos, tendo em conta todo o procedimento de análise especificado nesta secção, bem como a discussão dos mesmos.

4.3 Análise e discussão dos resultados

Para o estudo dos dados demográficos inicia-se por proceder à leitura do ficheiro em Excel no *RStudio*. Para tal é crucial a instalação e leitura de duas bibliotecas, nomeadamente, *XLConnectJars* e *XLConnect*, uma vez que estas são o ponto de partida para fazer a transição do ficheiro do Excel para o *software*. De seguida efetua-se a leitura das bibliotecas que serão utilizadas, sendo este processo igual para cada conjunto.

4.3.1 Conjunto de dados por grupo etário

Realizada a leitura da matriz do conjunto de dados por grupo etário que contém a informação absoluta, procedeu-se com a construção da nova matriz que integra toda a informação relativa dos dados. Iniciou-se o estudo com as metodologias numéricas e, para isso, aplicou-se a PBS obtendo-se as coordenadas *ilr*-transformadas.

A primeira função executada foi a *covMcd()*. Calculada várias vezes verificou-se que o número de *outliers* não variava, ficando assim calculado o valor do logaritmo do determinante da estimativa robusta da matriz de covariância dos dados. O número de municípios *outliers* obtido foi 34 para um valor do logaritmo de MCD de -13.00288 , e guardou-se a solução encontrada. Esta função tem a vantagem de representar graficamente uma comparação dos *outliers* obtidos pela distância de Mahalanobis robusta e clássica, tal como mostra a Figura 4.1.

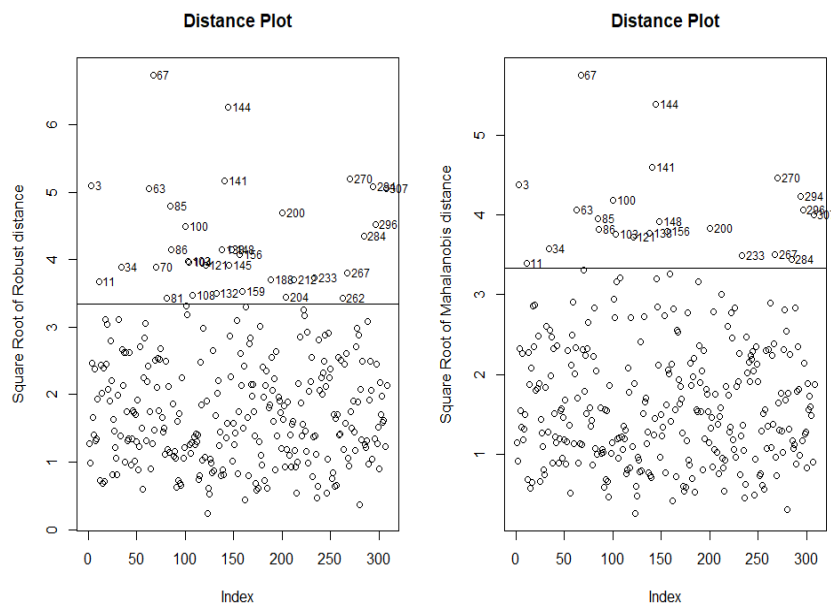
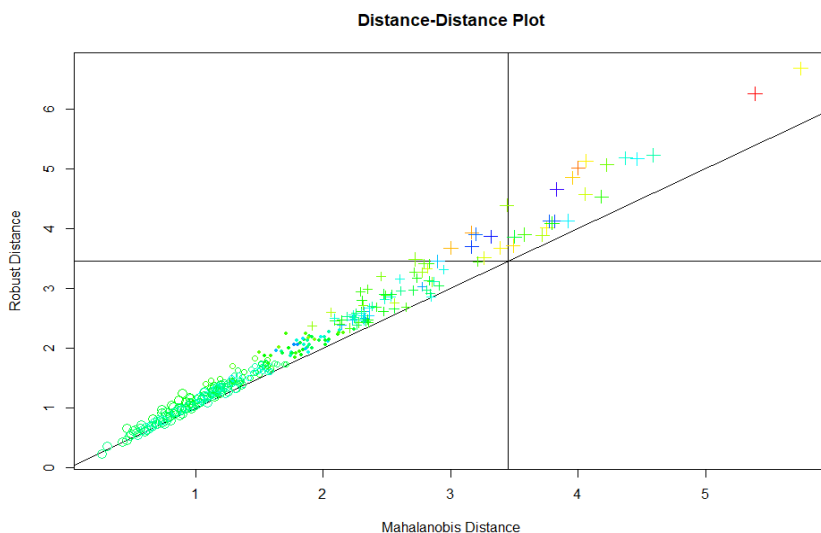
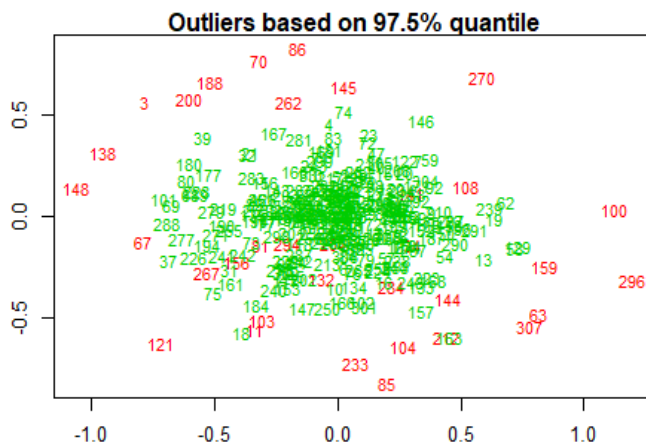


Figura 4.1: Comparação gráfica da raíz quadrada da distância de Mahalanobis robusta (gráfico da esquerda) e clássica (gráfico da direita) para o conjunto de dados por grupo etário. As observações identificadas por números são *outliers*.

Os gráficos evidenciam uma diferença pouco expressiva das duas distâncias, no entanto destaca-se que na distância de Mahalanobis robusta o número de *outliers* aumenta, e observações que não eram consideradas atípicas pela distância de Mahalanobis clássica convertem-se em atípicas.



Por fim, a última função foi a *outCoDa()*. Novamente, a repetição foi efetuada e no *input* escolheu-se a opção *method="robust"* de tal forma que se garantisse que o método aplicado para a detecção de *outliers* multivariados fosse robusto. A solução foi guardada e é dada por: 48 *outliers* para um valor do logaritmo de MCD de -13.16091 .

Relativamente às metodologias gráficas, na função *mvoutlier.CoDa()*, a solução guardada foi: 31 *outliers* com um valor do logaritmo de MCD de -13.01813 . Os biplots composicionais robustos para esta solução encontram-se na Figura 4.4.

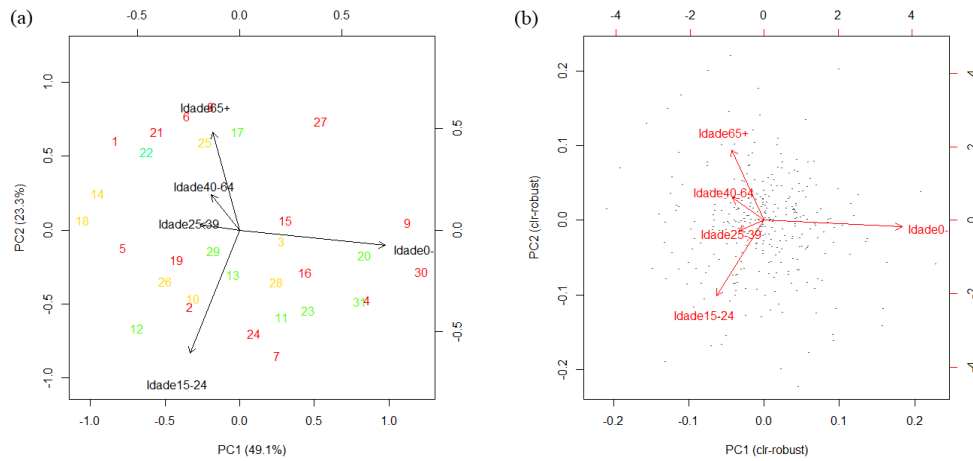


Figura 4.4: Biplots composicionais robustos (a) para visualização dos *outliers* bem como do comprimento dos *links* e (b) para visualização dos ângulos entre os *links* para o conjunto de dados por grupo etário.

A percentagem de variabilidade total retida pelo biplot, na Figura 4.4 (a), é de $49.1 + 23.3 = 72.4\%$ e para a sua interpretação é fundamental as ligações (*links*) entre as setas que fornecem informações sobre a variação relativa entre as variáveis. Desse biplot destaca-se que o menor *link* provém da ligação entre as variáveis *Idade 25-39* e *Idade 40-64*, evidenciando que estas são proporcionais entre si, possuindo log-razão $\ln\left(\frac{Idade\ 25-39}{Idade\ 40-64}\right)$ constante, comprovada pela matriz de variação composicional da Tabela 4.3. Assim, este resultado indica que a razão entre as percentagens de residentes que passaram a residir num município com idades entre os 25 e os 39 anos e os 40 e os 64 anos tende a ser a mais constante entre o restante rácio no conjunto dos 308 municípios. Ou seja, as percentagens de residentes que mudaram de município com estas idades devem manter-se proporcionalmente mais parecidas entre si, no conjunto dos 308 municípios.

Por outro lado, o maior *link* surge entre as variáveis *Idade 0-14* e *Idade 65+* demonstrando que estas possuem uma variabilidade muito grande entre si, confirmada pela Tabela 4.3. Consequentemente, significa que a razão entre as percentagens de residentes que mudaram de município com idades entre os 0 e os 14 anos e com mais de 65 anos apresenta maior variabilidade no conjunto dos 308 municípios. Neste conjunto onde se observou mobilidade de residência dos seus habitantes, haverá municípios em que a percentagem de residentes com idades dos 0 aos 14 anos será um dado valor proporcional à percentagem de residentes com idades a partir dos 65 anos mas tal valor de proporcionalidade diferirá entre os diferentes 308 municípios, sendo que para este par de grupos etários se observa a maior variabilidade.

Tabela 4.3: Matriz de variação composicional entre as variáveis dos grupos etários (a negrito destacam-se os valores do maior e do menor *link*).

	<i>Idade 0-14</i>	<i>Idade 15-24</i>	<i>Idade 25-39</i>	<i>Idade 40-64</i>	<i>Idade 65+</i>
<i>Idade 0-14</i>	0	0.167	0.157	0.152	0.182
<i>Idade 15-24</i>		0	0.074	0.079	0.117
<i>Idade 25-39</i>			0	0.043	0.087
<i>Idade 40-64</i>				0	0.083
<i>Idade 65+</i>					0

Na Figura 4.4 (b), o ângulo formado pelos *links* fornece informação acerca do coeficiente de correlação entre log-razões. Do biplot composicional robusto, verifica-se um ângulo de 90° entre os *links* de dois conjuntos de variáveis. O primeiro conjunto é formado pela variáveis: *Idade 65+*, *Idade 40-64*, *Idade 25-39* e *Idade 15-24*, sendo o segundo formado pelas variáveis: *Idade 0-14* e *Idade 25-39*. No primeiro conjunto fazem-se os *links* entre quaisquer pares de variáveis e no segundo conjunto o *link* do par das duas variáveis. Verifica-se que os *links* entre estes dois conjuntos formam um ângulo reto e, por isso, o rácio entre variáveis do primeiro conjunto são não correlacionadas com o rácio entre as partes *Idade 0-14* e *Idade 25-39*. A submatriz de correlações entre as log-razões, apresentada na Tabela 4.4, corrobora com as conclusões obtidas do biplot, pois a primeira linha dessa submatriz apresenta valores baixos e próximos de zero.

Relativamente à interpretação deste resultado, sabe-se que os indivíduos deslocam-se entre 2005 e 2011 com o propósito de mudarem de residência. O facto da razão entre as percentagens de residentes que mudaram de município para estes dois conjuntos de variáveis serem não correlacionadas leva a concluir que os indivíduos com estas idades apresentam lógicas de deslocamentos diferentes. Ou seja, os diferentes grupos etários têm problemas e questões diferentes que os levam a adotar comportamentos divergentes consoante a situação em que se encontram.

Os *links* do primeiro conjunto formam um ângulo de 0° salientando que as variáveis são correlacionadas. No entanto, tal conclusão não é diretamente comprovada pela submatriz de correlações da Tabela 4.4 visto que se observam valores relativamente baixos (por exemplo: -0.342 , -0.037 , -0.327 para alguns dos rácios). Estes valores deveriam ser elevados, próximos de 1. Tal situação poderá ser eventualmente explicada pela quantidade de variabilidade dos dados retida pelas duas primeiras componentes principais não ser suficiente para captar as correlações realmente observadas (Tabela 4.4).

As log-razões entre as variáveis *Idade 65+*, *Idade 40-64*, *Idade 25-39* e *Idade 15-24* são correlacionadas concluindo-se que a mobilidade de um determinado grupo ocorre com a mobilidade de um outro grupo etário (exceto na idade dos 0 aos 14 anos), evidenciando que indivíduos com estas idades (≥ 15 anos) possuem padrões de mobilidade relacionados entre os 308 municípios.

Tabela 4.4: Uma submatriz de correlações entre log-razões e partes no grupo etário.

	$\ln \frac{Idade\ 0-14}{Idade\ 25-39}$	$\ln \frac{Idade\ 15-24}{Idade\ 25-39}$	$\ln \frac{Idade\ 25-39}{Idade\ 40-64}$	$\ln \frac{Idade\ 40-64}{Idade\ 65+}$
$\ln \frac{Idade\ 0-14}{Idade\ 25-39}$	1.000	-0.299	-0.289	-0.063
$\ln \frac{Idade\ 15-24}{Idade\ 25-39}$		1.000	-0.342	-0.037
$\ln \frac{Idade\ 25-39}{Idade\ 40-64}$			1.000	-0.327
$\ln \frac{Idade\ 40-64}{Idade\ 65+}$				1.000

Os *outliers* detetados no biplot são mais fácil de serem analisados usando os gráficos de dispersão univariados da Figura 4.5.

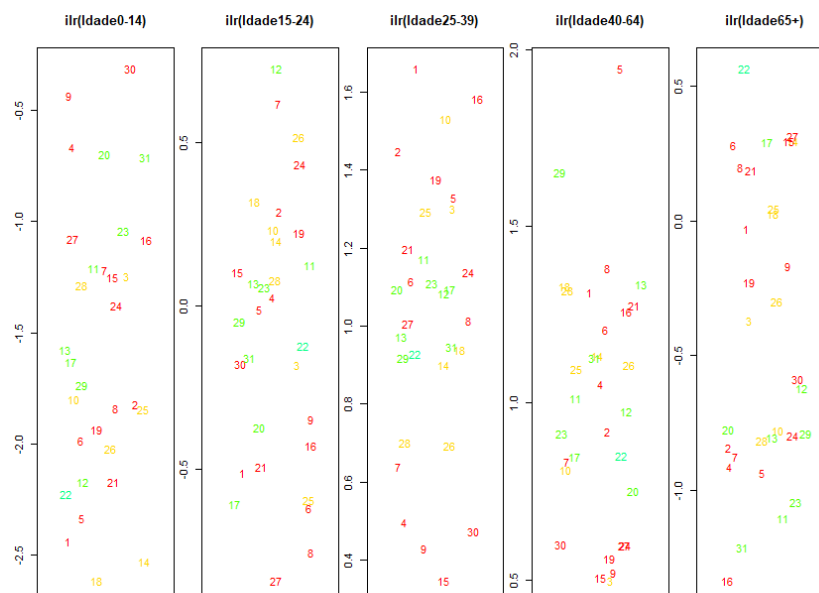


Figura 4.5: Gráficos de dispersão univariados para o conjunto de dados por grupo etário.

Os gráficos de dispersão univariados salientam os *outliers* dispersos para as diferentes variáveis em coordenadas *ilr*. Destaca-se no gráfico em *ilr(Idade 0-14)* que as observações 30, 9, 4, 20 e 31, correspondentes aos municípios de Vila Nova de Paiva, Figueira de Castelo Rodrigo, Carrazeda de Ansiães, Montalegre e Vizela, respetivamente, assumem valores altos confirmando este resultado com o biplot onde se visualiza que a variável *Idade 0-14* é dominante para estas observações. Este facto leva a concluir que estes cinco municípios foram “atrativos” pois existe uma maior percentagem de indivíduos com idades dos 0 aos 14 anos que passaram a residir nestes municípios. Por outro lado, apenas as três primeiras observações assumem valores baixos em *ilr(Idade 25-39)* indicando que estes municípios foram considerados repulsivos³ para indivíduos com idades dos 25 aos 39 anos.

Em *ilr(Idade 15-24)*, os valores elevados são atribuídos às observações 12 e 7 o que evidencia no biplot que a parte dominante destas observações é a variável *Idade 15-24*. Conclui-se que em termos de proporções nos cinco grupos etários, os municípios de Lajes das Flores e Crato, respetivamente, foram apazíveis para pessoas com idades entre os 15 e 24 anos que se deslocaram para residirem nestes municípios.

A observação 1, referente ao município de Aguiar da Beira, é um dos municípios que possui uma elevada percentagem de indivíduos com idades dos 25 aos 39 anos que passaram a habitar neste município. Porém, Aguiar da Beira tornou-se um local pouco atrativo para residir para cidadãos com idades dos 0 aos 14 anos.

Um facto interessante na Figura 4.5 surge em *ilr(Idade 40-64)* destacando que a maioria das observações atípicas tende a revelar valores médios ou baixos exceto nas observações 5 e 29, cujos valores são consideravelmente elevados. Esta análise sugere que estes municípios

³Entenda-se “repulsivo” no sentido de não atrair ou cativar residentes para os municípios.

atípicos podem ser considerados como repulsivos para pessoas com idades entre os 40 e os 64 anos. Apenas Castanheira de Pêra e Vila Nova de Foz Côa mostraram-se, em termos relativos, corresponder a localidades onde indivíduos com estas idades preferiram residir.

Em *ilr(Idade 65+)*, as observações 22 e 16 evidenciam que os municípios de Pinhel e Meda, respetivamente, são incompatíveis para pessoas com idades superiores a 65 anos. Existe uma maior percentagem de pessoas com estas idades, que entre 2005 e 2011, preferiram mover-se para o município de Pinhel para residir enquanto que uma baixa porção de indivíduos optou por se deslocar para Meda.

Efetuada o estudo de ambas as metodologias para o conjunto de dados em questão, a Tabela 4.5 evidencia os resultados obtidos relativamente às cinco funções implementadas para a deteção das observações atípicas.

Tabela 4.5: Síntese dos resultados obtidos pelas funções das metodologias numéricas e gráficas.

Funções	Número de <i>outliers</i>	Menor valor de $\ln(\text{MCD})$
<i>covMcd()</i>	34	-13.00288
<i>aq.plot()</i>	30	-13.01813
<i>dd.plot()</i>	34	-13.01813
<i>outCoDa()</i>	48	-13.16091
<i>mvoutlier.CoDa()</i>	31	-13.01813

Seleccionaram-se os *outliers* comuns às cinco funções usadas tendo-se obtido um total de 30 observações, listadas a seguir:

Municípios			
Aguiar da Beira	Figueira de Castelo Rodrigo	Melgaço	Torre de Moncorvo
Alcoutim	Freixo de Espada à Cinta	Mira	Trancoso
Arouca	Fronteira	Monchique	Vila de Rei
Carrazeda de Ansiães	Lajes das Flores	Montalegre	Vila Nova de Foz Côa
Castanheira de Pêra	Macedo de Cavaleiros	Pampilhosa da Serra	Vila Nova de Paiva
Castelo de Vide	Manteigas	Pinhel	Vizela
Crato	Marvão	Porto Moniz	
Cuba	Meda	Santa Marta de Penaguião	

Na lista seguinte encontram-se os restantes 18 municípios *outliers* que não são comuns às funções, mas que fazem parte do grupo das observações atípicas. Estes *outliers* são considerados dúbios, pois apesar de não pertencerem ao conjunto das observações comuns todos eles estão associados ao menor valor do logaritmo de MCD de cada uma das funções.

Municípios				
Almeida	Constância	Moimenta da Beira	Ribeira Grande	Viana do Castelo
Almodôvar	Figueiró dos Vinhos	Mora	Sernancelhe	Vila Flor
Arganil	Fornos de Algodres	Ponte da Barca	Tarouca	
Câmara de Lobos	Góis	Ribeira de Pena	Velas (R.A.A)	

Na Figura 4.6 visualiza-se o cartograma de Portugal Continental, destacando a cor verde os 28 municípios *outliers*, provenientes da primeira lista, para este território. Numa primeira observação salienta-se que a maioria destas observações atípicas pertencem a regiões do interior de Portugal, com exceção de Mira localizada na Beira Litoral e Monchique na região do Algarve.

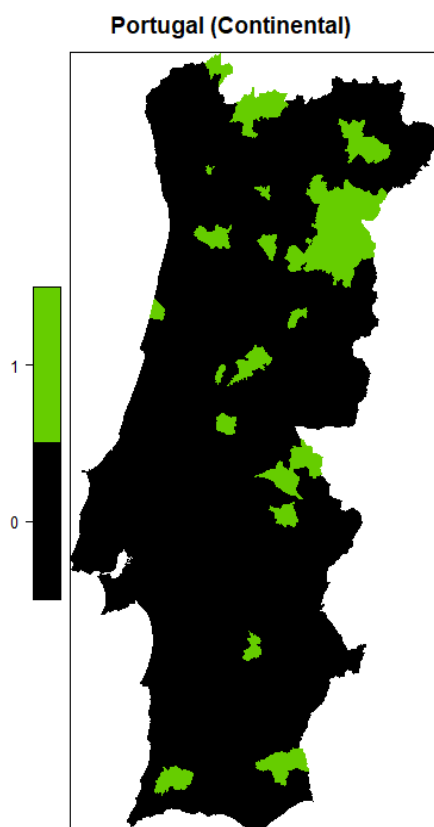


Figura 4.6: Cartograma de Portugal (Continental) com 28 municípios *outliers* a cor verde, para o conjunto de dados por grupo etário.

Face ao gráficos de dispersão univariados, da Figura 4.5, Mira e Monchique identificados pelas observações 18 e 19, respetivamente, situam-se na parte inferior do gráfico para as variáveis $ilr(Idade\ 0-14)$ e $ilr(Idade\ 40-64)$, respetivamente. Tal situação revela que para estes dois grupos etários ambos os municípios foram considerados repulsivos, pois uma pequena percentagem de cidadãos daquelas idades escolheu estes municípios para habitar.

O cartograma do Arquipélago da Madeira surge na Figura 4.7, na qual somente o município de Porto Moniz se destaca como *outlier*. Perante a interpretação da observação 23 deste município nos gráficos de dispersão univariados, esta destaca-se posicionando-se na parte inferior dos gráficos para as variáveis $ilr(Idade\ 40-64)$ e $ilr(Idade\ 65+)$. Assim, conclui-se que durante 2005 e 2011 este município tornou-se repulsivo para indivíduos com idades a partir dos 40 anos. Ou seja, na Região Autónoma da Madeira uma pequena percentagem de indivíduos moveu-se para Porto Moniz para residir neste município.

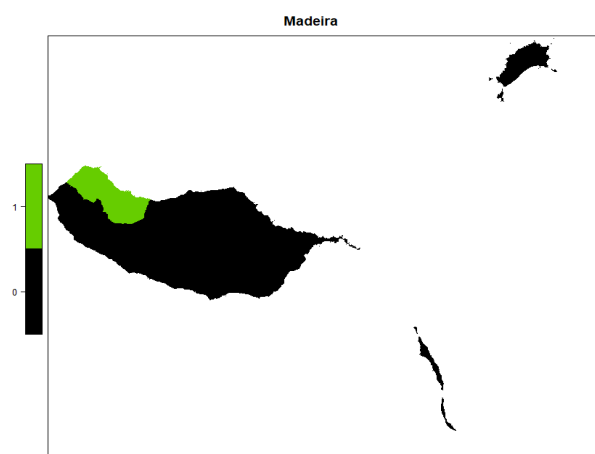


Figura 4.7: Cartograma do Arquipélago da Madeira com identificação do município *outlier* de Porto Moniz a cor verde, para o conjunto de dados por grupo etário.

A Figura 4.8 exhibe o cartograma do Arquipélago dos Açores revelando que unicamente o município de Lajes das Flores é um *outlier* para o grupo Ocidental. Além disso, nenhum outro município pertencente aos restantes grupos são *outliers*. Diante dos gráficos de dispersão univariados, este município foi interpretado segundo a observação 12 para a variável $ilr(Idade\ 15-24)$.

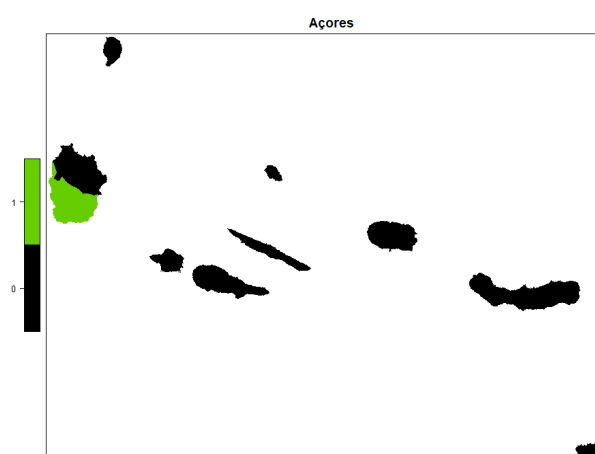


Figura 4.8: Cartograma do Arquipélago dos Açores com identificação do município *outlier* de Lajes das Flores a cor verde, no grupo Ocidental, para o conjunto de dados por grupo etário.

4.3.2 Conjunto de dados por habilitação académica

Os dados referente ao conjunto da habilitação académica tinham a particularidade de possuir componentes irregulares, nomeadamente zeros de contagem. Tal situação impedia a aplicação direta dos métodos estatísticos usuais conforme esclarecido no Capítulo 2. Deste modo, foi realizado o pré-processamento dos dados usando o algoritmo k -NN pela função *impKNNa()*, da biblioteca *robCompositions*, tendo sido a análise destes dados efetuada de acordo com a matriz composicional preenchida.

Dado que a análise se fundamenta em dados composicionais, a função *impKNNa()* tem especificidades para estudar com estes dados. Uma delas é o facto de que a distância de Aitchison é um requisito essencial no algoritmo para determinar a distância entre observações, pelo que no *input* da função se deve escolher *metric*="Aitchison". A outra incide na agregação e no ajuste dos k vizinhos mais próximos cuja escolha deve ser a mediana das observações, *agg*="median" e *adj*="median", respetivamente.

Concluindo todo o pré-processamento da matriz composicional procedeu-se com a aplicação da PBS a fim de se obter as coordenadas *ilr*-transformadas. Nas metodologias numéricas, para a função *covMcd()* o número de *outliers* foi 69 com um valor de $\ln(\text{MCD})$ de -24.58819 . Os gráficos obtidos para comparação das distâncias de Mahalanobis segundo esta função encontram-se na Figura 4.9.

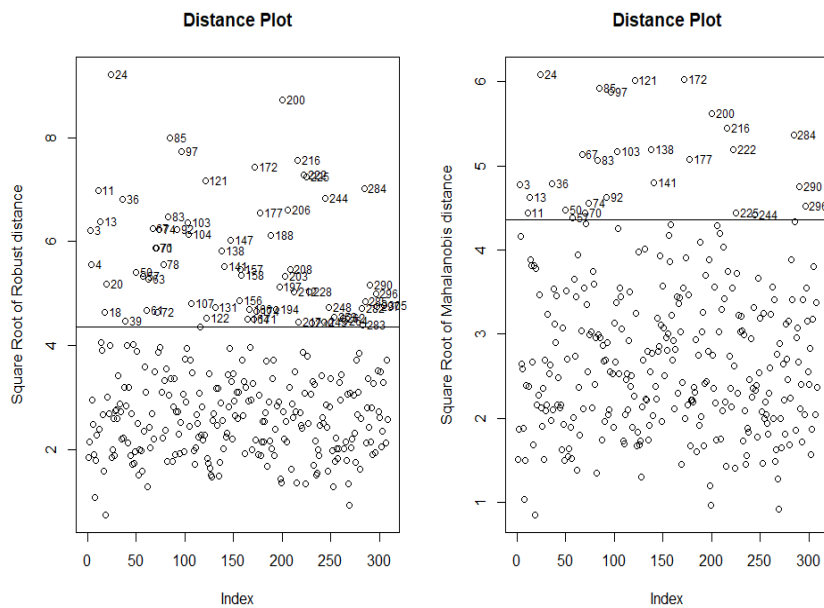


Figura 4.9: Comparação gráfica da raiz quadrada da distância de Mahalanobis robusta (gráfico da esquerda) e clássica (gráfico da direita) para o conjunto de dados por habilitação académica. As observações identificadas por números são *outliers*.

Pela visualização dos gráficos é notória uma saliente diferença das duas distâncias e, claramente, que o número de *outliers* detetado em ambas é relevante. Uma vez mais se verifica que a distância de Mahalanobis robusta para deteção de *outliers* multivariados prevalece sobre a distância de Mahalanobis clássica.

A função *aq.plot()* para os dados em foco deu como solução 62 *outliers* e um valor de $\ln(\text{MCD})$ de -24.61327 . A Figura 4.10 diferencia os *outliers* identificados pelo quantil $\chi^2_{9;0.975}$, observando uma indubitável mistura de observações atípicas e não atípicas.

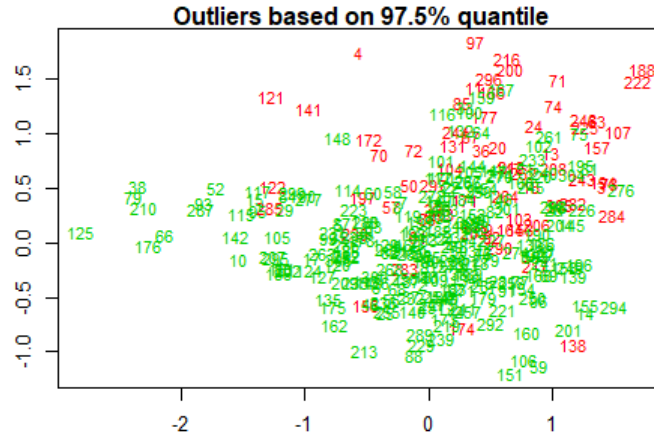


Figura 4.10: Gráfico que exibe os *outliers* (destacados a vermelho) de acordo com o quantil $\chi^2_{9;0.975}$ para o conjunto de dados por habilitação académica.

Relativamente à função *dd.plot()*, o número de *outliers* foi 69 para um valor de $\ln(\text{MCD})$ de -24.58819 , tendo-se obtido na Figura 4.11 o gráfico *D-D Plot*. Certifica-se que a solução encontrada para esta função é igual à solução de *covMcd()* e o *D-D Plot* revela que as observações tendem a distanciar-se da distância clássica para se dispersar na região da distância robusta.

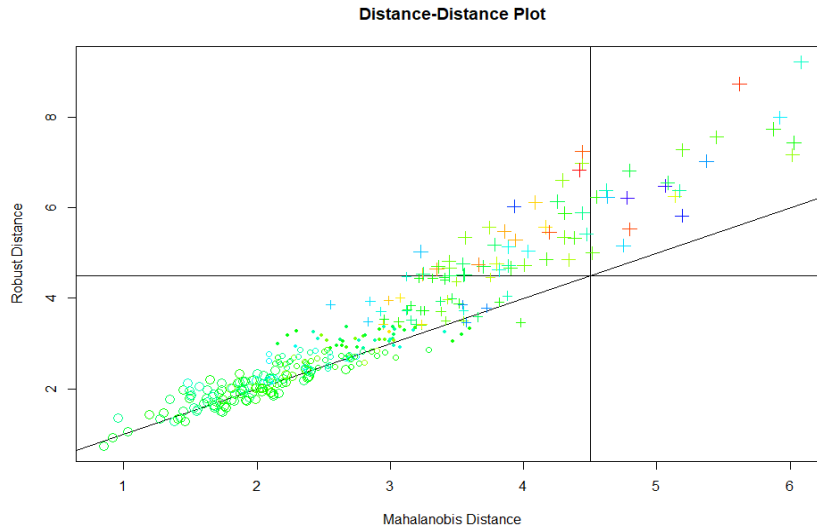


Figura 4.11: Gráfico que compara as distâncias de Mahalanobis robusta e clássica para o conjunto de dados por habilitação académica. Os símbolos “+” no 1º quadrante destacam as observações atípicas que se localizam fora do quantil $\chi^2_{9;0.975}$.

Por último, a função *outCoDa()* identificou 88 *outliers* com um valor de $\ln(\text{MCD})$ de -25.4167 .

Em relação às metodologias gráficas, a solução para a função *mvoutlier.CoDa()* deu 63 *outliers* com um valor de $\ln(\text{MCD})$ de -24.58819 . Os respectivos biplots composicionais robustos são exibidos na Figura 4.12.

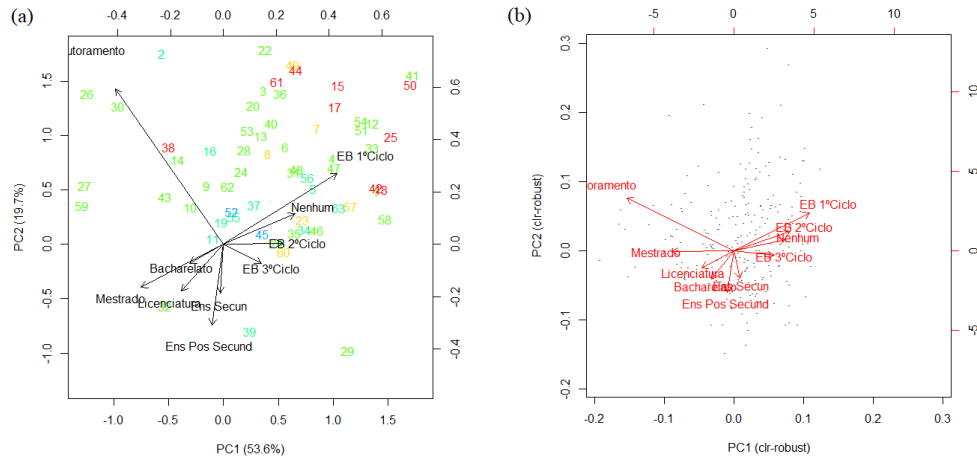


Figura 4.12: Biplots composicionais robustos (a) para visualização dos *outliers* bem como do comprimento dos *links* e (b) para visualização dos ângulos entre os *links* para o conjunto de dados por habilitação académica.

A percentagem de variabilidade total retida pelo biplot da Figura 4.12 (a) é de $53.6 + 19.7 = 73.3\%$ e pela sua visualização verifica-se que o menor *link* é demonstrado pela ligação entre as variáveis *EB 2º Ciclo* e *EB 3º Ciclo*, provado pela matriz de variação composicional da Tabela 4.6, $\tau_{3,4} = 0.062$. Tal facto realça que estas variáveis são proporcionais entre si e, por conseguinte, possuem log-razão $\ln\left(\frac{EB\ 2^\circ\ Ciclo}{EB\ 3^\circ\ Ciclo}\right)$ constante. Portanto, no conjunto dos 308 municípios, este resultado salienta que a razão entre as percentagens de cidadãos que mudaram de município com habilitação académica do Ensino Básico do 2º e 3º Ciclo se revela ser a mais constante. Quer dizer que as percentagens de residentes que mudaram de município com estas habilitações académicas devem manter-se proporcionalmente mais parecidas entre si, no conjunto dos 308 municípios.

Em contrapartida, o maior *link* advém da ligação entre as variáveis *EB 1º Ciclo* e *Doutoramento*, mas também entre as variáveis *EB 1º Ciclo* e *Mestrado*. Concluindo-se que estas duas ligações evidenciam o facto das variáveis possuírem uma variabilidade muito grande entre si. Consequentemente, a matriz de variação composicional comprova o sucedido mostrando o maior valor da matriz repetido duas vezes para as respectivas variáveis enunciadas, $\tau_{2,9} = \tau_{2,10} = 0.776$.

Assim sendo, este resultado indica que a razão entre as percentagens de residentes que mudaram de município com escolaridade no Ensino Básico do 1º Ciclo e com Doutoramento é muito variável na totalidade dos 308 municípios. Neste conjunto onde se observou mobilidade de residência dos seus habitantes, haverá municípios em que a percentagem de residentes com o Ensino Básico do 1º Ciclo será um dado valor proporcional à percentagem de residentes com Doutoramento mas tal valor de proporcionalidade diferirá entre os diferentes 308 municípios, sendo que para este par de habilitações académicas observa-se a maior variabilidade. De forma análoga, a mesma análise é efetuada para as variáveis *EB 1º Ciclo* e *Mestrado*.

Tabela 4.6: Matriz de variação composicional entre as variáveis das habilitações académicas (a negrito destacam-se os valores do maior e menor *link*).

	Nenhum	EB 1º Ciclo	EB 2º Ciclo	EB 3º Ciclo	Ens Secun	Ens Pos Secun	Bacharelato	Licenciatura	Mestrado	Doutoramento
Nenhum	0	0.076	0.097	0.100	0.189	0.395	0.373	0.294	0.571	0.582
EB 1º Ciclo		0	0.162	0.211	0.361	0.586	0.549	0.468	0.776	0.776
EB 2º Ciclo			0	0.062	0.155	0.327	0.361	0.218	0.478	0.559
EB 3º Ciclo				0	0.066	0.234	0.249	0.159	0.399	0.481
Ens Secun					0	0.172	0.160	0.104	0.306	0.428
Ens Pos Secun						0	0.295	0.204	0.324	0.599
Bacharelato							0	0.191	0.369	0.451
Licenciatura								0	0.179	0.368
Mestrado									0	0.530
Doutoramento										0

Relativamente aos ângulos entre os *links* é notório na Figura 4.12 (b) um ângulo de 90° entre dois conjuntos de variáveis. O primeiro conjunto é constituído pelas variáveis *Ens. Secundário* e *Doutoramento*, sendo o segundo composto pelas variáveis *Ens Secundário*, *EB 3º Ciclo*, *EB 2º Ciclo*, *Nenhum* e *EB 1º Ciclo*. Neste último faz-se a união dos quatro *links* das variáveis obtendo-se um único *link*. De seguida, verifica-se que ao unir este *link* com o *link* do primeiro conjunto existe um ângulo reto pelo que as variáveis são não correlacionadas. A submatriz de correlações da Tabela 4.7 prova tal conclusão, demonstrando na primeira linha valores muito próximos de zero.

Deste modo é admissível constatar que não havendo correlação entre a log-razão destas variáveis, então existem padrões discrepantes de deslocação dos indivíduos que apresentam estes diferentes graus de escolaridade. Por exemplo, a mobilidade da razão entre as percentagens de pessoas que mudaram de município com escolaridade no Ensino Secundário e com Doutoramento é distinta da razão entre as percentagens de pessoas que mudaram de município com escolaridade no Ensino Básico do 1º Ciclo e sem qualquer habilitação académica.

Um ângulo de 0° é projetado entre os quatro *links* das variáveis *Ens. Pos Secundário*, *EB 3º Ciclo*, *EB 2º Ciclo*, *Nenhum* e *EB 1º Ciclo*, revelando que estas estão correlacionadas. Contudo, na submatriz de correlações os valores 0.130, 0.096 e 0.302 da última coluna não são elevados, próximos de 1, logo tal conclusão não pode ser diretamente provada pela submatriz. Tal ocorrência poderá novamente dever-se às percentagens de variabilidade dos dados retidas pelas duas primeiras componentes principais no biplot.

Este resultado salienta o facto de que a razão entre as percentagens de pessoas que mudaram de município com habilitações académicas no Ensino Pos Secundário, Ensino Básico do 1º, 2º e 3º Ciclo e sem nenhuma habilitação académica apresentam padrões similares de deslocação que os levou a habitar noutros municípios.

Tabela 4.7: Uma submatriz de correlações entre log-razões e partes na habilitação académica.

	$\ln \frac{Ens\ Secun}{Doutoramento}$	$\ln \frac{EB\ 3^\circ\ Ciclo}{Ens\ Secun}$	$\ln \frac{EB\ 2^\circ\ Ciclo}{EB\ 3^\circ\ Ciclo}$	$\ln \frac{Nenhum}{EB\ 2^\circ\ Ciclo}$	$\ln \frac{EB\ 1^\circ\ Ciclo}{Nenhum}$	$\ln \frac{EB\ 3^\circ\ Ciclo}{Ens\ Pos\ Secun}$
$\ln \frac{Ens\ Secun}{Doutoramento}$	1.000	-0.038	-0.035	-0.028	0.060	-0.018
$\ln \frac{EB\ 3^\circ\ Ciclo}{Ens\ Secun}$		1.000	0.212	-0.024	0.432	0.515
$\ln \frac{EB\ 2^\circ\ Ciclo}{EB\ 3^\circ\ Ciclo}$			1.000	0.379	0.335	0.130
$\ln \frac{Nenhum}{EB\ 2^\circ\ Ciclo}$				1.000	-0.062	0.096
$\ln \frac{EB\ 1^\circ\ Ciclo}{Nenhum}$					1.000	0.302
$\ln \frac{EB\ 3^\circ\ Ciclo}{Ens\ Pos\ Secun}$						1.000

Para complementar a análise dos *outliers* identificados pelo biplot recorreu-se aos gráficos de dispersão univariados da Figura 4.13.

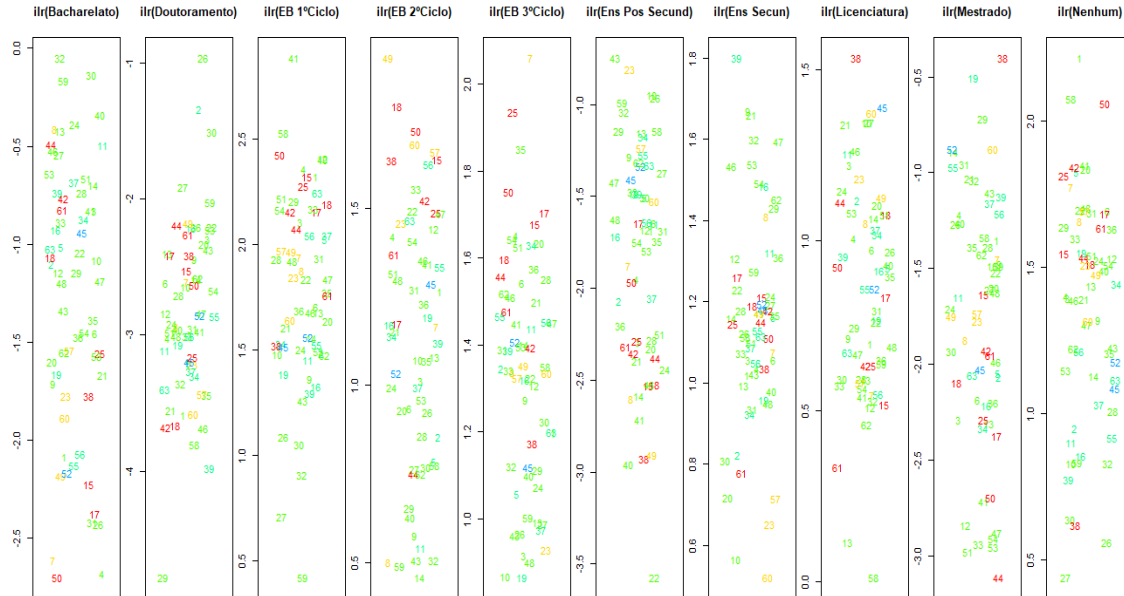


Figura 4.13: Gráficos de dispersão univariados para o conjunto de dados por habilitação académica.

Para um total de 63 observações atípicas, alguns dos gráficos de dispersão univariados tornam-se um pouco imprecisos de interpretar. De forma sucinta há uma tendência para os *outliers* assumirem uma percentagem elevada em *ilr(Bacharelato)* e *ilr(Ens. Pos Secund)*. A maior parte deles tende a permanecer em valores médios das variáveis *ilr* e uma parte dos *outliers* agrupa-se em valores baixos das variáveis *ilr(EB 2º Ciclo)*, *ilr(EB 3º Ciclo)* e *ilr(Mestrado)*.

Em *ilr(Licenciatura)* e *ilr(Mestrado)*, o município de Nordeste é considerado atrativo pois existe uma elevada percentagem de pessoas com habilitações académicas de Licenciatura e Mestrado que passaram a residir neste município, assinalada pela observação 38. Em contrapartida, esta observação assume valores baixos em *ilr(Nenhum)* indicando que o município tornou-se repulsivo para indivíduos que não tinham nenhuma habilitação académica e, por isso, uma pequena percentagem de pessoas escolheu Nordeste para residir.

Em *ilr(Bacharelato)*, a observação 50 possui valores baixos contrastando com *ilr(Nenhum)*. De 2005 para 2011 houve uma pequena porção de pessoas com Bacharelato que se deslocaram para Ribeira de Pena a fim de habitar neste município. Enquanto que durante esses seis anos uma grande parte das pessoas com nenhuma habilitação académica optou por passar a residir neste município.

A observação 59, alusiva ao município de Vila do Bispo, tem um destaque interessante em algumas das variáveis *ilr*. É de realçar que esta observação tem uma elevada percentagem em *ilr(Bacharelato)* e em *ilr(Ens. Pos Secund)*, porém dispõe de pequenas percentagens em *ilr(EB 1º Ciclo)*, *ilr(EB 2º Ciclo)* e *ilr(EB 3º Ciclo)*. Tais conclusões auxiliam a perceber que durante 2005 e 2011 este município foi cativante para indivíduos que tiraram um Bacharelato

e fizeram o Ensino Pos Secundário passando a residir neste município. Todavia, Vila do Bispo tornou-se repulsivo para cidadãos com escolaridade do Ensino Básico do 1º, 2º e 3º Ciclo.

Da visualização do biplot destacam-se as observações 39 e 29 que se afastam do conjunto de *outliers*. Respetivamente correspondem aos municípios de Odemira e Manteigas demonstrando que no gráfico de dispersão univariado, em *ilr(Doutoramento)*, ambas situam-se na parte inferior do gráfico assumindo valores baixos. Deste modo conclui-se que uma pequena parte de indivíduos com Doutoramento optaram por mover-se para estes municípios com a intenção de residirem. Realça-se que a observação 29 é a que se encontra na parte mais funda do gráfico salientando que neste município a percentagem de pessoas que se deslocaram para residir em Manteigas é mais pequena que na observação 39.

Sob outra perspetiva observa-se no biplot que as variáveis *Ens. Secundário* e *Ens. Pos Secundário* são dominantes para as duas observações anteriores, supondo que para os gráficos de dispersão univariados ambas assumam valores elevados tanto em *ilr(Ens Secund)* como em *ilr(Ens Pos Secund)*. Contudo, na Figura 4.13 há uma separação destas observações nas duas variáveis *ilr*. Em *ilr(Ens Pos Secund)* a observação 29 situa-se na parte superior do gráfico enquanto a 39 posiciona-se para baixo (sobreposta com outras observações). Assim é possível concluir que uma maior percentagem de indivíduos com Ensino Pos Secundário preferiu mudar-se para Manteigas a fim de habitar neste município. Em *ilr(Ens Secun)* verifica-se o contrário, a observação 39 assume valores elevados enquanto a 29 localiza-se a meio do gráfico. Do mesmo modo se conclui que indivíduos com escolaridade no Ensino Secundário preferiram mover-se para Odemira com a finalidade de residir neste município.

Realizado o estudo de ambas as metodologias para o presente conjunto de dados, a Tabela 4.8 mostra os resultados obtidos no que diz respeito às cinco funções implementadas para a deteção dos grupos das observações atípicas.

Tabela 4.8: Síntese dos resultados obtidos pelas funções das metodologias numéricas e gráficas.

Funções	Número de <i>outliers</i>	Menor valor de $\ln(\text{MCD})$
<i>covMcd()</i>	69	-24.58819
<i>aq.plot()</i>	62	-24.61327
<i>dd.plot()</i>	69	-24.58819
<i>outCoDa()</i>	88	-24.41670
<i>mvoutlier.CoDa()</i>	63	-24.58819

Para o estudo dos *outliers* selecionou-se as observações atípicas comuns às funções das metodologias e obteve-se um total de 62 observações. Os municípios *outliers* apresentam-se listados a seguir. Existem, ainda, 26 *outliers* dúbios que apesar de serem *outliers* não são comuns às funções, estando associados ao menor valor do logaritmo de MCD de cada uma delas, os quais estão listados após os primeiros.

Lista total dos 62 municípios *outliers* comuns às 5 funções do conjunto de dados por habilitação académica.

Municípios				
Aguiar da Beira	Castelo de Vide	Lajes do Pico	Pampilhosa da Serra	Sernancelhe
Alandroal	Castro Daire	Mação	Penalva do Castelo	Sever do Vouga
Alcoutim	Castro Marim	Manteigas	Penela	Tarouca
Alfândega da Fé	Celorico da Beira	Marvão	Pinhel	Vieira do Minho
Almeida	Cinfães	Mesão Frio	Ponta do Sol	Vila de Rei
Almodôvar	Corvo	Monchique	Ponte de Sor	Vila do Bispo
Alvito	Crato	Mondim de Basto	Portel	Vila Franca do Campo
Arronches	Estremoz	Monforte	Porto Moniz	Vila Nova de Paiva
Borba	Ferreira do Alentejo	Mortágua	Povoação	Vila Nova de Poiares
Calheta (R.A.A.)	Freixo de Espada à Cinta	Mourão	Ribeira de Pena	Vinhais
Campo Maior	Fronteira	Nordeste	Sabrosa	
Carrazeda de Ansiães	Gavião	Odemira	Santa Comba Dão	
Castanheira de Pêra	Lajes das Flores	Oleiros	Sardoal	

Lista dos 26 municípios *outliers* que não são comuns às 5 funções do conjunto de dados por habilitação académica.

Municípios			
Alvaiázere	Figueira de Castelo Rodrigo	Nisa	Terras de Bouro
Avis	Fornos de Algodres	Ourique	Valpaços
Baião	Góis	Pedrógão Grande	Viana do Alentejo
Belmonte	Idanha-a-Nova	Proença-a-Nova	Vila da Praia da Vitória
Boticas	Mértola	Sabugal	Vila Velha de Ródão
Castro Verde	Montalegre	Santa Cruz da Graciosa (R.A.A.)	
Celorico de Basto	Murça	São Vicente	

O cartograma de Portugal Continental é exibido na Figura 4.14, salientando a cor verde os 54 municípios *outliers*, referentes na primeira lista, para este território nacional. Do mapa visualiza-se que estes municípios encontram-se dispersos pelas várias regiões, afastando-se da zona costeira a Norte e Centro de Portugal. A interpretação de alguns destes *outliers* foi realizada anteriormente de acordo com a Figura 4.13.

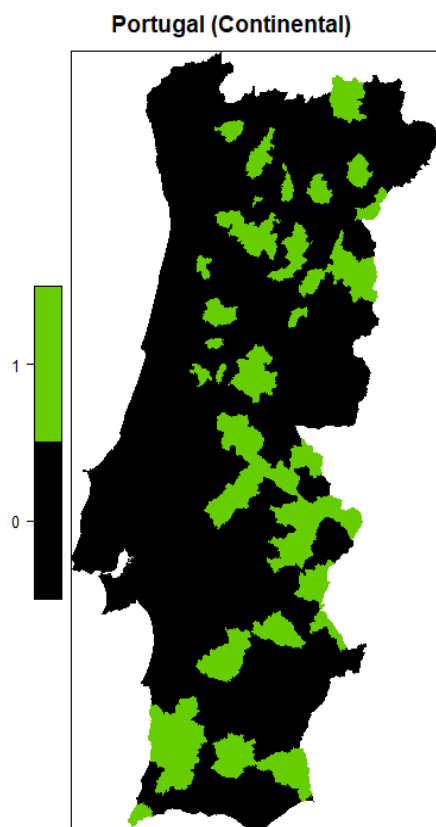


Figura 4.14: Cartograma de Portugal (Continental) com 54 municípios *outliers* a cor verde, para o conjunto de dados por habilitação académica.

A Figura 4.15 apresenta o cartograma do Arquipélago da Madeira expondo dois municípios *outliers*, nomeadamente Porto Moniz e Ponta do Sol. Da Figura 4.13, estes municípios são indicados pelas observações 48 e 45, respetivamente. Ambos posicionam-se em *ilr(EB 3º Ciclo)* na parte inferior do gráfico, contudo Porto Moniz apresenta uma percentagem menor de indivíduos com escolaridade no Ensino Básico do 3º Ciclo que se deslocaram para residir neste município. Relativamente à observação 45, esta surge na parte superior em *ilr(Ens Pos Secund)* e em *ilr(Licenciatura)*, porém a percentagem de indivíduos com uma Licenciatura que se moveram para habitar neste município é mais elevada do que a percentagem de indivíduos com escolaridade no Ensino Pos Secundário.

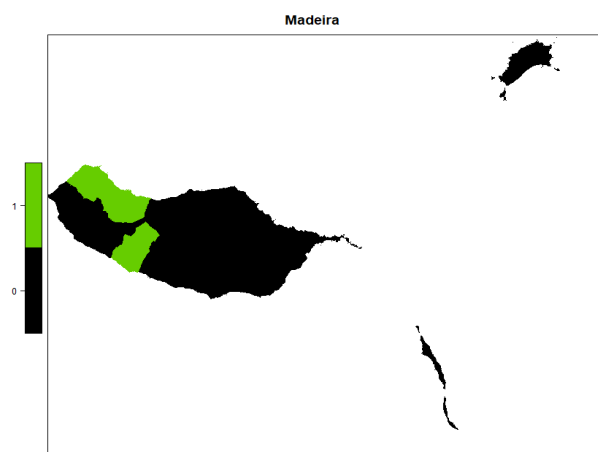


Figura 4.15: Cartograma do Arquipélago da Madeira com identificação dos municípios *outliers* de Porto Moniz e Ponta do Sol a cor verde, para o conjunto de dados por habilitação académica.

Para a Região Autónoma dos Açores, os três grupos, Ocidental, Central e Oriental, que constituem este Arquipélago possuem todos pelo menos um *outlier*. No grupo Ocidental tem-se os municípios do Corvo e Lajes das Flores, no grupo Central o município de Lajes do Pico e no grupo Oriental os municípios de Nordeste, Povoação e Vila Franca do Campo.



Figura 4.16: Cartograma do Arquipélago dos Açores com identificação a cor verde dos municípios *outliers* de Corvo e Lajes das Flores no grupo Ocidental, do município Lajes do Pico no grupo Central e dos municípios Nordeste, Povoação e Vila Franca do Campo no grupo Oriental, para o conjunto de dados por habilitação académica.

Os municípios de Lajes das Flores (observação 26) e do Pico (observação 27) apresentam valores baixos em $ilr(Nenhum)$ do gráfico da Figura 4.13, no entanto a percentagem de indivíduos que se moveu para o município de Lajes do Pico sem qualquer habilitação académica é bastante pequena comparando com a percentagem para Lajes das Flores. De igual modo, a observação 26 também apresenta valores baixos em $ilr(Bacharelato)$ e $ilr(EB\ 3^\circ\ Ciclo)$ e a observação 27 em $ilr(EB\ 1^\circ\ Ciclo)$. Mas, a percentagem de pessoas que se moveu para o município de Lajes das Flores é muito elevada em $ilr(Doutoramento)$. Corvo (observação 19) e

Povoação (observação 49) são os dois municípios que apresentam a percentagem mais elevada de cidadãos com Mestrado e com escolaridade no Ensino Básico do 2º Ciclo, respetivamente, que se deslocaram para residir nestes municípios. Vila Franca do Campo (observação 60) é o município que possui a percentagem mais baixa de pessoas com escolaridade no Ensino Secundário que passaram a morar neste município. Nordeste foi analisado previamente.

4.3.3 Conjunto de dados por situação profissional

Procedeu-se com a leitura da matriz do conjunto de dados por situação profissional que continha a informação absoluta e fez-se a elaboração da matriz composicional com a informação relativa. De seguida aplicou-se a PBS obtendo-se as coordenadas *ilr*-transformadas, e iniciou-se o estudo das metodologias numéricas.

A função *covMcd()* revelou a existência de 11 *outliers* para um valor de $\ln(\text{MCD})$ de -5.518727 , sendo os respetivos gráficos das distâncias de Mahalanobis exibidos na Figura 4.17.

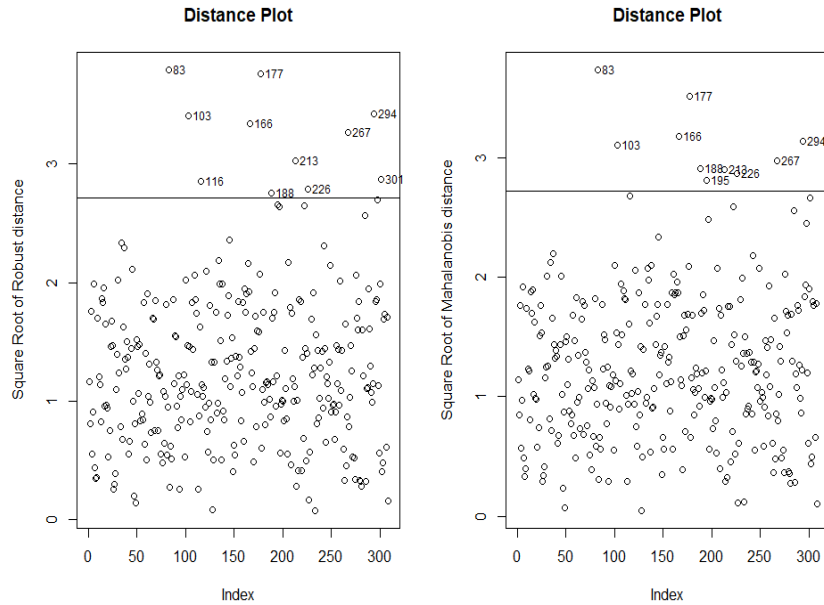


Figura 4.17: Comparação gráfica da raiz quadrada da distância de Mahalanobis robusta (gráfico da esquerda) e clássica (gráfico da direita) para o conjunto de dados por situação profissional. As observações identificadas por números são *outliers*.

Os gráficos demonstram uma diferença quase inexistente entre as duas distâncias. Observa-se que a distância de Mahalanobis robusta deteta 11 observações atípicas enquanto a clássica revela 10 observações. Ambas possuem 9 *outliers* em comum onde apenas três deles não são comuns às distâncias, as observações 116 e 301 são detetadas pela distância robusta e a observação 195 pela distância clássica. Deste modo, constata-se que a distância robusta predomina sobre a clássica.

A função *aq.plot()* identificou a mesma solução encontrada do comando anterior e a Figura 4.18 exibe as observações atípicas detetadas pelo quantil $\chi^2_{2;0.975}$. Um facto interessante deste gráfico é que os *outliers* localizam-se na região periférica das observações não atípicas não

revelando junção de observações, contrastando com os dois gráficos homólogos obtidos nas subsecções anteriores.

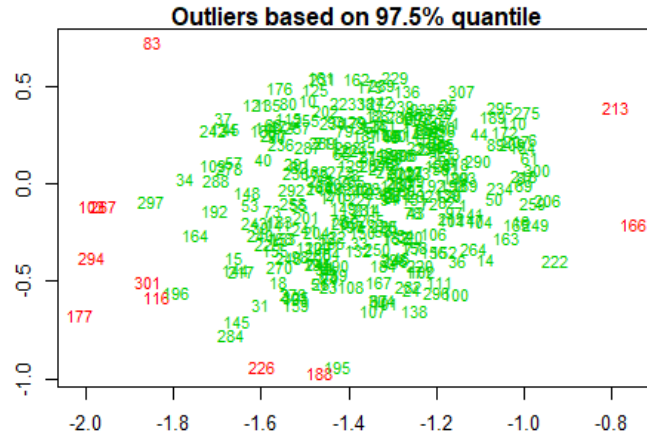


Figura 4.18: Gráfico que exhibe os *outliers* (destacados a vermelho) de acordo com o quantil $\chi^2_{2;0.975}$ para o conjunto de dados por situação profissional.

Para a função *dd.plot()* obteve-se 14 *outliers* e o valor do logaritmo de MCD foi exatamente igual aos dois comandos anteriores. No gráfico *D-D Plot*, na linha a tracejado, as observações tendem a permanecer sobre a linha aproximadamente até metade, mas de seguida separam-se e enquanto algumas mantêm-se no espaço da distância robusta três delas afastam-se para a divisão da distância clássica.

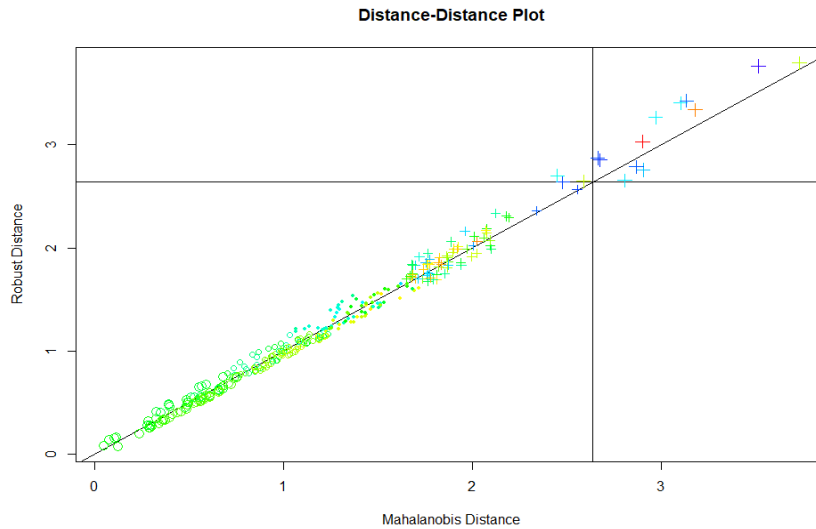


Figura 4.19: Gráfico que compara as distâncias de Mahalanobis robusta e clássica para o conjunto de dados por situação profissional. Os símbolos “+” no 1º quadrante destacam as observações atípicas que se localizam fora do quantil $\chi^2_{2;0.975}$.

Em último, para a função *outCoDa()* o número de *outliers* detetado foi 19 com um valor de $\ln(\text{MCD})$ de -5.539568 .

Quanto às metodologias gráficas, a função *mvoutlier.CoDa()* identificou 15 *outliers* para um valor de $\ln(\text{MCD})$ de -5.518727 . É de destacar que, para encontrar o menor valor do logaritmo do determinante da matriz de covariância dos dados, nesta função esse valor foi exatamente o mesmo em todas as dez vezes que houve repetição da função. Por conseguinte, o código não foi alterado tal como ocorreu para os outros dois conjuntos de dados. Os biplots composicionais robustos surgem na Figura 4.20.

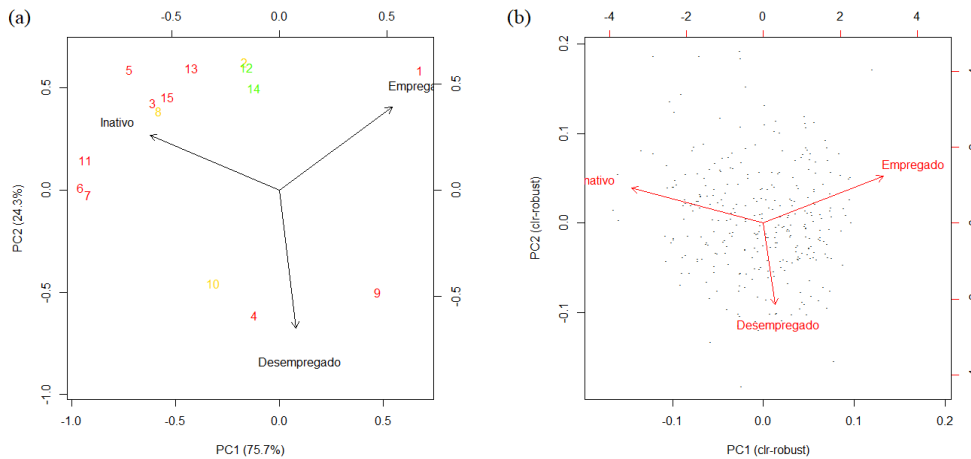


Figura 4.20: Biplots composicionais robustos (a) para visualização dos *outliers* bem como do comprimento dos *links* e (b) para visualização dos ângulos entre os *links* para o conjunto de dados por situação profissional.

Observe-se que a percentagem de variabilidade dos dados com as duas primeiras componentes principais é 100% em ambos os biplots, o que seria de esperar porque a matriz dos dados é constituída por três partes restringida à soma ser um. Logo, o espaço dos dados está contido em \mathbb{R}^2 .

Por visualização do biplot na Figura 4.20 (a) é perceptível que os três *links* têm comprimentos muito similares. Recorrendo diretamente à matriz de variação composicional, exibida na Tabela 4.9, observa-se que a variabilidade das três possíveis log-razões são todas baixas.

Tabela 4.9: Matriz de variação composicional entre as variáveis da situação profissional (a negrito destacam-se os valores do menor e maior *link*, respetivamente).

	<i>Empregado</i>	<i>Desempregado</i>	<i>Inativo</i>
<i>Empregado</i>	0	0.093	0.206
<i>Desempregado</i>		0	0.134
<i>Inativo</i>			0

Comparando os valores da matriz de variação composicional, o menor é entre as variáveis *Empregado* e *Desempregado* com um valor de $\tau_{1,2} = 0.093$, o que evidencia que estas variáveis são proporcionais entre si e, por isso, possuem log-razão $\ln\left(\frac{\text{Empregado}}{\text{Desempregado}}\right)$ constante. Assim, a razão entre as percentagens de residentes que mudaram de municípios estando empregados e desempregados reflete ser a mais constante no conjunto dos 308 municípios. Isto é, as

percentagens de residentes que mudaram de município, encontrando-se nas respetivas situações profissionais enunciadas, devem manter-se proporcionalmente mais parecidas entre si no conjunto dos 308 municípios.

Por outro lado, o maior valor da matriz de variação é entre as variáveis *Empregado* e *Inativo* com um valor de $\tau_{1,3} = 0.206$. Tal facto vem demonstrar que estas duas variáveis têm uma variabilidade maior que a log-razão referida anteriormente. Como consequência, significa que a razão entre as percentagens de residentes que mudaram de município apresentando-se empregados e inativos é mais variável no total dos 308 municípios. Neste conjunto, onde se observou mobilidade de residência dos seus habitantes, haverá municípios em que a percentagem de residentes empregados será um dado valor proporcional à percentagem de residentes inativos mas tal valor de proporcionalidade diferirá entre os diferentes 308 municípios, sendo que para este par de variáveis da situação profissional constata-se a maior variabilidade.

Do biplot da Figura 4.20 (b) não se observam ângulos entre *links* de 90° ou de 0° pelo que da análise do biplot não se deduzem interpretações de existência de correlações entre log-razões fortes ou quase nulas, respetivamente. Contudo, calculando a submatriz de correlações entre as log-razões possíveis das três partes que constituem o conjunto de dados (Tabela 4.10) observa-se que a submatriz de correlações exibe um único valor bastante próximo de zero, 0.094, entre $\ln\left(\frac{Empregado}{Desempregado}\right)$ e $\ln\left(\frac{Inativo}{Desempregado}\right)$. Tal sugere que relativamente à percentagem de desempregados, a percentagem de empregados e de inativos que mudam de município, bem como de residência, não se correlacionam. Este resultado leva a crer a mobilidade (em termos de novo município para residir) inerente aos empregados e inativos, comparativamente com os desempregados, resulta de diferentes perspetivas, onde se deduz que os fundamentos que impulsionam os indivíduos empregados e inativos a mudar de município, bem como de residência, serão distintos. Por sua vez, -0.744 é o valor da submatriz de correlações com maior valor absoluto (diferente de 1) implicando que existe uma correlação relativamente forte positiva entre $\ln\left(\frac{Empregado}{Inativo}\right)$ e $\ln\left(\frac{Desempregado}{Inativo}\right)$ ($= -\ln\left(\frac{Inativo}{Desempregado}\right)$). Assim, relativamente à percentagem de inativos, a percentagem de empregados e de desempregados que mudam de município, bem como de residência, está correlacionada positivamente. Este resultado leva a crer que tal mobilidade inerente aos empregados e aos desempregados, comparativamente com a dos inativos, se relacionam entre si, sugerindo que a mudança de município bem como de residência dos indivíduos desempregados nos 308 municípios, relativamente à dos inativos, poderá ser influenciada positivamente pela mobilidade dos indivíduos empregados.

Tabela 4.10: Uma submatriz de correlações entre log-razões e partes na situação profissional.

	$\ln \frac{Empregado}{Desempregado}$	$\ln \frac{Inativo}{Desempregado}$	$\ln \frac{Empregado}{Inativo}$
$\ln \frac{Empregado}{Desempregado}$	1.000	0.094	0.596
$\ln \frac{Inativo}{Desempregado}$		1.000	-0.744
$\ln \frac{Empregado}{Inativo}$			1.000

Em relação aos *outliers* presentes no biplot, para a interpretação destes analisou-se os gráficos de dispersões univariados da Figura 4.21.

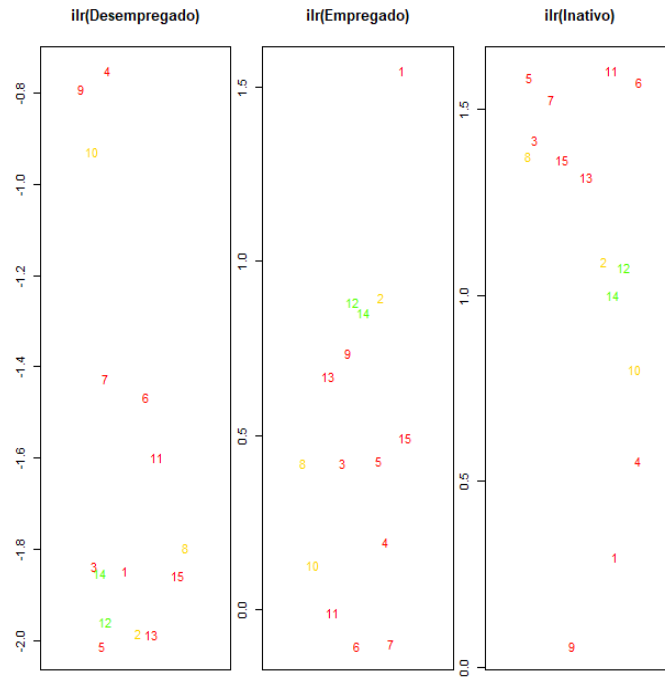


Figura 4.21: Gráficos de dispersão univariados para o conjunto de dados por situação profissional.

Dado que a função *mvoutlier.CoDa()* identificou apenas 15 *outliers*, os gráficos de dispersão univariados tornam-se acessíveis de interpretar e são muito menos confusos. Da sua visualização destacam-se as observações 4, 9 e 10 em *ilr(Desempregado)* que se situam na parte superior do gráfico apresentando valores altos. Pelo biplot observa-se que a variável *Desempregado* é dominante para estas observações. Assim, conclui-se que os municípios de Mourão, Porto Santo e Ribeira de Pena, respetivamente, foram aprazíveis uma vez que existe uma elevada percentagem de indivíduos desempregados que passaram a residir nestes municípios. A maior parte das restantes oito observações, em *ilr(Desempregado)*, tendem a situar-se na parte inferior do gráfico evidenciando que estes oito municípios *outliers* são repulsivos para indivíduos desempregados.

Em *ilr(Empregado)* exclusivamente a observação 1, referente ao município do Corvo, assume valores elevados pelo que existe uma elevada percentagem de pessoas empregadas que moveram-se para este município entre 2005 e 2011 com a finalidade de residirem. Tal conclusão confirma-se com o biplot, pois a variável *Empregado* é somente dominante para a observação 1. As observações 4, 10, 11, 6 e 7 evidenciam valores baixos em *ilr(Empregado)* constatando-se que uma pequena porção de indivíduos empregados passou a habitar nos municípios de Mourão, Ribeira de Pena, Sabugal, Pampilhosa da Serra e Penamacor.

Por fim, em *ilr(Inativo)* há uma clara disposição da maioria das observações presente na parte superior do gráfico, destacando que nestas a variável *Inativo* se torna dominante, tal como mostra o biplot. Assim, entre 2005 e 2011 os indivíduos que não possuíam qualquer

contrato de trabalho deslocaram-se para estes municípios atrativos passando a habitar. Em contrapartida, os municípios de Mourão, Corvo e Porto Santo, exibidos pelas observações 4, 1 e 9, foram considerados repulsivos pois uma pequena percentagem de indivíduos inativos escolheu estes municípios para residir.

Para o conjunto de dados desta subsecção existem apenas três variáveis representativas dos dados, tornando assim possível a exibição num diagrama ternário deste conjunto. Para tal recorreu-se à função *ternaryDiag()*, da biblioteca *robCompositions*. Esta função tem algumas particularidades que o analista pode escolher para representar os seus dados. Uma delas é delinear uma elipse em torno do conjunto, *line="ellipse"*, e definir o parâmetro que determina a elipse. Para a deteção de *outliers* multivariados o parâmetro escolhido foi *tol=0.975*. O gráfico da Figura 4.22 é o diagrama ternário para o conjunto de dados por situação profissional.

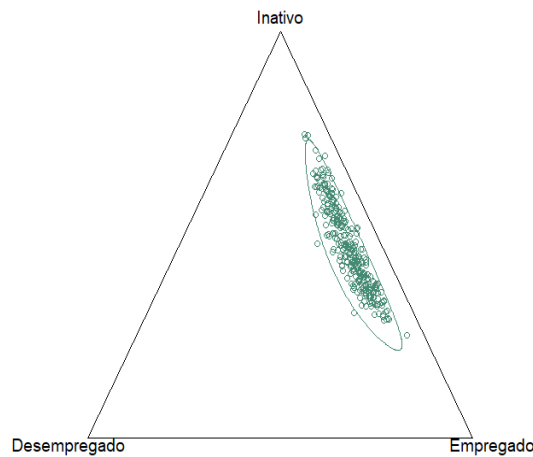


Figura 4.22: Diagrama ternário para o conjunto de dados por situação profissional.

Um dos principais problemas dos diagramas ternários é o facto de que a mente humana está apenas acostumada a pensar em termos da distância Euclidiana, e não na distância de Aitchison que precisa de ser considerada quando se olha para dados num diagrama ternário. Por isso, a interpretação do diagrama será tida em conta apenas com as características enunciadas na Subsecção 3.2.1.

Pela visualização do diagrama, os dados exibem a tendência de se alinhar entre as variáveis *Empregado* e *Inativo*, formando um padrão linear paralelo à aresta direita do gráfico. Tal situação condiz com as características (ii) e (iv) exibidas nas Figuras 3.4 (b) e 3.5 (a), indicando a dominância das componentes associadas à aresta direita o que se conclui que as proporções da componente *Desempregado* nas composições são constantes. Observa-se, ainda, a presença de *outliers* que se situam no lado exterior da elipse. Porém, a função não permite saber quais são essas observações.

Finalizado o estudo de ambas as metodologias para os dados da situação profissional, a Tabela 4.11 revela os resultados obtidos acerca das cinco funções implementadas para a deteção de *outliers*.

Tabela 4.11: Síntese dos resultados obtidos pelas funções das metodologias numéricas e gráficas.

Funções	Número de <i>outliers</i>	Menor valor de $\ln(\text{MCD})$
<i>covMcd()</i>	11	-5.518727
<i>aq.plot()</i>	11	-5.518727
<i>dd.plot()</i>	14	-5.518727
<i>outCoDa()</i>	19	-5.539568
<i>mvoutlier.CoDa()</i>	15	-5.518727

Para a interpretação das observações atípicas obteve-se 11 municípios *outliers* comuns às funções das metodologias, tal como mostra a lista seguinte:

Municípios		
Corvo	Oleiros	Torre de Moncorvo
Freixo de Espada à Cinta	Pampilhosa da Serra	Vila Nova de Foz Côa
Idanha-a-Nova	Porto Santo	Vila Velha de Ródão
Mourão	Sabugal	

A listagem a seguir revela os 8 *outliers* dúbios associados ao menor valor do logaritmo de MCD de cada uma das funções.

Municípios	
Arouca	Ribeira de Pena
Melgaço	São Roque do Pico
Penamacor	Vila de Rei
Penedono	Vila Nova de Poiares

A Figura 4.23 demonstra o cartograma de Portugal Continental revelando os 9 municípios *outliers* deste território, presentes na primeira lista. Observa-se, claramente, que estes municípios localizam-se em regiões do interior de Portugal, nomeadamente Trás-os-Montes e Alto Douro, Beira Interior e Alentejo.

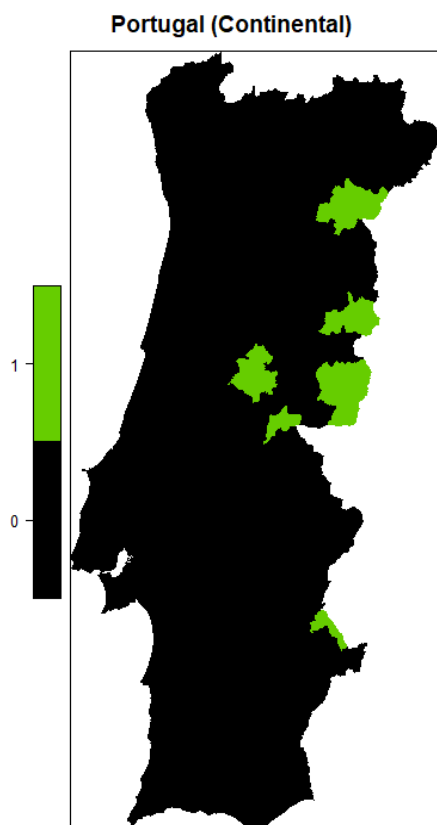


Figura 4.23: Cartograma de Portugal (Continental) com 9 municípios *outliers* a cor verde, para o conjunto de dados por situação profissional.

Transpondo a análise destes nove municípios para os gráficos de dispersão univariados, da Figura 4.21, salienta-se que a maioria deles admitem posicionamentos semelhantes para as diferentes variáveis *ilr*. Destes nove municípios, somente em *ilr(Empregado)* é que Mourão (observação 4) tem comportamento similar aos restantes, sendo que em *ilr(Desempregado)* e *ilr(Inativo)* este sobressai-se com um comportamento contrário.

Para o Arquipélago da Madeira, o cartograma é exibido na Figura 4.24 onde exclusivamente o município de Porto Santo é considerado um *outlier*. A interpretação desta observação, tendo em conta os dados relativos à situação profissional, foi discutida anteriormente.

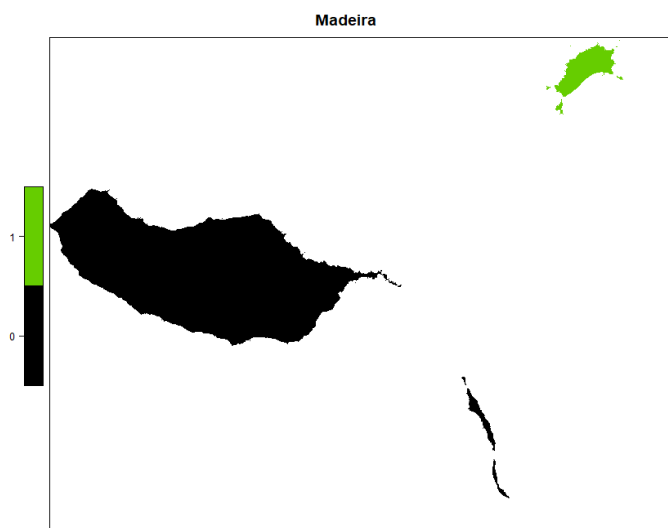


Figura 4.24: Cartograma do Arquipélago da Madeira com identificação do município *outlier* de Porto Santo a cor verde, para o conjunto de dados por situação profissional.

A Figura 4.25 mostra o cartograma do Arquipélago dos Açores evidenciando que existe apenas um *outlier* no grupo Ocidental identificado pelo município do Corvo. Nos gráficos de dispersão univariados, a observação 1, referente a este município, foi analisada em *ilr(Empregado)* e *ilr(Inativo)*. Constatase que em *ilr(Desempregado)* a observação adota o mesmo comportamento que em *ilr(Inativo)*, pelo que as conclusões retiradas são semelhantes nas duas variáveis.

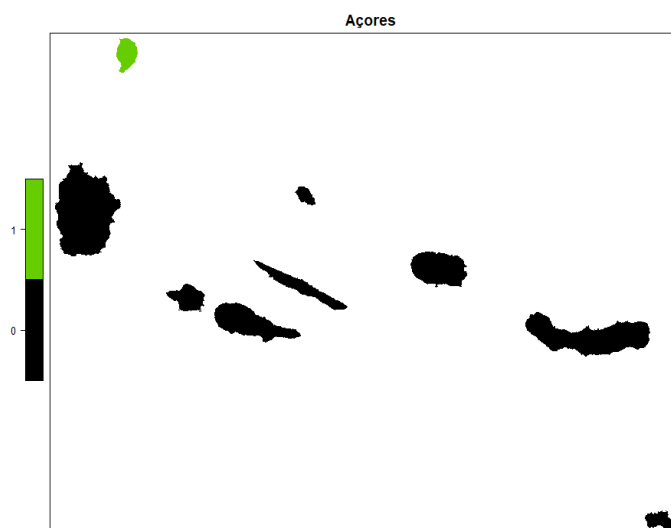


Figura 4.25: Cartograma do Arquipélago dos Açores com identificação do município *outlier* do Corvo a cor verde, no grupo Ocidental, para o conjunto de dados por situação profissional.

Capítulo 5

Conclusão e trabalho futuro

Os trabalhos de Aitchison, em 1986, foram os impulsionadores para a descoberta de dados composicionais. Durante os anos 80, outros autores desenvolveram trabalhos na sequência do estudo de Aitchison e, desde então, a análise de dados composicionais tem ganho uma dimensão suprema apresentando-se em crescente desenvolvimento. A análise destes dados ganha interesse na medida em que se pretende avaliar o rácio entre componentes (variação relativa) e não a diferença entre elas (variação absoluta). Para tal, a finalidade de uma análise multivariada de dados composicionais requer alguma precaução visto que uma transformação apropriada é crucial para que os dados possam ser interpretados no espaço Euclidiano, dado que a geometria de Aitchison é o espaço amostral dos dados composicionais.

Frequentemente, em diversos conjuntos de dados, a presença de *outliers* multivariados constitui um acontecimento interessante para a análise uma vez que revela o aparecimento de fenómenos atípicos camuflados pelo desconhecimento de tais observações. Os dados composicionais não fogem à regra pelo que o estudo da presença de dados atípicos nesta classe de dados se revela também interessante. A identificação de *outliers* multivariados em dados composicionais era um dos objetivos desta dissertação, recorrendo a metodologias numéricas e gráficas. Constatou-se que a transformação *ilr* é essencial para a deteção de observações atípicas, no entanto uma conjugação desta com a transformação *clr* permite que, posteriormente, os dados atípicos possam ser visualizados em biplots composicionais robustos. Averiguou-se também que a distância de Mahalanobis robusta é o principal método para a deteção destas observações e que o estimador MCD é adequado para que uma abordagem robusta desta distância possa ser garantida. Por sua vez, ferramentas exploratórias gráficas permitiram também a identificação de *outliers* multivariados, nomeadamente os biplots composicionais robustos enquanto os diagramas ternários possibilitaram apenas a sua visualização para o conjunto de dados por situação profissional. Para interpretação das observações atípicas, os gráficos de dispersão univariados foram as ferramentas fulcrais para justificar a deteção destas observações.

A matriz de dados composicionais relativa ao conjunto de dados por habilitação académica continha zeros de contagem que foram imputados pelo algoritmo k -NN, para $k = 3$. A imputação desses valores foi realizada com base na seleção de quatro valores diferentes para k , e como os resultados obtidos eram iguais, a escolha de qualquer um dos valores era indiferente. Ora, para que esta escolha não seja baseada exclusivamente numa decisão do analista e para que os resultados sejam mais fidedignos sugere-se que um dos trabalhos futuros seja o aperfeiçoamento do estudo do k ótimo, por exemplo, recorrendo à validação cruzada.

Da análise efetuada aos dados demográficos obteve-se um total de 30 *outliers* para o conjunto do grupo etário, 62 *outliers* para o conjunto da habilitação académica e 11 *outliers* para o conjunto da situação profissional. Curiosamente, os municípios *outliers* têm tendência para se localizar em regiões do interior de Portugal Continental. Somente dois desses municípios são comuns aos conjuntos, tais como Freixo de Espada à Cinta e Pampilhosa da Serra. Porém, não foram identificados municípios comuns para as Regiões Autónomas da Madeira e dos Açores.

Dos três conjuntos, a análise realizada ao conjunto da situação profissional revelou conclusões mais interpretáveis. De um modo geral, constatou-se a existência de um padrão nos municípios *outliers* revelando que a maior parte desses municípios tendem a ser menos atrativos para indivíduos que se encontram na atividade económica empregados, mas também aqueles que se encontram desempregados. Contrariamente, indivíduos que não possuam nenhum contrato de trabalho consideram esses mesmos municípios atrativos escolhendo-os para residir. Relativamente aos restantes dois conjuntos, os resultados não evidenciaram um padrão global que pudesse ser interpretado de modo explícito. O conjunto do grupo etário demonstra que os municípios *outliers* foram atrativos tanto para residentes mais novos como para residentes mais idosos. No entanto, curiosamente, indivíduos com idades entre os 40 e os 64 anos consideraram que os municípios *outliers*, identificados neste conjunto, não eram atrativos para habitar. O conjunto da habilitação académica possui um padrão diversificado, no sentido em que não há uma clara distinção do modo como se possa tipificar/categorizar o grau académico dos novos residentes em cada município *outlier*.

Em suma, os municípios identificados como *outliers* são considerados dados atípicos uma vez que o modo como “atraem” residentes é diferente dos restantes municípios em termos de distribuição do grupo etário, habilitação académica e situação profissional. Portanto, o estudo destes municípios beneficiou a clarificar que a percentagem de pessoas que entraram pode ser muito elevada, mas também muito baixa fazendo com que os municípios identificados sejam mais atrativos ou repulsivos consoante a distribuição do grupo etário, habilitação académica e situação profissional. Do ponto de vista demográfico, as particularidades das composições dos municípios *outliers* detetados são sugeridas para futuros estudos, nomeadamente com equipas multidisciplinares envolvendo estatísticos e demógrafos.

Perante os resultados obtidos salienta-se que, para o conjunto de dados por situação profissional, os gráficos de dispersão univariados revelaram conclusões mais objetivas apresentado uma melhor padronização dos *outliers* identificados. Além disso, a qualidade de representação dos biplots para este conjunto é perfeita ($75.7 + 24.3 = 100\%$) pelo que o padrão de variação dos dados apresentado nos biplots para dados originais é fiável. O mesmo não poderá ser dito relativamente aos restantes dados, pois as percentagens são de $49.1 + 23.3 = 72.4\%$ e $53.6 + 19.7 = 73.3\%$ para o primeiro e segundo conjunto, respetivamente, influenciando os resultados da análise das correlações entre rácios. Deste modo, um dos trabalhos que se propõe para futuro é a visualização de outros planos, ou seja, a construção de biplots composicionais robustos com base noutras componentes principais de forma a que os resultados obtidos com a matriz de correlações possam ser devidamente interpretados.

Uma das limitações deste estudo corresponde ao facto da abordagem relativa considerada não ter em conta o total de residentes que habita em cada município, ou seja, não tem em conta a densidade populacional do município. Por exemplo, um município onde se verifique uma percentagem populacional de 5% pode ser pouco ou muito volumoso, isto é, esta percentagem pode corresponder a 1 ou 100 pessoas pelo que terá uma interpretação demográfica diferente.

As composições dos três conjuntos, dos 308 municípios, foram analisadas separadamente para três fatores distintos: grupo etário, habilitação acadêmica e situação profissional. No entanto, uma análise sob o ponto de vista de tabelas de composição (*compositional tables*) pode revelar inéditas conclusões. Por exemplo, a análise pode ser efetuada de acordo com a conjugação dos conjuntos em dois fatores. Assim sendo, uma sugestão de investigação futura corresponde a estudar a agregação dos conjuntos em tabelas composicionais com dois fatores que permita o estudo das interações. Poder-se-ia analisar pares de conjuntos: grupo etário versus habilitação acadêmica, ou grupo etário versus situação profissional ou, ainda, habilitação acadêmica versus situação profissional. Esta última revela-se importante em termos de políticas públicas na capacidade de empregabilidade dos municípios tendo em conta a escolaridade dos potenciais empregados.

Referências Bibliográficas

- [1] N. V. do Prado. «Abordagens para análise de dados composicionais». Tese de doutoramento. Universidade de São Paulo, 2017.
- [2] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, 1986.
- [3] A. Buccianti, G. Mateus-Figueras e V. Pawlowsky-Glahn. *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Vol. 264. Special Publications, 2006.
- [4] R. C. A. de Sousa. «Análise Estatística de Dados Composicionais». Tese de mestrado. Universidade de Aveiro, 2016.
- [5] V. Pawlowsky-Glahn e A. Buccianti. *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, 2011.
- [6] V. Pawlowsky-Glahn, J. J. Egozcue e R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.
- [7] J. Aitchison e M. Greenacre. «Biplots of compositional data». Em: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 51.4 (2002), pp. 375–392.
- [8] J. Aitchison. «A concise Guide to Compositional Data Analysis». Em: 2nd Compositional Data Analysis Workshop – CoDaWork’05. 2005. URL: http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf.
- [9] V. Pawlowsky-Glahn e J.J. Egozcue. «Geometric approach to statistical analysis on the simplex». Em: *Stochastic Environmental Research and Risk Assessment* 15 (2001), pp. 384–398.
- [10] J.J. Egozcue et al. «Isometric Logratio transformations for Compositional Data Analysis». Em: *Mathematical Geology* 35.3 (2003), pp. 279–300.
- [11] P. Filzmoser, K. Hron e C. Reimann. «Principal component analysis for compositional data with outliers». Em: *Environmetrics* 20 (2009), pp. 621–632.
- [12] D. M. Hawkins. *Identification of outliers*. Chapman & Hall, 1980.
- [13] V. Barnett e T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1978.
- [14] R. A. Maronna, R. D. Martin e V. J. Yohai. *Robust Statistics: Theory and Methods*. 2006.
- [15] P. Filzmoser, A. Ruiz-Gazen e C. Thomas-Agnan. «Identification of local multivariate outliers». Em: *Springer* 55 (2014), pp. 29–47.
- [16] E. Dos S. Ferreira. «Métodos Biplot aplicados a dados de Biologia Molecular». Tese de mestrado. Universidade de Aveiro, 2010.

- [17] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc. Boston, MA, 2016. URL: <http://www.rstudio.com/>.
- [18] *GitHub*. <https://github.com/Leticia-Leite/CompositionalData>.
- [19] P. Filzmoser, K. Hron e M. Templ. *Applied Compositional Data Analysis With Worked Examples in R*. Springer, 2018.
- [20] V. Pawlowsky-Glahn, J.J. Egozcue e R. Tolosana-Delgado. «Lecture Notes on Compositional Data Analysis». Em: (2007), pp. 1–96.
- [21] M. Templ, K. Hron e P. Filzmoser. «Exploratory tools for outlier detection in compositional data with structural zeros». Em: *Journal of Applied Statistics* 44.4 (2017), pp. 734–752.
- [22] J.J. Egozcue e V. Pawlowsky-Glahn. «Groups of Parts and Their Balances in Compositional Data Analysis». Em: *Mathematical Geology* 37.7 (2005), pp. 795–828.
- [23] P. Filzmoser, K. Hron e C. Reimann. «Univariate analysis of environmental (compositional) data: Problems and possibilities». Em: *The Science of the total environment* 407 (2009), 6100–6108.
- [24] K. Hron. «Classical and robust statistical methods for a comprehensive statistical treatment of compositional data». Habilitation thesis. Masaryk University, Faculty of Science, 2012. URL: https://is.muni.cz/do/rect/habilitace/1431/Hron/habilitace/habilitation_thesis-Hron.pdf?lang=en.
- [25] P. Kynčlová, P. Filzmoser e K. Hron. «Compositional biplots including external non-compositional variables». Em: *Statistics* 50 (2016), pp. 1–18.
- [26] E. Fiservá e K. Hron. «On the Interpretation of Orthonormal Coordinates for Compositional Data». Em: *Mathematical Geosciences* 43 (2011), pp. 455–468.
- [27] K. Hron, M. Templ e P. Filzmoser. «Exploratory compositional data analysis using the R-package robCompositions». Em: *Analysis and Applications* 18 (1997), pp. 1–8.
- [28] V. Barnett e T. Lewis. *Outliers in Statistical Data*. 3^a ed. Wiley, New York, 1994.
- [29] R. Gnanadesikan e J. R. Kettenring. «Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data». Em: *Biometrics* 28.1 (1972), pp. 81–124.
- [30] D. Peña e F. Prieto. «Multivariate Outlier Detection and Robust Covariance Matrix Estimation». Em: *Technometrics* 43.3 (2001), pp. 286–310.
- [31] R. Maronna e R. Zamar. «Robust Estimates of Location and Dispersion for High-Dimensional Datasets». Em: *Technometrics* 44.4 (2002), pp. 307–317.
- [32] P. Filzmoser e K. Hron. «Outlier Detection for Compositional Data Using Robust Methods». Em: *Mathematical Geosciences* 40 (2008), pp. 233–248.
- [33] P. Filzmoser, K. Hron e C. Reimann. «Interpretation of multivariate outliers for compositional data». Em: *Computers and Geosciences* 39 (2012), pp. 77–85.
- [34] P. Filzmoser, Robert G. Garrett e C. Reimann. «Multivariate outlier detection in exploration geochemistry». Em: *Computers & Geosciences* 31 (2005), pp. 579–587.
- [35] P. J. Rousseeuw e K. V. Driessen. «A Fast Algorithm for the Minimum Covariance Determinant Estimator». Em: *Journal Technometrics* 41 (1999), pp. 212–223.

- [36] K. Hron, M. Templ e P. Filzmoser. «Imputation of missing values for compositional data using classical and robust methods». Em: *Computational Statistics & Data Analysis* 54 (2010), pp. 3095–3107.
- [37] K. Gerald van den Boogaart e R. Tolosana-Delgado. *Analyzing Compositional Data with R*. Springer, 2013.
- [38] V. Pawlowsky-Glahn e J.J. Egozcue. «BLU estimators and compositional data». Em: *Mathematical Geology* 34.3 (2002), pp. 259–274.
- [39] K. R. Gabriel. «The Biplot Graphic Display of Matrices with Application to Principal Component Analysis». Em: *Biometrika* 58.3 (1971), pp. 453–467.
- [40] C. Eckart e G. Young. «The approximation of one matrix by another of lower rank». Em: *Psychometrika* 1.3 (1936), pp. 211–218.
- [41] R. Wedlake. «Robust Principal Component Analysis Biplots». Tese de mestrado. Universidade de Stellenbosch, 2008. URL: <http://scholar.sun.ac.za/handle/10019.1/2491>.
- [42] A. B. Nieto et al. «A Methodology for Biplots Based on Bootstrapping with R». Em: *Revista Colombiana de Estadística* 37 (2014), pp. 367–397.
- [43] M. Templ, K. Hron e P. Filzmoser. *robCompositions: an R-package for robust statistical analysis of compositional data*. John Wiley e Sons, 2011, pp. 341–355. URL: <https://cran.r-project.org/web/packages/robCompositions>.
- [44] M. Maechler et al. *robustbase: Basic Robust Statistics*. R package version 0.93-3. 2018. URL: <http://CRAN.R-project.org/package=robustbase>.
- [45] P. Filzmoser e M. Gschwandtner. *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*. R package version 2.0.9. 2018. URL: <https://CRAN.R-project.org/package=mvoutlier>.

Apêndice A

Outliers: função `mvoutlier.CoDa()`

A.1 Conjunto de dados por grupo etário

Tabela A.1: Identificação dos *outliers* detetados pela função *mvoutlier.CoDa()* para o conjunto de dados por grupo etário.

Número	Município	Número	Município
1	Aguiar da Beira	17	Melgaço
2	Alcoutim	18	Mira
3	Arouca	19	Monchique
4	Carrazeda de Ansiães	20	Montalegre
5	Castanheira de Pêra	21	Pampilhosa da Serra
6	Castelo de Vide	22	Pinhel
7	Crato	23	Porto Moniz
8	Cuba	24	Santa Marta de Penaguião
9	Figueira de Castelo Rodrigo	25	Tarouca
10	Freixo de Espada à Cinta	26	Torre de Moncorvo
11	Fronteira	27	Trancoso
12	Lajes das Flores	28	Vila de Rei
13	Macedo de Cavaleiros	29	Vila Nova de Foz Côa
14	Manteigas	30	Vila Nova de Paiva
15	Marvão	31	Vizela
16	Meda		

A.2 Conjunto de dados por habilitação académica

Tabela A.2: Identificação dos *outliers* detetados pela função *mvoutlier.CoDa()* para o conjunto de dados por habilitação académica.

Número	Município	Número	Município	Número	Município
1	Aguiar da Beira	22	Ferreira do Alentejo	43	Penela
2	Alandroal	23	Freixo de Espada à Cinta	44	Pinhel
3	Alcoutim	24	Fronteira	45	Ponta do Sol
4	Alfândega da Fé	25	Gavião	46	Ponte de Sor
5	Almeida	26	Lajes das Flores	47	Portel
6	Almodôvar	27	Lajes do Pico	48	Porto Moniz
7	Alvito	28	Mação	49	Povoação
8	Arronches	29	Manteigas	50	Ribeira de Pena
9	Borba	30	Marvão	51	Sabrosa
10	Calheta (R.A.A.)	31	Mesão Frio	52	Santa Comba Dão
11	Campo Maior	32	Monchique	53	Sardoal
12	Carraceda de Ansiães	33	Mondim de Basto	54	Sernancelhe
13	Castanheira de Pêra	34	Monforte	55	Sever do Vouga
14	Castelo de Vide	35	Mortágua	56	Tarouca
15	Castro Daire	36	Mourão	57	Vieira do Minho
16	Castro Marim	37	Nisa	58	Vila de Rei
17	Celorico da Beira	38	Nordeste	59	Vila do Bispo
18	Cinfães	39	Odemira	60	Vila Franca do Campo
19	Corvo	40	Oleiros	61	Vila Nova de Paiva
20	Crato	41	Pampilhosa da Serra	62	Vila Nova de Poiares
21	Estremoz	42	Penalva do Castelo	63	Vinhais

A.3 Conjunto de dados por situação profissional

Tabela A.3: Identificação dos *outliers* detetados pela função *mvoutlier.CoDa()* para o conjunto de dados por situação profissional.

Número	Município	Número	Município
1	Corvo	9	Porto Santo
2	Freixo de Espada à Cinta	10	Ribeira de Pena
3	Idanha-a-Nova	11	Sabugal
4	Mourão	12	Torre de Moncorvo
5	Oleiros	13	Vila Nova de Foz Côa
6	Pampilhosa da Serra	14	Vila Nova de Poiares
7	Penamacor	15	Vila Velha de Ródão
8	Penedono		

Apêndice B

Poster

Detection of outliers municipalities in Portugal: a compositional analysis of occupational status and academic qualification

Letícia Leite¹, Adelaide Freitas^{1,2}, Cristina Gomes^{3,4}

¹Department of Mathematics; ²Center for Research & Development in Mathematics and Applications (CIDMA); ³Department of Social, Political and Territorial Sciences; ⁴Research Unit on Governance, Competitiveness and Public Policies (GOVCOPP).
University of Aveiro, Portugal

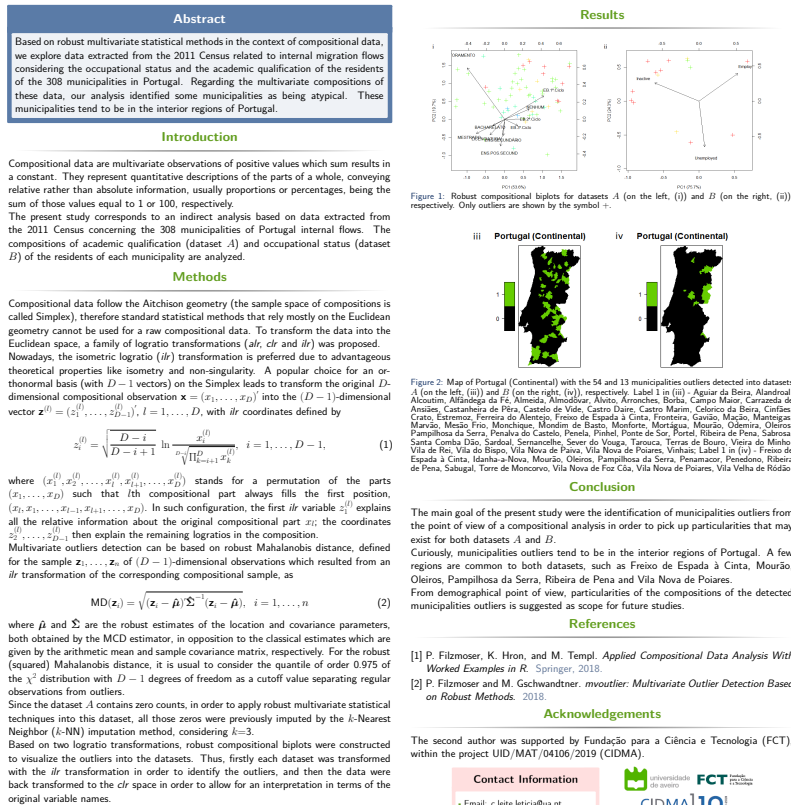


Figura B.1: Poster apresentado nas XXVI Jornadas de Classificação e Análise de Dados, na Escola Superior de Tecnologia e Gestão de Viseu, em Abril de 2019.